Célpont-ligandum kölcsönhatások számítása

rövid formátumú akadémiai doktori értekezés

Hetényi Csaba

Pécsi Tudományegyetem
Általános Orvostudományi Kar
Farmakológiai és Farmakoterápiai Intézet
2024

Tartalomjegyzék

Az értekezés alapját képező saját közlemények jegyzéke

[D1] **Hetenyi, C***.; van der Spoel, D. Blind Docking of Drug-Sized Compounds to Proteins with up to a Thousand Residues. *FEBS Lett.* **2006**, *580* (5), 1447–1450. https://doi.org/10.1016/j.febslet.2006.01.074.

[D2] **Hetenyi, C***.; van der Spoel, D. Toward Prediction of Functional Protein Pockets Using Blind Docking and Pocket Search Algorithms. *Protein Sci.* **2011**, *20* (5), 880–893. https://doi.org/10.1002/pro.618.

[D3] Balint, M.; Jeszenoi, N.; Horvath, I.; Abraham, I. M*.; **Hetenyi, C***. Dynamic Changes in Binding Interaction Networks of Sex Steroids Establish Their Non-Classical Effects. *Sci Rep* **2017**, *7*, 14847. https://doi.org/10.1038/s41598-017-14840-9.

[D4] Balint, M.; Jeszenoi, N.; Horvath, I.; van der Spoel, D.; **Hetenyi, C***. Systematic Exploration of Multiple Drug Binding Sites. *J. Cheminformatics* **2017**, *9*, 65. https://doi.org/10.1186/s13321-017-0255-6.

[D5] **Hetenyi, C***.; Balint, M. Systematic Exploration of Binding Modes of Ligands on Drug Targets. In S*tructural bioinformatics: methods and protocols*; Methods in Molecular Biology, Humana Press Inc: Totowa, Springer Science+Business Media, LLC, part of Springer Nature 2020; Vol. 2112, pp 107–121. https://doi.org/10.1007/978-1-0716-0270-6_8.

[D6] Zsido, B. Z.; Borzsei, R.; Pinter, E.; **Hetenyi, C***. Prerequisite Binding Modes Determine the Dynamics of Action of Covalent Agonists of Ion Channel TRPA1. *Pharmaceuticals* **2021**, *14* (10), 988. https://doi.org/10.3390/ph14100988.

[D7] Balint, M.; Zsido, B. Z.; van der Spoel, D.; **Hetenyi, C***. Binding Networks Identify Targetable Protein Pockets for Mechanism-Based Drug Design. *Int. J. Mol. Sci.* **2022**, *23* (13), 7313. https://doi.org/10.3390/ijms23137313.

[D8] Zsido, B. Z.; **Hetenyi, C***. Molecular Structure, Binding Affinity, and Biological Activity in the Epigenome. *Int. J. Mol. Sci.* **2020**, *21* (11), 4134. https://doi.org/10.3390/ijms21114134.

[D9] Org, T.; Chignola, F.; **Hetenyi, C**.; Gaetani, M.; Rebane, A.; Liiv, I.; Maran, U.; Mollica, L.; Bottomley, M. J.; Musco, G.; Peterson, P. The Autoimmune Regulator PHD Finger Binds to Non-Methylated Histone H3K4 to Activate Gene Expression. *EMBO Rep.* **2008**, *9* (4), 370–376. https://doi.org/10.1038/embor.2008.11.

[D10] Börzsei, R.; Bayarsaikhan, B.; Zsidó, B. Z.; Lontay, B.; **Hetényi, C***. The Structural Effects of Phosphorylation of Protein Arginine Methyltransferase 5 on Its Binding to Histone H4. *Int. J. Mol. Sci.* **2022**, *23* (19), 11316. https://doi.org/10.3390/ijms231911316.

[D11] Balint, M.; Horvath, I.; Meszaros, N.; **Hetenyi, C***. Towards Unraveling the Histone Code by Fragment Blind Docking. *Int. J. Mol. Sci.* **2019**, *20* (2), 422. https://doi.org/10.3390/ijms20020422.

[D12] Zsidó, B. Z.; Bayarsaikhan, B.; Börzsei, R.; **Hetényi, C***. Construction of Histone–Protein Complex Structures by Peptide Growing. *Int. J. Mol. Sci.* **2023**, *24* (18), 13831. https://doi.org/10.3390/ijms241813831.

[D13] Börzsei, R.; Zsidó, B. Z.; Bálint, M.; Helyes, Z.; Pintér, E.; **Hetényi, C***. Exploration of Somatostatin Binding Mechanism to Somatostatin Receptor Subtype 4. *Int. J. Mol. Sci.* **2022**, *23* (13), 6878. https://doi.org/10.3390/ijms23136878.

[D14] Zsidó, B. Z.; **Hetényi, C***. The Role of Water in Ligand Binding. *Current Opinion in Structural Biology* **2021**, *67*, 1–8. https://doi.org/10.1016/j.sbi.2020.08.002.

[D15] Zsidó, B. Z.; Bayarsaikhan, B.; Börzsei, R.; Szél, V.; Mohos, V.; **Hetényi, C***. The Advances and Limitations of the Determination and Applications of Water Structure in Molecular Engineering. *Int. J. Mol. Sci.* **2023**, *24* (14), 11784. https://doi.org/10.3390/ijms241411784.

[D16] Jeszenoi, N.; Horvath, I.; Balint, M.; van der Spoel, D.; **Hetenyi, C***. Mobility-Based Prediction of Hydration Structures of Protein Surfaces. *Bioinformatics* **2015**, *31* (12), 1959–1965. https://doi.org/10.1093/bioinformatics/btv093.

[D17] Jeszenoi, N.; Balint, M.; Horvath, I.; van der Spoel, D.; **Hetenyi, C*.** Exploration of Interfacial Hydration Networks of Target Ligand Complexes. *J. Chem Inf. Model.* **2016**, *56* (1), 148–158. https://doi.org/10.1021/acs.jcim.5b00638.

[D18] Jeszenoi, N.; Schilli, G.; Balint, M.; Horvath, I.; **Hetenyi, C***. Analysis of the Influence of Simulation Parameters on Biomolecule-Linked Water Networks. *J. Mol. Graph.* **2018**, *82*, 117–128. https://doi.org/10.1016/j.jmgm.2018.04.011.

[D19] Zsido, B. Z.; Borzsei, R.; Szel, V.; **Hetenyi, C***. Determination of Ligand Binding Modes in Hydrated Viral Ion Channels to Foster Drug Design and Repositioning. *J. Chem Inf. Model.* **2021**, *61* (8), 4011–4022. https://doi.org/10.1021/acs.jcim.1c00488.

[D20] Fodor, K.; Harmat, V.; **Hetenyi, C**.; Kardos, J.; Antal, J.; Perczel, A.; Patthy, A.; Katona, G.; Graf, L. Extended Intermolecular Interactions in a Serine Protease-Canonical Inhibitor Complex Account for Strong and Highly Specific Inhibition. *J. Mol. Biol.* **2005**, *350* (1), 156–169. https://doi.org/10.1016/j.jmb.2005.04.039.

[D21] **Hetenyi, C***.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins. *J. Am. Chem. Soc.* **2006**, *128* (4), 1233–1239. https://doi.org/10.1021/ja055804z.

[D22] Balogh, B.; **Hetenyi, C***.; Keseru, M. G.; Matyus, P*. Structure-Based Calculation of Binding Affinities of Alpha(2A)-Adrenoceptor Agonists. *ChemMedChem* **2007**, *2* (6), 801–805. https://doi.org/10.1002/cmdc.200600251.

[D23] **Hetenyi, C***.; Maran, U.; Karelson, M. A Comprehensive Docking Study on the Selectivity of Binding of Aromatic Compounds to Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1576–1583. https://doi.org/10.1021/ci034052u.

[D24] Zhang, H.; Tan, T.; **Hetenyi, C**.; van der Spoel, D. Quantification of Solvent Contribution to the Stability of Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9* (10), 4542–4551. https://doi.org/10.1021/ct400404q.

[D25] Zhang, H.; Tan, T.; **Hetenyi, C**.; Lv, Y.; van der Spoel, D. Cooperative Binding of Cyclodextrin Dimers to Isoflavone Analogues Elucidated by Free Energy Calculations. *J. Phys. Chem. C* **2014**, *118* (13), 7163–7173. https://doi.org/10.1021/jp412041d.

[D26] Horvath, I.; Jeszenoi, N.; Balint, M.; Paragi, G.; **Hetenyi, C***. A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level. *Int. J. Mol. Sci.* **2019**, *20* (18), 4384. https://doi.org/10.3390/ijms20184384.

[D27] Szél, V.; Zsidó, B. Z.; Jeszenői, N.; **Hetényi, C***. Target–Ligand Binding Affinity from Single Point Enthalpy Calculation and Elemental Composition. *Phys. Chem. Chem. Phys.* **2023**, *25* (46), 31714–31725. https://doi.org/10.1039/D3CP04483A.

[D28] **Hetenyi, C***.; Maran, U.; Garcia-Sosa, A. T.; Karelson, M. Structure-Based Calculation of Drug Efficiency Indices. *Bioinformatics* **2007**, *23* (20), 2678–2685. https://doi.org/10.1093/bioinformatics/btm431.

[D29] Garcia-Sosa, A. T.; **Hetenyi, C**.; Maran, U. Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions. *J. Comput. Chem.* **2010**, *31* (1), 174–184. https://doi.org/10.1002/jcc.21306.

[D30] Garcia-Sosa, A. T.; Maran, U.; **Hetenyi, C***. Molecular Property Filters Describing Pharmacokinetics and Drug Binding. *Curr. Med. Chem.* **2012**, *19* (11), 1646–1662. https://doi.org/10.2174/092986712799945021.

Az élet az anyagnak egy sajátsága, szerkezetének következménye.[1]

Szent-Györgyi Albert

Rövidítések jegyzéke

1SCF	önkonzisztens tér számítás egy szerkezeti végpontra
AIRE	autoimmun regulátor fehérje
AMBER	Assisted Model Building and Energy Refinement, egy MM erőtér
BD	blind docking, a teljes célpontra kiterjesztett kötési mód keresés
DL	drug-likeness, gyógyszerszerűség
DNS	dezoxiribonukleinsav
EI	efficiency index, hatékonysági index
$\mathbf{E_{teljes}}$	teljes (molekulamechanikai) potenciális energia
FBD	fragment blind docking, BD molekulafragmensekre alkalmazva
HPC	high performance computing, szuperszámítástechnika
HTS	high throughput screening, nagy áteresztőképességű szűrés
ITC	isothermal titration calorimetry, izotermális titrációs kalorimetria
KEM	krio-elektronmikroszkópia
MD	molekuláris dinamika
MEP50	metiloszóma protein 50
MM	molekulamechanika
NHA	number of heavy atoms, a hidrogéntől eltérő atomok száma
NMR	mágneses magrezonancia
$\mathbf{N_{tor}}$	azon torziós szögek száma, amelyek mentén a ligandum szabadon elforog
OPLS	Optimized Potentials for Liquid Simulations, egy MM erőtér
PHD	plant homeodomain, növényi homeodomén
PM7	Parametric Method 7, egy szemiempírikus QM parametrizálás
PMF	potential of mean force, az átlagos erő potenciálja
PRMT5	protein arginin metiltranszferáz 5
PS	pocket search, kötőzsebkereső
PTM	poszt-transzlációs módosulás
QM	kvantummechanika
RKR	röntgen-krisztallográfia
SGCI	sivatagi sáska kimotripszin inhibitor
SSTR4	a szomatosztatin 4-es altípusú receptora
TRP	tranziens receptor potenciál
$\mathbf{\Delta G_b}$	kötődési szabadentalpiaváltozás
$\mathbf{\Delta H_b}$	kötődési entalpiaváltozás
$\mathbf{\Delta S_b}$	kötődési entrópiaváltozás

1. Az értekezés tárgya

A gyógyszertervezés célja a farmakodinámia szempontjából egy olyan ligandum meg- vagy kitalálása, amely erősen és szelektíven kötődik a célponthoz, hogy azon hatását kifejthesse. Farmakokinetikai elvárás, hogy a ligandum eljusson az adminisztráció helyétől a célponthoz, majd pedig a megfelelő idő elteltével távozzon a szervezetből. A továbbiakban ligandum alatt értem nem csupán a gyógyszereket, de az élettanilag fontos endogén vegyületeket, xenobiotikumokat és általában a farmakológiai aktivitást mutató vegyületeket. A célpont többnyire egy fehérjemolekulát takar majd, de a dolgozat eredményei könnyen adaptálhatóak más bio- vagy makromolekuláris partnerekre is. A ligandumok teljes farmakológiai (farmakodinámiai és farmakokinetikai) hatékonysága alapvetően függ a szerkezetüktől és a célponttal, valamint testünk egyéb molekuláival való kölcsönhatásaik térbeli elrendeződésétől és erősségétől.

Nem véletlen tehát, hogy munkám során Szent-Györgyi Albertnek az előző lapon olvasható gondolata volt az egyik fogódzóm. Az életfunkciók és a farmakológiai jelenségek valóban mind anyagszerkezeti alapokon nyugszanak, ezért leírásuk elsődlegesen a kémia hatáskörébe tartozik. Az anyag szerkezetének atomi felbontású ismerete nyújtja a biztos talajt a magasabb szinteken végbemenő élettani folyamatok legpontosabb értelmezéséhez és a patológiás események molekuláris szintű, gyógyszeres kezeléséhez is.

Ennek megfelelően, dolgozatom elején az anyagszerkezetre vonatkozó eredményeimet mutatom be, a célpont-ligandum kölcsönhatásokra fókuszálva. A ligandum farmakológiai profiljának felrajzolásakor, az atomi felbontású szerkezetek birtokában az első logikus lépés a célponttal kialakítható kölcsönhatások erősségének a meghatározása. Erre lehetőséget nyújt a szerkezeti alapú energiaszámítás és végső soron a termodinamikai potenciálfüggvények változásainak meghatározása a kölcsönhatás során. Dolgozatom második része ezért a célpont-ligandum kölcsönhatások energiájával foglalkozik. Mivel a szerkezeti számítások is igénylik egyes energiatípusok használatát, a szerkezet és energia területei természetesen átfednek. Az eredményeket így elsődleges jellegük szerint soroltam e két külön fejezetbe. Az energiáról szóló fejezetben azokat az eredményeimet is sorra veszem, amelyek az előzőekre épülve a ligandum farmakológiai hatékonyságával kapcsolatosak.

A racionális, célpont-alapú gyógyszertervezés tehát a szerkezet-energia kapcsolatok feltárásával és alkalmazásával jut el jó eséllyel a farmakológialiag hatékony ligandumokig. A továbbiakban az említett kapcsolatokra épülő, új eszközök kidolgozásáról, valamint a meglévők praktikus alkalmazásairól számolok be. Az MTA Kémiai Osztály Doktori Bizottsága vezetőségével történt egyeztetés alapján a jelen dolgozat a rövid értekezés formát követi. Ennek megfelelően tömörítve adom meg az elért eredmények értékelő összefoglalását. Az arányosság megtartása végett az egyes fejezetek bevezető szakaszaiban az irodalmi hátteret és a felhasznált módszertant is csupán vázlatosan ismertetem, a vonatkozó, válogatott alapművek és összefoglaló tanulmányok tételes hivatkozásával.

A gyógyszerkutatás elméleti módszertana az értekezésben áttekintett közel két évtized alatt nagy változásokon ment keresztül, köszönhetően a rendelkezésre álló informatikai eszköztár fejlődésének is. Kezdetben jórészt személyi számítógépek álltak rendelkezésünkre és az alkalmazott szoftverek párhuzamosítása gyerekcipőben járt. Ma már a hazai szuperszámítógépes (*high performance computing, HPC*) központban is kihasználhatjuk a jól megírt, párhuzamosított kódok adta gyorsaságot, amely a hatékony, reprodukálható munkának is szükséges alapfeltétele lett.

2. Szerkezet

E fejezetben olyan munkák eredményeiről számolok be, amelyeket konkrét gyakorlati igények motiváltak: többnyire a kísérletes szerkezetmeghatározás korlátai. A munka során a célpont-ligandum komplexek szerkezeti számításainak fejlesztésére összpontosítottunk. Folytattuk és lényegében befejeztük az egyetemi doktori disszertációmban megkezdett munkát a számítógépes dokkolás területén, majd ennek eredményeire építettük a prerekvizit (előfeltételi) kötőhelyek feltérképezését. A fehérje-peptid komplexek és a hidrátszerkezet számítására irányuló eljárások a szerkezetmeghatározás nagy kihívásaira adnak válaszokat. Mielőtt rátérek az említett eredmények ismertetésére, röviden igyekszem bemutatni a motiváló előzményeket, valamint a felhasznált elméleti és módszertani hátteret.

2.1. A kísérletes szerkezetmeghatározás korlátai

A célpont-ligandum komplexek atomi felbontású szerkezetének meghatározására több kísérletes eljárás is használható. Bár az egyes technikák állandó fejlődésen mennek keresztül, még ma sem megoldott a gyógyszertervezéshez elvárható nagy számú célpont-ligandum komplex szerkezet gyors és pontos előállítása. A következőkben a kísérleti eljárások ide vonatkozó főbb korlátait tekintem át a tervezés szemszögéből, mivel egyrészt ezek szolgáltak fő motivációként munkámban, másrészt viszonylag kevés tanulmány elemzi őket – ugyanakkor az eljárások méltatásáról számos összefoglaló mű készült már. A korlátokról írok elsősorban, de ezzel véletlenül sem azt szeretném sugallni, hogy a mérésekre már nincs vagy a közeljövőben nem lesz szükség. Sőt, a szerkezeti biokémia jelenlegi állapotának ismeretében és napi tapasztalataim alapján is teljes mértékben osztom például Moore, Hendrickson, Henderson és Brunger professzorok véleményét[2] napjaink népszerű, a gépi tanulás elvén működő AlphaFold[3] nevű számítógépes eljárásáról, amelyet nemrég a fehérjeszerkezet-meghatározás végső megoldásaként ünnepeltek. Véleményük ide vonatkozó részét szó szerint idézem: *„... although structural predictions by AlphaFold and RoseTTAfold may be accurate enough to assist with experimental structure determination, they* ***alone cannot provide*** *the kind of detailed understanding of molecular and chemical interactions that is required for studies of molecular mechanisms and for* ***structure-based drug design****."* Az elméleti megközelítések tehát mindig is rá lesznek utalva a kísérleti eredményekre legalább a validálásuk végett – a kísérletek viszont a jó elmélet adta keretek között hasznosulnak igazán. *Nota bene*, az említett, ismeretalapú, gépi tanulásos eljárások eleve nem is létezhetnének kísérletileg kimért szerkezetek nélkül, hiszen ezeken történik a betanításuk.

A kísérleti módszerek listájának legelején említeném a krio-elektronmikroszkópiát (KEM), amely a legfiatalabb eljárás (kémiai Nobel-díj, 2017)[4], és a nagy méretű részecskék (fehérje-komplexek, riboszóma, vírusok stb.)[5–8] kimérésében lenyűgöző eredményeket ért el. Ugyanakkor fontos szerkezeti részletekben, mint például a ligandum kötődési módjának, vagy a vízmolekulák helyzetének[9] meghatározásában a célponton sajnos még korlátozott a teljesítőképessége[10–12]. E korlátokat számszerűen elsősorban az jelzi, hogy a 2 Å-nál kisebb felbontást – amely a molekulatervezés e fontos részleteihez szükséges lenne – elméletileg a rendelkezésre álló KEM apparátusok el tudják érni, de rutinszerűen a gyakorlat még nem produkálja[13], elsősorban mintaelőkészítési, homogenitási és stabilitási problémák miatt.

A mai napig a röntgen-krisztallográfiás (RKR)[14,15] eljárások adják a Fehérje Adatbankban[16] elérhető célpont-ligandum komplex szerkezetek zömét. A RKR és a KEM várhatóan még jó darabig komplementer technikák lesznek[11,17]. Bár a célpont-ligandum komplexek és a gyógyszertervezés terén a RKR szerepe vitathatatlan, e technika is számos, zavaró korláttal rendelkezik[18,19], amelyek közül a következőkben néhány általános korlátot, majd pedig a célpont-ligandum komplexekre és a hidrátszerkezetre vonatkozó speciálisabb problémákat érintek.

Az első és legnagyobb probléma rögtön a fehérje célpont előállíthatósága és kristályosíthatósága. A genetikusok jelenleg kb. 20 ezerre teszik[20] a humán DNS-ben kódolt fehérjék számát. (Megjegyzem, hogy pár évtizeddel ezelőtt ez az adat nagyságrendekkel nagyobb volt[21] és a pontos számot ma sem ismeri senki, de növekedni már biztos nem fog, ami – a gyógyszertervezés szempontjából nézve – elég aggasztó.) E nem túl nagy fehérjekészletnek csak olyan 10-14 %-a alkalmas szerkezetileg a gyógyszertervezésre (*druggable*)[22] és ennek is csak egy része köthető kórélettanilag valamely betegséghez, így véső soron, jó ha mintegy 1500 humán fehérje alkalmas tervezési célpontként. Ugyanakkor a fehérjekészlet[18] és az aktív gyógyszerek célpontjai esetében[22] is ezek jó harmada membránfehérje (receptor, ioncsatorna, transzpoter). Mivel a membránfehérjék kristályosítása oldhatósági okokból körülményes[23], így az említett csoport nagy hányadának a szerkezetmeghatározása nem rutin feladat. Megjegyzendő, hogy a KEM a fehérjék oldhatósági problémáitól kevésbé függ[24], így bizonyos membránfehérjék esetében kiválthatja a RKR-t, de a minta inhomogenitása ez esetben is kritikus tényező lehet. A membránfehérjék példája mellett a fehérjék izolálásának vagy expressziójának, tisztításának[25] és kristályosítás „művészetének" számos egyéb buktatója is van, amelyekre itt nem térek ki. Ráadásul, a kristályos állapot természetszerűleg különbözik a fehérjék élettani közegétől és így gyakran kristályosodási műtermékek keletkezhetnek, például olyan intermolekuláris kapcsolatok, amelyek a sejtben vagy oldatban nem jönnének létre. Ezt a jelenséget az említett membránfehérjék meghatározásánál használt detergensek felerősíthetik[26] és mindezek miatt külön metodika vált szükségessé[27] e műtermékek felderítésére. A kristály birtokában maga a mérési adatgyűjtés viszonylag gyorsan lezajlik. Ennek kapcsán a nagy energiájú sugárzás okozta mintaroncsolódás és részben az ennek megakadályozására gyakran alkalmazott alacsony hőmérsékletű (90-120 K-en végzett) mérés során jelentkező kriosztatikus műtermékek problémáját említem[28]. E nem kívánt műtermékek főleg az oldószerhez kapcsolt folyamatok során képződhetnek és érinthetik a hidratációs egyensúlyok eltolódása mellett az erősen hidratált oldalláncok konformációs változásait és a gyenge ligandum kötődéseket is. Természetszerűleg hasonló problémák a KEM esetében is előfordulhatnak. A kristályosítás és emellett a kriosztatikus körülmények miatt természetes, hogy a RKR bizonyos kivételektől eltekintve[29,30] nem térképezi fel a rendszerek időbeli evolúcióját, különösen nem azok hosszabb távú molekuláris dinamikáját, amely az élettani körülmények között természetes jelenség, akár nagyobb, például domén mozgások szintjén is. Ehelyett pillanatfelvételt rögzít[31], amelyről többnyire hallgatólagosan elfogadjuk, hogy a legjobban megközelítheti a valóságot. A mérési adatok interpretálása és az atomi pozícióknak az elektronsűrűség-térképbe történő illesztése viszonylag gyors, algoritmizált folyamat, amelyhez több kiváló szoftver[32–35] alkalmazható. A fehérjék inherens molekuláris dinamikája (az anizotróp mozgások miatti diszkrét konformációs állapotokból eredő szerkezeti heterogenitás) azonban a szoftverek által biztosított gyors illesztések mellett is eredményez egy természetes bizonytalanságot[36] a kapott szerkezeti modell tekintetében és ezt – különösen az alacsonyabb felbontású szerkezetek interpretálásakor – célszerű észben tartani. Emellett a problémásabb elektronsűrűség-térképek illesztésekor szükséges lehet a kutató manuális beavatkozása is, ami természetesen lassítja és további hibával terhelheti a folyamatot.

A fent vázolt általános problémák mellett a célpont-ligandum komplexek esetében a nagyáteresztőképességű RKR mérések[37] fejlődése viszont bíztató tendencia. Az adatgyűjtés egy kristályról a szinkrotronban kevesebb mint 2 percet vesz igénybe, majd a szerkezet megoldása és finomítása további egy órát[11]. Az ilyen nagy áteresztőképességű projektek leginkább a fragmens alapú tervezést segítik[38], amennyiben rendelkezésre áll például 500, különböző ligandum-fragmenst tartalmazó fehérjekristály, úgy az adatgyűjtés és kiértékelés ezekre egy nap és egy hét alatt elvégezhető[11]. Az áteresztőképesség korlátja a célpont-ligandum komplex szerkezetek esetében tehát nem maga a diffrakciós adatok gyűjtése és feldolgozása, hanem a kristályok előállítása, amint azt az általános korlátoknál már említettem. A komplex kristályait leginkább úgy állítják elő, hogy a ligandumot a fehérjével együtt kristályosítják, vagy pedig a kész fehérje kristályt a ligandum oldatába áztatják. Mindkét esetben számos paraméter ideális beállítása szükséges, amit hosszú próbálgatás előz meg és még így is nagyon alapos manuális ellenőrzés szükséges ahhoz, hogy a kristályosodási műtermékeket kizárják és csak a valódi komplex-szerkezeteket tartalmazó kristályokat értékeljék ki[39].

A célpont-ligandum komplexek hidrátszerkezet-meghatározásának legfontosabb technikája már régóta[40] szintén a RKR. A tervezés szempontjából legfontosabb vízmolekulák a célpont felszínén és a célpont-ligandum interfészben helyezkednek el, amelyek pontosabb kiméréséhez is legalább az említett 2 Å felbontás elérése lenne szükséges[41–43].

A vízszerkezet finomítására több számítógépes eljárás születik folyamatosan[44,45]. A viszonylag régóta folyó fejlesztések[46] ellenére a mai napig nem tekinthető rutin feladatnak a vízmolekulák komplexbeli pozíciójának meghatározása, elsősorban azok nagy mobilitása és összességében sok szabadsági foka miatt. A teljes rendszer mérete befolyásolja a vízszerkezet pontosságát is, kisebb fehérjék esetén az atomi pozíciók illesztése az elektronsűrűség-térképbe könnyebb, mint nagy rendszereknél[42]. A vízmolekulák esetében további nehézséget jelent, hogy az egyedülálló oxigén atomjaik elektronsűrűség-csúcsai rendszerint kisebbek, mint a célpont szomszédos atomcsoportjaié, valamint a hidrogén atomoknak is igen kicsi szórása, ami nem javít a helyzeten[42]. Ilyen módon egy víz rendszerint sokkal kisebb értelmezhető jelet produkál, mint a környezete és sokszor nehéz attól elkülöníteni. Az izoelektronos ionok is tovább nehezítik az adott csúcs hozzárendelését[45]. A fehérjék hidrátszerkezete a kristályban sokszor nagyon eltér az oldatban tapasztalhatótól[47], amelynek okozói sokszor a kristályban létrejövő művi fehérje-fehérje kapcsolatok. E kapcsolatok interfész felszíne elérheti akár az oldatban hozzáférhető teljes fehérjefelszín 30-40 %-át is kisebb fehérjék esetében[48], tehát a vízszerkezet igen nagy hányada műterméknek tekinthető és korlátozottan használható ez esetekben. A kriosztatikus körülmények okozta – már említett – problémák a vízszerkezet esetében fokozottan érvényesek[28]. Előfordul ugyanakkor az is, hogy a szerkezeti finomítások során a kedvezőbb illesztési statisztika túlzott mértékben befolyásolja a végeredményt[43]. A neutron-diffrakciós technikák[49,50] sokat segíthetnek a hidrátszerkezetek megbízhatóságának javításában[51], ugyanakkor e technikák komplexitása és hozzáférhetősége lényegesen korlátozza[52,53] még azok használatát.

Végezetül, de nem utolsó sorban megemlíteném a mágneses magrezonancia spektroszkópiát (NMR)[54,55], amelynek főként kisebb fehérjék esetében van jelentősége és igen hasznos információkat nyújt a rendszerek molekuláris mozgásairól, valamint a célpont-ligandum kölcsönhatások forró pontjairól is. A hidrátszerkezet esetében számszerűen kevesebb eredményt szolgáltatott eddig, viszont annak dinamikáját ígéretes módszerekkel jellemzi[56,57].

Mindent összevetve, a kísérletes szerkezetmeghatározási eljárások használatának korlátai monetáris és technikai jellegűek. Az elmúlt évtizedekben mindkét területen pozitív elmozdulás volt észlelhető, egyrészt a gyógyszerfejlesztésbe áramló tőke, másrészt a folyamatos módszerfejlesztői munka révén. E tendencia remélhetően megmarad, ugyanakkor az egyes módszereknek mára kirajzolódtak olyan technikai korlátai is, amelyek csak nagyon nehezen, vagy egyáltalán nem haladhatók meg. Különösen e reménytelennek tűnő esetekben, valamint a szerkezetmeghatározás és -finomítás gyorsaságának növelésében (a monetáris erőforrások hatékonyabb felhasználásában) segítenek a következő szakaszokban ismertetett elméleti megközelítések.

2.2. A felhasznált elméleti megközelítések

A következő szakaszokban a munkám során felhasznált elméleti eljárásokról nyújtok egy vázlatos áttekintést. E fejezetben a molekulamechanikai (MM) szintre szorítkozom, a kvantumkémai (QM) módszereket a **3. fejezet**ben fogom érinteni az energiaszámításhoz kapcsolódóan. Az MM szintű modellezés részletes ismertetésére sem vállalkozom e szűk keretek között. E területről magyar[58] és angol[59,60] nyelvű könyvek, könyvfejezetek, összefoglaló szakcikkek nagy számban elérhetők. A molekulamechanika elméletének csupán a legfontosabb kiindulópontjait és az eredményeim értelmezéséhez szükséges elemeit tárgyalom, valamint külön szólok a vízmodellekről.

2.2.1. Molekulamechanika

A molekulamechanika kiindulópontjának D. H. Andrews 1930-as cikkét[61] szokás tekinteni, amely posztulátumaiban megfogalmazza a Raman spektrumok értelmezésének klasszikus mechanikai alapjait. Ezt követték N. L. Allinger és mások úttörő munkái[62–65], és az 1950-es évektől a fejlődés a számítástechnikai háttér kiépülésével párhuzamosan egyre rohamosabb lett[66]. Egy nagy ugrással megemlítem még a 2013-as kémiai Nobel-díjat, amely a terület általános elismertségét fémjelzi[67–69].

A molekulamechanikai eljárások fizikai-kémiai kódját az erőterek (force field), matematikai motorját pedig a kereső algoritmusok adják. E két alkatrész egyaránt nélkülözhetetlen és ma is nagy fejlődésen mennek át[70]. Az erőtér molekulamechanikai értelemben egy potenciálisenergia-függvényt és a benne szereplő, atomtípusoktól függő paraméterkészletet jelenti. A molekulamechanika klasszikus megközelítése tehát atomtípusokra épül és (szemben a kvantummechanikával) nem foglalkozik az elektronszerkezet leírásával. Az erőtereket osztályokba szokás sorolni a bennük alkalmazott megközelítések alapján[71–73].

Az 1. osztály erőtereit használjuk leggyakrabban a biomolekuláris célpontok és ligandum-komplexeik számításában. Jellemzőjük, hogy a kötés nyújtási és kötésszöghajlítási tagokra harmonikus megközelítést alkalmaznak, kereszt-tagokat nem használnak, valamint a másodlagos kötésekkel atompáronként számolnak el. Mindezek a számítógépidő nagymértékű csökkenését teszik lehetővé az akár több 10-100 ezer atomos (vagy még nagyobb) rendszerekre. Az *Assisted Model Building and Energy Refinement* (AMBER, **1. egyenlet**)[74–78] klasszikus példája az 1. osztályba tartozó erőtereknek.

$$E_{teljes} = \sum_{kötések} K_r\left(r - r_{eq}\right)^2 + \sum_{szögek} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{torziók} \frac{V_n}{2}[1 + \cos(n\varphi - \gamma)] + \\ + \sum_{i<j} \frac{q_i q_j}{\varepsilon R_{ij}} + \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \qquad (1)$$

A teljes potenciális energiát (E_{teljes}) – amely értelemszerűen a rendszerben lévő összes atom helyvektorától függ – a kovalens kötésben részt vevő atomok közötti (*bonding*, felső sor), valamint az alsó sorban feltüntetett másodlagos (*nonbonding*) kölcsönhatások összegeként írja fel. A felső sorban K_r és K_θ a kötés nyújtásra, kötésszög deformációra vonatkozó erőállandók, V_n a rotációs energiagát nagyságától függő torziós paraméter, r_{eq} és θ_{eq} az egyensúlyi kötéstávolság és kötésszög, n a multiplicitás és γ a torziós fázisszög. Az aktuális geometriát a r kötéstávolságok, θ kötésszögek, és ϕ torziós szögek írják le, valamint a másodlagos kötések (alsó sor) esetében a R_{ij} kötéstávolságok az i és j sorszámú atomok között.

A másodlagos kötések közül a parciális töltések (q) közötti, páronkénti kölcsönhatást leíró Coulomb-formulát és a Lennard-Jones potenciált alkalmazza az erőtér (alsó sor), ahol ε a dielektromos állandó, A és B a kölcsönható i-j atompárok típusától függő Lennard-Jones paraméterek. Az **1. egyenlet**et a szerzők által többször (újra)közölt formában adtam meg. A hidrogénkötésekre az újabb AMBER erőterek már nem tartalmaznak külön kifejezést. Korábban a Lennard-Jones 12-10-es függvényt használták erre a célra. Az AMBER ma is az egyik leggyakrabban használt erőtér, folyamatosan jelennek meg újabb fejlesztései[79–81] és általánosított verziója[82], amelyek mintegy félszáz atomtípust tartalmaznak. Ennél jóval több, ténylegesen használható atomtípust csak ritkán találunk erőterekben. Bár például az Universal Force Field[83] névlegesen a teljes periódusos rendszerre alkalmazható, ez a valóságban csak korlátozottan igaz. Az erőterek parametrizálásának aktuális kérdéseire nem térnék itt ki részletesen, ennek a meglehetősen összetett területnek a jelenlegi állapotáról például van der Spoel professzor közölt nemrég egy áttekintést[70]. Az AMBER mellett gyakrabban használt erőterek az 1. osztályban a CHARMM[84,85] és az OPLS[86,87]. A 2. osztályba tartozó erőterek rendszerint nagyobb számítási igényű függvényeket és kereszt-tagokat tartalmaznak, mint például az MM1-3[88–91], MM4[92], CFF[93,94], MMFF[95]. A 3. osztályba tartozó erőterek polarizációs és hiperkonjugációs effektusok kezelésére képesek[96,97], ilyen például az AMOEBA[98,99] erőtér. Amennyiben az adott molekuláris rendszerhez rendelkezésünkre áll az erőtér és benne a szükséges paraméterkészlet, úgy elvégezhető a rendszer szerkezeti optimalizálása. Az optimalizálás során az említett E_{total} függvényen (**1. egyenlet**) kell elvégezni a minimumhelyek megkeresését. Az optimalizálási probléma alapvetően két kihívással küzd a biológiai rendszerek nagysága és komplexitása miatt. A számítási kapacitás korlátja egy tecnikai kihívás, amely egy N atomból álló rendszer esetében, például a fent említett (**1. egyenlet**) másodlagos kötéseknél $O(N^2)$ rendű operációt takar. E kapacitási probléma párhuzamosított kódok alkalmazásával és a hazánkban is elérhető HPC infrastruktúra alkalmazásával részben kezelhető. A másik, elméleti probléma a nagy számú, egymástól sokszor nehezen megkülönböztethető mélységű minimumvölgyek feltérképezése. Ez lokális vagy globális[100] kereső módszerek segítségével tehető meg. A lokális keresők, mint például a gradiens módszerek közé tartozó legmeredekebb csökkenés[101,102], konjugált gradiens[103–105], az E_{teljes} második deriváltját is felhasználó Newton-Raphson[106] vagy a kvázi-Newton[107] eljárások egy, a kiindulási ponthoz közeli minimum megtalálására alkalmasak. Ezzel szemben a globális keresők tágabb értelmezési tartományon dolgoznak, a minimumvölgyek között energiagátakon is átjuthatnak és alapvetően sztochasztikus vagy determinisztikus stratégiákat követnek. A sztochasztikus stratégiák alapvetően (pszeudo) véletlenszámok generálására[108] épülnek és eredetileg Neumann János nevéhez fűződnek, aki a monakói kaszinóvárosról Monte-Carlo módszereknek nevezte el ezeket. A véletlenszám-generátorok többnyire egyenletes vagy normális eloszlású számokat állítanak elő az egyes programozási nyelvek függvényeibe vagy különálló programokba implementálva[109]. A véletlenszámok felhasználást nyerhetnek a determinisztikus eljárásokat megelőző kiindulási molekulageometriák vagy sebességek előállításakor is, valamint az evolúciós alapú, genetikus algoritmusokban[110,111] molekula-sokaságok mintáinak (a generációknak) a legyártásakor. A gyógyszertervezésben a Monte-Carlo módszerek és a genetikus algoritmusok főként a gyors molekuláris dokkoló eljárásokban hódítottak teret[112–115], amelyek rendszerint egy pontozó (scoring) függvényen keresnek minimumokat. E pontozó függvények sokszor az E_{teljes} függvény (**1. egyenlet**) egyszerűsített és/vagy kiegészített verziói (lásd még **3. fejezet**). Munkánk során igen gyakran alkalmaztuk a globális keresők másik nagy csoportjába tartozó determinisztikus eljárást, a molekuláris dinamikát (MD)[116–120] is. Amíg a Monte-Carlo szimulációk során az atomok időbeli mozgását nem tudjuk nyomonkövetni, az MD a newtoni mechanika törvényei szerint a rendszer teljes trajektóriáját szolgáltatja számunkra. A globális keresések hatékonysága tovább növelhető a konformáció-sokaságokat hatékonyan mintázó eljárásokkal[121] mint például a szimulált anelláció[122], vagy a replika kicserélődés[123].

2.2.2. Vízmodellek

Ahogy a **2.1 szakasz**ban is láttuk, a hidrátszerkezet kimérése ma sem triviális feladat és időben kivitelezhetetlen lenne elméleti, informatikai segítség nélkül. Ráadásul a pontos gyógyszertervezés során szükségünk lenne mind a mobilisabb tömbfázisbeli (*bulk*), mind a határfelületi, statikusabb hidrátszerkezetek vízkölcsönhatásainak modelljeire. M. Levitt már az 1983-as tanulmányában[118] is részletesen kitér a fehérje-számítások során alkalmazott vízmodellek fontosságára és az *in vacuo* számítások korlátaira. Megemlíti azt is, hogy a víznek, mint közegnek merőben más hatása van az apoláris és a poláris molekularészletek[124] kölcsönhatásaiban. Utóbbi esetben a nagy dipólusmomentummal rendelkező vízmolekuláknak az elektrosztatikus kölcsönhatásokra kifejtett árnyékoló hatása érvényesül. E hatást legegyszerűbben a távolságtól (R_{ij}, **1. egyenlet**) függő relatív permittivitás (dielektromos állandó) függvényekkel[125–129] lehet figyelembe venni, az E_{teljes} függvényben a Coulomb-tagban az ε helyén alkalmazva ezeket.

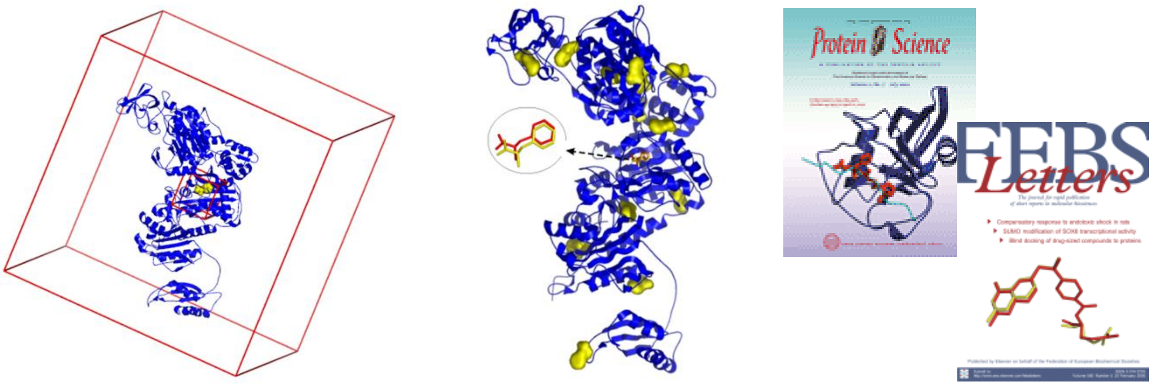
A kontinuum oldószermodellek között az előbbieknél szofisztikáltabb formalizmussal rendelkeznek az általánosított Born[130–133]- vagy a Poisson-Boltzmann[134–136]-egyenletekre épülő implicit[137] vízmodellek. Ezek a szolvatációs szabadentalpiaváltozás elektrosztatikus komponensének számítására alapoznak. Az implicit vízmodelleknél – ahogy az elnevezések is sugallják – tehát a vízmolekulák atomi szinten nincsenek reprezentálva a számítás során, hanem hatásuk együttesen, kontinuumként számolódik el. A vízmolekulák atomi reprezentációjának hiánya természetszerűleg az implicit modellek használhatóságát jócskán korlátozza[138] főleg olyan kölcsönhatásoknál, amikor az oldószer főszerepet játszik. Ilyenek például a gyakran előforduló hálózatos rendszerek, amelyeket a vízmolekulák egymással és az oldott anyaggal[139] alkotnak, vagy a hidrofób effektus[140–144] alapját képező klatrát (kalitka) szerkezetek változásainak leírása az apoláris csoportok körül. Az explicit vízmodellek esetében e kölcsönhatások modellezésére esély kínálkozik, mivel ezek a vízmolekula valós geometriáját és töltésviszonyait atomi szinten reprezentálják. A legnépszerűbb explicit vízmodellek többnyire merev vízmolekulát alkalmaznak, ponttöltésekkel és páronkénti, additív kölcsönhatásokkal[145] számolnak. Ilyenek a *Transferable Intermolecular Potential* (TIP)[146] és a *Simple Point Charge* (SPC)[147] családok legtöbb tagja. A vízmodellek területe messze nem lezárt, amit jól jeleznek a flexibilitás és polarizálhatóság irányába[138] tett fejlesztések, valamint a megjelent összehasonlító tanulmányok[148,149] konklúziói.

2.3. Eredmények

A továbbiakban az elmúlt évek saját eredményeit tárgyalom a szerkezet témakörben megjelent közleményeim alapján, amelyeket az egyéb irodalomtól eltérően szögletes zárójellel hivatkozok meg a szövegben és a dolgozat elején külön jegyzékben soroltam fel.

2.3.1. A ligandumok kötési módjainak feltérképezése [D1-D7]

A gyógyszertervezés központi eleme a ligandumok hatásának (farmakodinámia) előrejelzése. A ligandumok rendszerint egy vagy több célponthoz kötődve fejtik ki hatásukat. E célpontok legtöbbször fehérjék, kisebb hányadban nukleinsavak vagy maga a sejtmembrán. A kötődéskor a ligandum felvesz egy adott térbeli pozíciót, orientációt és konformációt, amelyeket együttesen kötési módnak nevezhetünk. A ligandumok hatásának előrejelzésekor a kötési módok kiszámítása kulcsfontosságú. Amennyiben a kötőzseb ismert, a kötési módot fókuszált molekuláris dokkolással állíthatjuk elő, amelynek során a számításba a kötőzseb környékét vonjuk csak be (**1. ábra**).



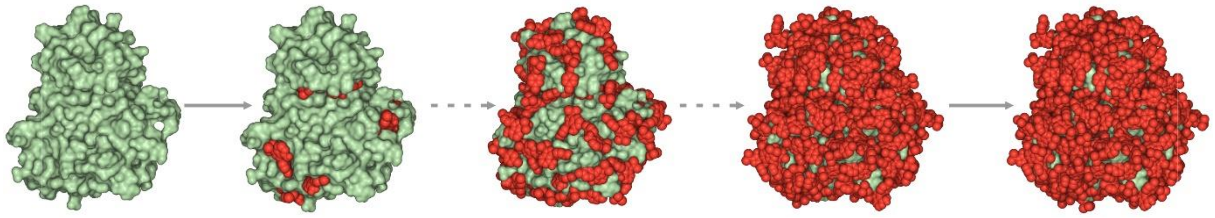
1. ábra A fenilalanil-tRNS-szintetáz (kék szalagok, balra és középen) célponton a fókuszált dokkoláskor a keresés a kötőzseb körüli szűk térrészre (kis piros doboz, balra) korlátozódik, míg a *blind docking* (BD) során a teljes célmolekula felszínére (nagy piros doboz) kiterjed. A BD rendszerint több kötési módot eredményez egy adott ligandumra a célpont teljes felszínén (az L- fenilalanin ligandum [D1] tanulmányban megtalált kötési módjai sárga felszínnel jelölve, középen), így lehetőség nyílik az allosztérikus kötési módok feltérképezésére is. A dokkolással megtalált, a pontozó függvény által az élre sorolt (sárga vonalas ábrázolás, kiemelve, középen) és a kísérletileg kimért (piros vonalak) ortosztérikus kötési módok kiváló egyezést mutatnak. A korábbi[150] és a jelen disszertációban tárgyalt [D1] tanulmányunkat is címlapra emelték (jobbra).

Ez történik például a legtöbb nagy áteresztőképességű szűrő (high throughput screening, HTS) projektben. Korábbi munkánk során a keresést a teljes fehérje célpont felszínére kiterjesztettük, és peptid ligandumokon teszteltük e megközelítésünket, amelyet *blind docking* (BD)[150] névre kereszteltünk. A BD elnevezése abból adódott, hogy a keresés vakon indul el a célmolekula körüli térben, a molekuláris dokkolást nem korlátozzuk a lehetséges kötőzsebre (**1. ábra**), mivel annak helyét nem ismerjük. E tanulmányban[150] vizsgáltuk meg a BD lehetőségét szisztematikusan[151,152] kis peptid ligandumokon, majd alkalmaztuk a megközelítésünket az Alzheimer-kór fontos célpontjára, a β-amiloidra[153,154]. A tanulmányt[150] többek között M. Parrinello csoportja is hivatkozta[155]. A mai napig többen alkalmaznak és fejlesztenek[156–160] BD eljárásokat a korábbi és alább ismertetendő munkáinkra építve. Tekintettel a BD iránti érdeklődésre és a probléma megoldatlanságára, a PhD értekezésemben megkezdett munkát[150,153,154] tovább folytattuk és ennek eredményeit a [D1-D5] közleményekben publikáltuk. Először azt vizsgáltuk meg, milyen célmolekula méretig releváns egyáltalán BD-t alkalmazni. Az első, idevágó [D1] tanulmányban viszonylag kisebb méretű, kompaktabb (gyógyszer jellegű) ligandumokkal dolgoztunk, így a kereső számára a kihívást főleg a célmolekula mérete jelentette. A munkát az AutoDock 3[161] nevű gyors molekuláris dokkoló eljárással végeztük, amely a globális kereséshez (**2.2.1. fejezet**) genetikus algoritmust alkalmaz. Az AutoDock szemi-sztochasztikus keresője és igen egyszerű pontozó függvénye ellenére az eredményeink azt mutatták, hogy kisebb, gyógyszer-méretű ligandumok esetében akár ezer aminosavas fehérjéken is érdemes elvégezni a keresést a teljes célmolekulán. Apo fehérjéket is vizsgálva azt találtuk, hogy kis mértékű indukált illeszkedés mellett a BD még elfogadható eredményeket ad (ez a kis molekulás gyógyszereknél gyakran teljesül). Kimutattuk, hogy a BD egyszerre több kötőhely feltérképezésére is alkalmas, különösen, ha egymás után több körben alkalmazzuk úgy, hogy a már megtalált ligandum kötési módot a célmolekula részeként kezeljük a következő körben, akkumulatív módon. Ez az eredmény adta a későbbiekben az alapötletet a BD probléma szisztematikus megoldásához a [D4] tanulmányban (lásd részletesen később). Az említett több kötőhelyes szituáció számos esetben előfordul gyógyszertanilag fontos receptoroknál[162], így ezek feltérképezése alapvető. A fent részletezett [D1] tanulmányunk megkapta a *Febs Letters* folyóirat címlapját (**1. ábra**), hasonlóan a korábbi alapcikkhez[150].

A BD alkalmazásai során időközben több lehetőség és korlát felmerült, amelyeket egy újabb *Protein Science* tanulmányban elemeztünk ki [D2]. Itt főként olyan rendszereket vizsgáltunk, amelyeken a BD korábban sikertelennek bizonyult. Az AutoDock4[163,164] mellett az EADock[165,166] programot, valamint kötőzseb-kereső (pocket search, PS) módszereket is vizsgáltunk. Utóbbiak között voltak egyszerűsített erőtér alapján (Sitehound[167], Q-SiteFinder[168]), valamint pusztán geometriai elven (Pocket-Finder[169], Pass[170]) működők is. Amíg az AutoDock4 előzetes PS nélkül használható BD céljára, az EADock egy Ligsite-os[171] PS után fókuszált dokkolást végez, tehát két különböző BD stratégiát is vizsgáltunk. Érdekes módon azt kaptuk, hogy egyes PS módszerek használatának létjogosultsága van a precízebb BD mellett is. Ez abból adódik, hogy nem függenek az adott ligandumtól, hanem a kötőzsebet rendszerint próba atomokkal töltik meg és ezekre együttesen számolnak (például a Sitehound) egy teljes kölcsönhatási energia értéket, amely több ligandum(konformáció) kötési lehetőségét fejezi ki az adott zsebhez egy számértékkel. A tanulmányban emellett a másodlagos ligandumok és kötőhelyek, valamint a szerkezeti vízmolekulák szerepét is vizsgáltuk a BD során. Utóbbiak (az explicit vízmodell) hiánya alapvető korlátja a gyors BD eljárásoknak. Mindkét faktor esetében elmondható, hogy bár technikailag (látszólag) problémát jelentenek a BD probléma megoldása szempontjából, ugyanakkor vizsgálatuk a ligandum kötési módjáról és a kötőzsebek funkciójáról hasznos információkat szolgáltat (lásd még a **2.3.3 fejezetet**). A tanulmányban [D2] végül arra jutottunk, hogy egy ortoszterikus kötőhely azonosítását nagy biztonsággal jelzi, ha azt az első osztályban (*Rank 1*) legalább 2 különböző BD eljárással vagy egy BD és egy PS eljárással megtaláltuk (konszenzusos találatok). A PS eljárások ilyen módon igen jól használhatók a BD eredmények megerősítésére.

A BD megközelítésünket mi is alkalmaztuk [D3] a szex szteroidok kötési hálózatainak feltérképezésre a humán ösztrogén receptor α-n. Itt a klasszikus kötőzseb mellett a BD azonosította az alternatív kötőzsebet a vizsgált szteroidok (17-β-ösztradiol és egy ösztrén) esetében. Az alternatív kötőzseb létezését korábbi tanulmányok[172–174] is vizsgálták. A mi tanulmányunkban azonban úgy azonosítottuk a klasszikus és alternatív kötőzsebeket az első két legerősebb kötődést mutató osztályban, hogy a teljes ösztrogén receptoron végeztük el a keresést a BD megközelítésnek megfelelően, tehát minden előzetes információ, irányítás/befolyásolás nélkül. Ezen túlmenően kimutattuk, hogy a klasszikus kötőzsebbe történő ligandumkötődés, valamint a koaktivátor peptid hiánya elősegítik a szteroidok alternatív kötőzsebbe történő kötődését. A ligandumok disszociációjának követése explicit vízmodelles MD számításokkal lehetővé tette a kötési dinamika részletes leírását beleértve az egyes zsebeknél képződő komplexek kinetikus stabilitásának kvantitatív összehasonlítását, valamint kötőzsebeknél megfigyelhető csapóajtó-mechanizmus atomi szintű magyarázatát.

A BD megközelítéssel foglalkozó tanulmányaink számos kutatócsoport érdeklődését felkeltették[175–210] és az ortoszterikus kötőhely-keresésen túl a BD alkalmazást nyert alloszterikus[211,212] (másodlagos, többszörös)[213–217] kötőhelyek esetében is. Ezen a ponton felmerült a kérdés, hogy lehetne-e egy olyan eljárást kidolgozni, amely a gyors dokkoló módszerek korlátait (a globális kereső algoritmusok tökéletlensége, az explicit vízmodell hiánya, a célpont merevsége) meghaladva egy végső megoldást nyújtana a BD problémára az összes lehetséges kötési mód megtalálásával. A kérdés megválaszolására szisztematikus módszert dolgoztunk ki [D4], amelyet Wrap ‘n’ Shake névre kereszteltünk. Ahogy a módszer neve is jelzi, alapvetően a csomagoló (Wrapper) és a rázó (Shaker) fázisokból tevődik össze. A legnagyobb kihívás a csomagoló algoritmus kidolgozása volt. Ennek során ugyanis a célpont teljes felszínét a ligandumnak egy monomolekuláris rétegével vonjuk be minél hézagmentesebben, több BD ciklus során (**2. ábra**).



2. ábra A Wrap ‘n’ Shake módszer csomagoló algoritmusa a célpontot (zöld) a ligandum (piros) másolataival több ciklusban, monomolekulás, hézagmentes réteggel vonja be.

Ahogy a fentiekben utaltam rá, már a korábbi [D1] tanulmányunkban megmutatkozott a BD több ciklusban történő alkalmazásának haszna, amelynek során a már dokkolt ligandum molekulákat a célmolekula részeként kezeljük. Gyorsan kiderült azonban, hogy a ligandum másolatainak korrekt elhelyezése egy monomolekuláris rétegbe nem triviális, olyan módon, hogy azok ugyan a célmolekulával kölcsönhassanak de egymással ne képezzenek aggregátumokat a célmolekula felszínén. A cél érdekében – egy sor próbálkozást követően – egy új atomtípust definiáltunk azoknak a célmolekula atomoknak, amelyek a már dokkolt ligandumok körül helyezkednek el. Ezeknél az atomoknál az elektrosztatikus kölcsönhatást kikapcsoltuk a töltések nullázásával (**1. egyenlet**, második sor) és a Lennard-Jones kölcsönhatást pedig gyenge taszítóra kalibráltuk, amellyel a kívánt (hézagmentes) monomolekuláris, N ligandumból álló rétegbe végül sikerült a célmolekulákat becsomagolni. Az így előállt célmolekula-ligandum$_N$ komplexről aztán explicit vizes MD számításokban és szűrő lépésekben ráztuk le a gyengén kötött ligandum-másolatokat és a megmaradt kötési módokat listáztuk. E monomolekuláris beterítés és a disszociatív MD szűrési lépések végül egy jól reprodukálható módszert eredményeztek, amely tekinthető a BD probléma szisztematikus megoldásának gyógyszer méretű ligandumok esetében. A Wrap ‘n’ Shake módszerünk jó visszhangot[218–221] kapott. Technikai részleteit és használati útmutatását a népszerű Methods is Molecular Biology sorozatban részletesen publikáltuk [D5].

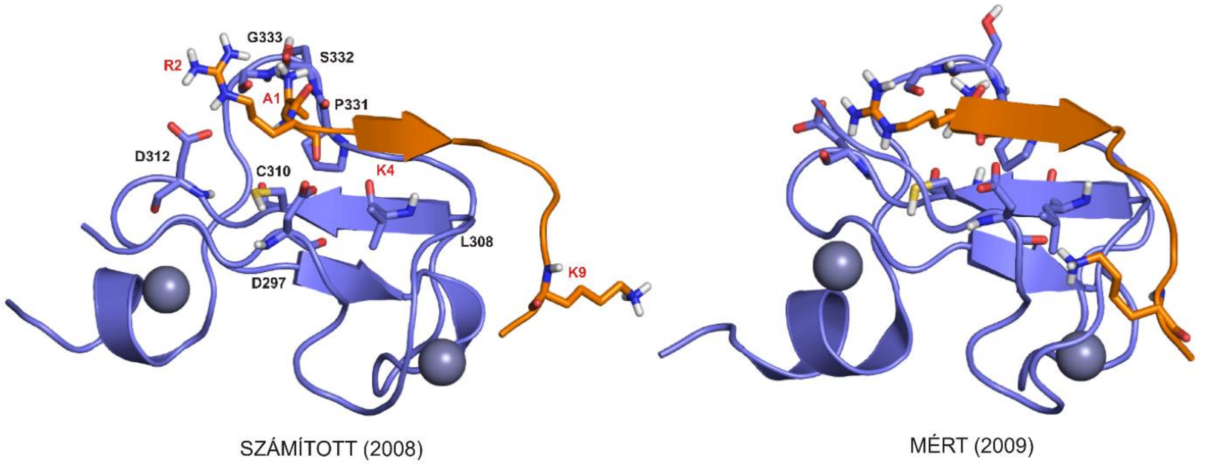
A ligandumok ortoszterikus kötőhelyhez történő vándorlása során gyakran több prerekvizit (előfeltételi) kötőhelyhez is aszociálódnak, amelyek rendszerint kisebb kötési erősséget nyújtanak a végső kötőhelynél. E prerekvizit kötési módok jórészt tranziensek és így kísérletes technikákkal nehezen kimérhetők. Ugyanakkor kétségtelenül fontosak, hiszen a ligandum horgonyzását biztosítják és így a kötődés szükséges állomásai. Tanulmányainkban [D6, D7] megmutattuk, hogy a kísérletes technikákat igen jól kiegészítik a számítások e prerekvizit kötési módok feltérképezésében. A kovalens kötésmódú ligandumoknál a kovalens kötés kialakulása előtt nyilvánvalóan szükség van egy másodlagos kötésekkel kialakuló kötési módra, amely a ligandum robbanófejét (warhead) a célmolekula reaktív csoportja felé írányítja. Ez a szituáció áll elő a tranziens receptor potenciál ankyrin 1 (TRPA1) nevű polimodális nociszenzor esetében is[222–224], amelyet kovalensen kötő agonisták[225,226] aktiválnak. A prerekvizit kötőhelyek feltérképezése után megmutattuk [D6], hogy a kovalens agonisták esetében az ezekhez történő kötődés módja előrejelzi a végső, kovalens kötési mód kialakulását. Megállapítottuk, hogy a prerekvizit kötési módok az agonisták asszociációs/disszociációs mechanizmusainak fontos mérföldkövei, például az A-hurok régió felnyílásának szabályozásán keresztül. Ezen túlmenően a mechanizmus-alapú tervezéshez is fontos információt szolgáltatnak, új célpont aminosavakat azonosítanak az ortoszterikus kötőhely aminosavain túl. A TRP receptorok egyébként a fájdalomérzet molekuláris mechanizmusának központi szereplői és vizsgálatuk az utóbbi évtizedekben előtérbe került, amelyet a 2021-es Nobel-díj[227] is fémjelez.

A prerekvizit kötőhelyek a mechanizmus-alapú gyógyszertervezés fontos elemei, amennyiben egymással és az ortoszterikus kötőhellyel fennálló kapcsolataikat jellemezni tudjuk. E célból kidolgoztuk [D7] a NetBinder nevű eljárást, amellyel nem csupán feltérképezzük, de osztályozzuk és hálózatba is kapcsoljuk a prerekvizit kötőhelyeket. Az eljárás tesztelésére a miozin II enzimet választottuk, amelynek hosszú, belső kötőcsatornájában a blebbisztatin nevű inhibitor vándorlása során több prerekvizit kötőhelyen megtapad és végül a csatorna mélyén lévő végső kötőzsebekhez jut el[217,228]. A fent részletezett Wrap ‘n’ Shake módszer segítségével a kötőhelyeket azonosítottuk, majd kötési hálózatot, ebből pedig kötési útvonalakat állítottunk elő, teljes képet kapva a blebbisztatin kötési mechanizmusáról. Az eljárás segítségével a teljes kötési mechanizmust feltártuk, a fehérje felszínéről elindulva egészen a végső (ortoszterikus) kötőzsebig. Az MD alkalmazása lehetővé tette a fehérje konformációváltozások precíz követését is. Kijelöltük a kulcsfontosságú prerekvizit kötőhelyeket is, amelyek a későbbi tervezésben mint új célpont használhatók fel, ezzel is növelve az inhibitortervezés lehetőségeit a miozin II-höz hasonlóan hosszú, belső üreggel rendelkező célpontokon (ioncsatornák, egyes receptorok).

2.3.2. Fehérje-peptid komplexek számítása [D8-D13]

A fehérje-peptid komplexek szerkezetének kiszámítása a mai napig nem triviális feladat. A peptidek a kisméretű gyógyszermolekulákhoz képest sokkal kezelhetetlenebb ligandumok, a méretük, nagy flexibilitásuk (torziós szabadsági fokuk) és kiterjedt hidratáltságuk miatt. A kisebb peptidek esetében a gyors dokkoló módszerek sikeresek lehetnek[150], ám a tetrapeptidektől nagyobb ligandumoknál gyakran nem eredményeznek pontos szerkezeteket[229–231]. Amíg a fizikai-kémiai alapokon működő dokkoló eljárások korlátait az erőtér és a kereső eljárások tökéletlensége okozza (lásd előző fejezetek), addig az ismeret alapú (*knowledge-based*) módszereknél a tanuló szettek mérete és összetétele szabják meg a teljesítőképességük természetes határát. A hisztonok fehérjékkel képzett komplexei tipikus példák a fent említett nehézségekre. A hisztonok fontosságát az adja, hogy kulcsszerepet játszanak az epigenetikai regulációban azáltal, hogy a kromatinban a DNS-sel, valamint különféle fehérje partnerekkel is kölcsönhatnak és poszt-transzlációs módosulásaik (PTM) definiálják az ún. „hiszton kódot”[232,233], amelynek a genetikai kódhoz mérhető a jelentősége számos betegség pathomechanizmusában[234,235] – erről bővebben összefoglaló tanulmányunkban [D8] írtunk.

A hisztonok terminális peptid szakaszai a nukleoszómák felszínéről antennaként kinyúlva[236] hatnak kölcsön a kódot olvasó/író fehérjékkel. E terminális peptid szakaszok lineáris, rendezetlen szerkezetűek, így a komplexeik kimérése elég nagy kihívást jelent és sok esetben csak a láncvégi aminosavakra szorítkozik. Ugyanakkor a hiszton kód megfejtéséhez és a kapcsolódó gyógyszertervezésekhez számos ilyen komplex szerkezet előállításra lenne szükség tekintettel a nagy számú PTM variációra (többféle módosulás, több aminosav pozícióban). Ezek az igények inspirálták munkánkat e területen, amelynek során többféle megközelítéssel állítottunk elő ilyen problémás fehérje-peptid komplex szerkezeteket. Peterson professzor Tartuban (Észtország) keresett meg azzal az eredménnyel, hogy az autoimmun regulátor (AIRE) transzkripciós aktivátor fehérje a kísérleteik szerint a H3-as hisztonhoz µM-os affinitással kötődik, egész pontosan annak H3K4me0, azaz a K4 helyen metilálatlan formájához. Ugyanakkor kimutatták azt is, hogy a metil csoportok számának növelésével a kötéserősség fokozatosan elenyészik és a H3K4me3 esetben gyakorlatilag már nem mérhető. Tapasztalataikat atomi szintű komplex szerkezettel nem tudták alátámasztani, az NMR mérésekkel csupán a kötésben részt vevő aminosavakra tudtak rámutatni amelyek az AIRE PHD ujjára estek. A meglévő szerkezetek kombinálásával, hasonlósági alapon és MM/MD szinten modellezésssel előállítottam [D9] a hiszton H3 peptid AIRE PHD ujjal alkotott szerkezetét atomi felbontásban, amelyen már jól látszott a kötődés módja. Megállapítottam, hogy a hiszton H3 peptid gerince az AIRE PHD ujjban meglévő, két sávból álló antiparallel β-redőzött réteg mellé orientálódva egy harmadik, antiparallel lefutású sávot képezve kapcsolódik. A peptidgerincek közti H-hidakon túlmenően a két molekula kölcsönhatását (**3. ábra**) az egyes oldalláncok közti sóhidak és hidrofób effektusok tovább erősítik.



3. ábra Az autoimmun regulátor (AIRE, kék) fehérjének a H3-as hiszton N-terminális peptidjével (narancs) alkotott komplexének számított és mért atomi felbontású szerkezetei

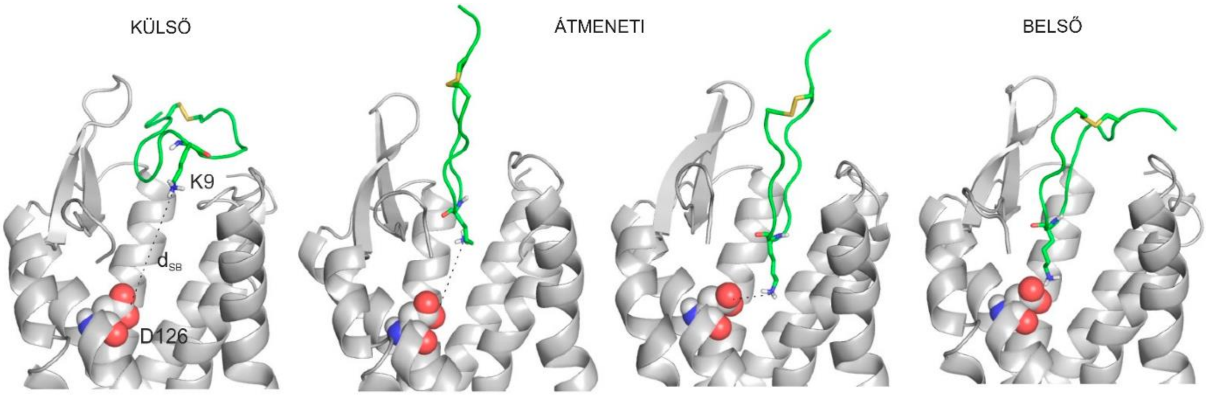
A számított szerkezetet NMR-es és irányított mutagenezises kísérletek már ebben [D9] a közleményben megerősítették, majd az NMR-es mérések további finomításával a következő évben a kísérletesen kimért, atomi felbontású szerkezet is közlésre került[237], amely kiváló egyezést mutatott az egy évvel korábban közölt [D9] számított szerkezettel (**3. ábra**). A H3K4me0 komplex szerkezet előállításán túl molekuláris dinamikai számításokkal sikerült azt is megmutatni [D9], hogy a K4-metilációk során a metil csoportok térigénye miatt nem tud kialakulni a kötés az AIRE PHD ujj kötőzsebével, a H3K4me3 esetében már fizikailag nem fér be a kisebb méretű PHD kötőzsebbe a K4me3 oldallánc.

Szintén modellezéssel, a hiszton H4 peptid fragmensek *in situ*, a kötési felszínen történt összeépítésével állítottuk elő [D10] a hiszton H4 protein arginin metiltranszferáz 5 (PRMT5) és metiloszóma protein 50 (MEP50) partnerekkel alkotott komplexének atomi felbontású szerkezetét. A PRMT5 értelemszerűen író funkciót tölt be a fent említett hiszton kód kialakításában és főként a hiszton lineáris, N-terminális szakaszán lévő aminosavakra helyezi el az oldallánc-módosító metil csoportokat. A tanulmány célja az volt, hogy magyarázatot adjunk a PRMT5 T80 helyen történt foszforilációja során kísérletileg tapasztalt megnövekedett metiltranszferáz aktivitásra, valamint arra, hogy a PRMT5 miért csak a hiszton H4 szabad (nukleoszómához nem kötött) formáját metilálja. A komplex konstrukciója mellett a vad típusú és foszforilált szerkezetek MD számításait is elvégeztük és a célkitűzésben megfogalmazott kérdésekre ezek segítségével adtunk választ. Az PRMT5-MEP50-H4 komplex megépített szerkezetéből jól látszik [D10], hogy az említett PRMT5:T80-as aminosav a hiszton oldalon a H4:R45 aminosavval és környezetével kerül kapcsolatba. E kölcsönhatás az aktív centrumtól távol (a H4 átellenes oldalán) alakul ki és a foszforilált T80 esetében sóhíd formájában tovább erősödik, stabilizálva az enzim-szubsztrát komplexet és növelve a metiltranszferáz aktivitást. A hiszton H4 nukleoszómában kimért szerkezetéből megállapítottuk, hogy az említett H4:R45 aminosav részt vesz a hiszton DNS-hez történő horgonyzásában, elsősorban a DNS foszfát csoportjaival kialakított sóhidakon keresztül, nagyon hasonlóan a foszforilált PRMT5:T80 fent említett esetéhez. Értelemszerűen így a nukleoszómában kötött H4 esetében pont a PRMT5-höz történő kötődésben résztvevő H4:R45 aminosav kerül blokkolásra, amelyhez így a PRMT5 nem tud hozzáférni és ez a szerkezeti magyarázata annak, hogy a H4 szabaddá válása szükséges a metiláláshoz.

A fentiekben elemzett két tanulmányunkban [D9, D10] sikeresen tudtuk kombinálni a meglévő kísérletileg kimért, homológ és/vagy töredékes szerkezeteket új hiszton komplexek előállítására. Kimért szerkezetek azonban a nagy számú lehetséges PTM miatt nem mindig állnak rendelkezésre és sokszor csak a célmolekula (író/olvasó fehérje) hozzáférhető a Fehérje Adatbankban, valamint a kötődés tényét tudják megállapítani kísérletileg, például kötődési esszék segítségével. Ilyenkor a hiszton peptid ligandumot elméleti úton lehet a célmolekulához illeszteni. A hisztonok kötésben részt vevő, terminális peptid szakasza a bevezető sorokban említettek miatt különösen problematikus ligandumok a dokkolás szempontjából. Mivel a tapasztalatok azt mutatják, hogy a „klasszikus" gyors dokkoló eljárásokkal többnyire a tetrapeptid mérettartományig lehet elfogadhatóan dokkolni a peptid ligandumokat, így a probléma megoldására a terminális hiszton peptidet kisebb méretű fragmensekre vágtuk a fragmens blind docking (FBD) eljárást közlő tanulmányunkban [D11]. Itt a célmolekulák teljes felszínét beterítettük a fenti Wrap 'n' Shake módszerben kifejlesztett módon di-, tri-, és tetrapeptidekkel. Azt tapasztaltuk, hogy a legpontosabban a dipeptid fragmensek dokkolódtak, így az FBD következő szakaszában ezeket kapcsoltuk, majd hegesztettük össze a kovalens kapcsolódási pontokon. A fragmensek kapcsolása nem triviális ez esetben, hiszen a célmolekula teljes felszínét beterítettük mindkét összekapcsolandó (a hiszton szekvenciájában egymást követő) dipeptid fragmenssel. Egy automatizált algoritmussal a dipeptid fragmensekből párokat, triádokat és tetrádokat képeztünk és ezeket rangsoroltuk. Ezt követően a kialakítandó peptidgerinc szomszédos torziós szögei mentén történő szisztematikus keresés után hegesztettük össze a dipeptid darabokat az amid kötés mentén, majd a komplex szerkezetet MM szintű optimalizációval finomítottuk.

A fragmens alapú dokkoló eljárásoknak, így az FBD-nek is a legkényesebb része a fragmensek kovalens összekapcsolása. Ez fokozottan igaz, ha a célpont nagy felszínén vagy – mint az FBD esetében – a teljes célponton történik a keresés és ha a célpont-ligandum kölcsönhatás nem túl erős, mint a hiszton ligandumok esetében, ahol a K_d legtöbbször a µM-os tartományba esik. Mivel a dipeptid hiszton fragmensek kötési módját az AutoDock 4.2.6 jól és gyorsan megtalálja [D11], így e fragmensekre építettük a PepGrow [D12] nevű protokollunkat, amely az említett, problematikus összekapcsoló lépés kihagyásával dolgozik. A PepGrow a bedokkolt hiszton H3 dipeptid fragmenseket horgonyként kezeli, amellyel a hiszton ligandum a célmolekulához kapcsolódik. E horgonyzó dipeptidekből (magokból) „növeszti meg" a teljes hiszton ligandumot *in situ* a célmolekula kötőzsebében a homológia modellezésben alkalmazott gyors fehérjelánc építő program segítségével, több száz kötési módot generálva le rövid idő alatt egy-egy dipeptid magból kiindulva. Ezt követően az összes legenerált komplex szerkezetre kiszámításra kerül a célmolekula-ligandum kölcsönhatási energia, amely az alapját képezi a reprezentáns kötési mód kiválasztásának. Ilyen módon tehát a PepGrow kiküszöböli a problematikus kapcsolási lépést a fragmens dokkolás során, helyette a „fragmensből növesztés" stratégiáját követve. Protokollunkat tíz másik módszerrel összehasonlítva a vizsgált hiszton komplexek esetében a legjobb eredményeket szolgáltatta. Ugyanakkor a szisztematikus összehasonlításainkból az is kiderül, hogy a jelenleg alkalmazott gyors dokkoló eljárások mindegyike igen rossz a megkapott kötési módok rangsorolásában. Ez leginkább a pontozó függvényeik (**2.2.1. fejezet**) korlátaira vezethető vissza, amit súlyosbít az explicit vízmodell (**2.2.2. fejezet**) hiánya, valamint a célmolekula flexibilitásának korlátozott figyelembe vétele is.

Az említett hiányosságok részben feloldhatók, ha a PepGrow által előállított célpont-ligandum komplex szerkezetet MD segítségével finomítjuk tovább. Ezt a megközelítést alkalmaztuk a szomatosztatin nevű endogén peptid 4-es altípusú receptorával (SSTR4) képzett komplexének előállításakor [D13] is. Ez esetben sem a receptor, sem a komplex szerkezete nem állt rendelkezésre a Fehérje Adatbankban. A Wrap 'n' Shake beterítő eljárását alkalmazva, a 14 aminosav hosszúságú szomatosztatin peptid ligandum csúcsi részén lévő FWKT tetrapeptidet tartalmazó fragmenssel térképeztük fel a kötőhelyeket az SSTR4 teljes felszínén. Magának az SSTR4 célpontnak a szerkezetét is homológiamodellezéssel állítottuk elő. A legjobb kölcsönhatási energiát mutató dokkolt fragmensből indítottuk el a PepGrow protokollban alkalmazott növesztési lépést és jutottunk el a szomatosztatin molekula külső (prerekvizit) kötési módjához (**4. ábra**).



4. ábra A szomatosztatin peptid (zöld) 4-es altípusú receptorához (SSTR4, szürke) történő kötődési folyamatának főbb állomásai. A kötődés során a célpont D126 és a ligandum K9 aminosavai közötti sóhíd fokozatosan kialakul (az ábrán d_{SB}-vel jelzett távolság lecsökken).

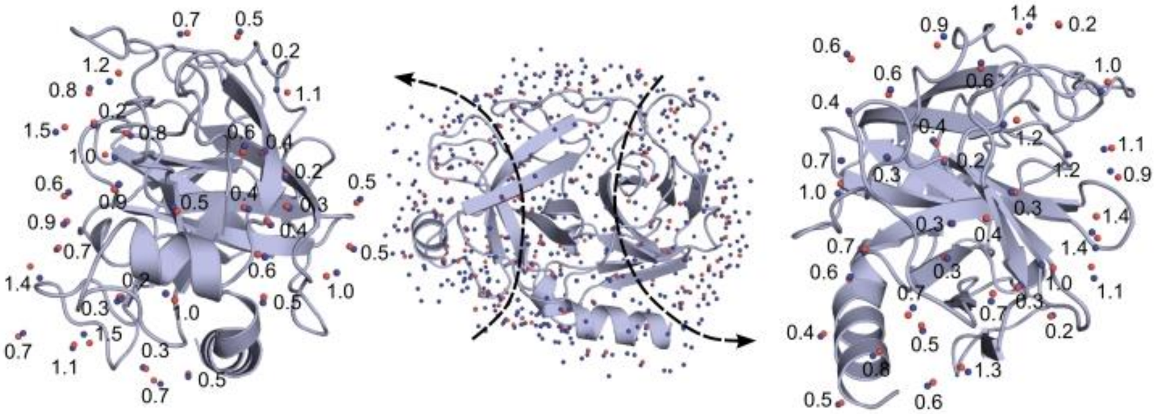
A tetrapeptid komplexből kiindulva ezután MD segítségével feltártuk a belső (ortoszterikus) kötési módot is. Ilyen módon, a PepGrow és az MD kombinálásával a szomatosztatin teljes kötődési mechanizmusát végig tudtuk kísérni az SSTR4 célpontra. Azon túl, hogy e munkánk közölte [D13] az első SSTR4-szomatosztatin szerkezetet, e megközelítésünknek további előnye – a **2.3.1 szakasz**ban leírtakhoz hasonlóan – hogy az ortoszterikus kötőhelyen túl feltárja a prerekvizit kötőzsebeket is és lehetővé teszi a tervezést e célterületekre is. A kísérletileg elérhető szerkezetek, mint például az SSTR2 altípusnál[238] csak az ortoszterikus kötési módot írják le.

2.3.3. A hidrátszerkezet számítása célpont-ligandum kapcsolatokban [D14-D19]

A célpont-ligandum kölcsönhatásokat a víz nagymértékben képes befolyásolni, mint közeg és mint kölcsönható partner is. E szerepkörök fizikai-kémiai alapjait a felvezető, **2.2 fejezet**ben már érintettem és utaltam rá, hogy a víz viselkedését leíró oldószermodellek a mai napig nagy fejlődésen mennek keresztül. Ennek megfelelően a hidrátszerkezet (változásainak) számítása a célpont-alapú gyógyszertervezés módszertanának egy lezáratlan – és ezért izgalmas – fejezetét képezi. Ez részben a hidrátszerkezet kísérleti meghatározásának korlátaiból (**2.1 fejezet**), részben a vízmolekula sokarcúságából adódik. A víz egyfajta molekuláris jolly jokerként egyszerre rendelkezik nagyfokú mobilitással és – ha a körülmények úgy hozzák – a maximálisan kialakítható négy H-kötése segítségével kulcsfontosságú összekapcsoló, hidaló szereppel is. A vízmolekulák „társas viselkedésének" topográfikus problematikáját talán legjobban B. Halle tartózkodási időre vonatkozó megjegyzése[47] mutatja be: *„An unusually long residence time for a hydration water molecule, therefore, does not indicate particularly strong protein–water interactions, but rather a topography that prevents the water molecule from exchanging by a cooperative mechanism. The simplest example of such restrictive topography is a deep pocket on the protein surface."* A fehérjefelszín lokális topográfiája valóban igen fontos tényező a hidrátszerkezet helyi (kinetikus) stabilitásában és így a ligandum kötődésekor felmerülő változásaiban is.

A hidrátszerkezetnek a ligandum kötődésekor felmerülő, a fentiekből következő, aktuális kérdéseit a *Current Opinion in Structural Biology* folyóiratban megjelent [D14] cikkünkben részletesen is tárgyaltuk. A ligandum célmolekulához történő kötődésekor az apo célmolekula felszínén elhelyezkedő vízmolekulák sorsa alapvetően a maradás (konzervált vizek) vagy a távozás lehet aszerint, hogy a ligandum és a célmolekula között hidakat tudnak képezni, vagy pedig a ligandum kilöki őket társaik közé az oldat belsejébe. A ligandum tervezése során tehát mind az apo célmolekula felszínének, mind a komplex (holo) interfésznek a vízszerkezetére szükség van, hogy a ligandum szerkezetét optimalizálni tudjuk a (de)hidratáció szempontjából is. Tekintettel a kísérletes szerkezetmeghatározás korlátaira (**2.1 fejezet**), napjainkra számos elméleti módszer áll rendelkezésre a vízszerkezet számításához. E módszerek durván a statikus és a dinamikus kategóriákba sorolhatók, aszerint, hogy alkalmaznak-e MD-t és/vagy explicit vízmodellt (**2.2 fejezet**) a vízszerkezet kiszámításánál. A módszerek egy aktuális jegyzékét és összehasonlításukat a [D15] cikkben adtuk közre.

Munkánk során egy dinamikus módszer kidolgozását határoztuk el. A MobyWat nevű módszer és program [D16] első verzióját fehérje célpontok felszíni hidrátszerkezetének számításához készítettük el. A módszer explicit vizes MD számításokra épül, amelyeket a Gromacs[239] nevű open source programmal végeztünk el. Itt említeném, hogy a MobyWat bármely más MD programcsomagból származó szerkezeteket fel tud természetesen dolgozni. A Gromacs trajektóriákat a hordozható xdr-kompatibilis xtc fájl formátumban is képes olvasni és egy saját bináris fájltípust is használ a predikció során kialakított vízhalmazok tárolására és kezelésére. Az explicit vízmodelles MD-alapú megközelítés előnye, hogy nem csak a célpont-víz, de a víz-víz kapcsolatokat is számítjuk és így pontosabb vízpozíciókat használhatunk fel a predikcióban. A program többféle klaszterező algoritmust is alkalmaz a prediktált vízpozíciók kinyerésére a trajektóriából. Egyrészt a vízmolekulákat azonosítójuk alapján egyedileg végigkövetve, másrészt csak az adott térbeli pozíció betöltöttségét követve, vagy e kettő kombinálásával is létre tudja hozni az ún. predikciós listát, amely a vízmolekulák oxigén atomjainak koordinátáit tartalmazza. A programhoz egy validációs mód is készült, amelynek segítségével a kísérletesen kimért szerkezet birtokában automatikusan kiszámítható a sikerességi hányad, amely azt mutatja meg, hogy a kísérletileg meghatározott víz oxigén pozíciók hány százalékát sikerült adott pontossággal prediktálni. A MobyWattal végzett validációnk azt mutatták, hogy 20 fehérjemolekula több mint 1500 vízpozíciójának átlagban több mint 80 %-át sikerült pontosan kiszámítanunk (**5. ábra**). Ezen túlmenően megvizsgáltuk adott rendszernél az MD trajektóriák különbözőségéből eredő reprodukálhatóságot is és a módszer robosztusnak bizonyult, ha eltérő kiindulási sebességeloszlásokat alkalmaztunk. A módszer a prediktált vízpozíciókat mobilitásuk alapján rangsorolja, így az adott vízpozíció kinetikus stabilitásáról is számszerű információt ad. Érdekes módon relatíve kevés eljárás van egy teljes fehérjefelszín hidrátszerkezetének kiszámítására. Összevetve más módszerekkel [D15] a MobyWat teljesítménye a legjobbak között van. Általánosan elmondható egyébként, hogy a legtöbb módszer a konzervált vízpozíciókat számítja a legbiztosabban, ami érthető, hiszen e vízmolekulák maradnak a kötőzsebben és képezik majd a hidat a célmolekula és a ligandum között a komplex kialakulása után (kevéssé mobilisak, erősen kötődnek, jó topográfiai beágyazottsággal rendelkeznek).



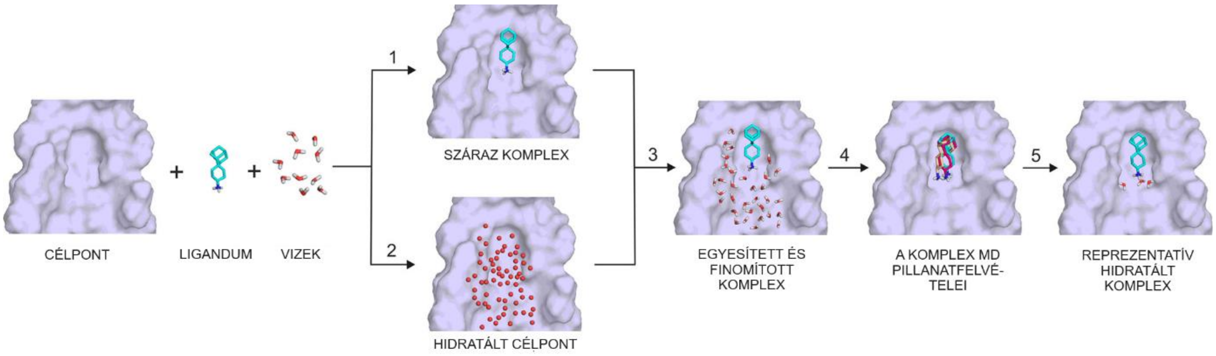
5. ábra A MobyWat által prediktált (kék gömbök jelzik az oxigén atomokat) és RKR-rel mért (piros gömbök) vízpozíciók egyezése a szarvasmarha hasnyálmirigy tripszin (PDB kód: 1s0q) felszínén. A távolságadatok a két szélső, nagyított részleten Å-ben vannak feltüntetve (1 Å = 10^{-10} m). Az egyezések igen jók, többnyire 1 Å alattiak, összességében 84 %-os sikerességi hányadot eredményezve ezen a fehérjén.

A célmolekula-ligandum interfészek hidrátszerkezetét a teljes felszíni hidrátszerkezetnél még nagyobb, átlagosan 90 % feletti sikerrel számítottuk ki a MobyWat alkalmazásával [D17]. Azt tapasztaltuk, hogy az interfész régiók számításában a ligandum jelenléte egyszerre segíti és hátráltatja is az explicit vízmodelles MD trajektóriákra épülő stratégiánkat. Segíti annyiban, hogy a hidaló vízmolekulák helyzete itt már a ligandum oldaláról is stabilizálódik, tehát a kis mobilitásuk alapján könnyű kiválasztani őket a predikciós listára. A hátrány abból fakad, hogy ha a „száraz” célpont-ligandum komplexből indítjuk el a szokásos módon az MD felkészítését (doboz generálása, vizek elhelyezése a dobozba, azaz szolvatáció), akkor előfordul, hogy az interfész régióba a ligandum jelenléte miatt nem helyez el az egyenletes grid alapján szolvatáló algoritmus vízmolekulákat és az így keletketkező, ligandum alatti, zárt kavitásokba az oldatból (a ligandum jelenléte miatt) sem tudnak később, az MD alatt vizek bediffundálni. E problémát észlelve, két lépéses eljárást dolgoztunk ki, amelynek első lépésében a ligandum nélküli célmolekula felszínét borítottuk be vizekkel, majd a ligandumot visszahelyezve egy nagyon szűk toleranciával csupán a kovalens távolságban ütköző vizeket távolítottuk el, és ezt a „megtömött” interfészt egy második MD lépésben relaxáltuk, majd sor kerül a MobyWat-os predikcióra. Ilyen módon kavitások nélküli, teljes mértékben hidratált célpont-ligandum interfészeket tudtunk előállítani, a fent említett nagy sikerességi hányaddal (sok rendszernél elértük a 100 %-ot is). A validációt 31 komplex 344 vízmolekuláján végeztük el, jól reprodukálható, robosztus eredményeket kaptunk itt is.

A teljes hidrátszerkezet előállítása lehetőséget adott arra, hogy a hálózatelmélet alapelemeit adaptáljuk a hidrátszerkezet kölcsönhatási rendszerére [D17]. Felrajzoltuk a vízhálózatok gráfjait, amelyekben definiáltuk a statikus és dinamikus csúcsokat, éleket, valamint alhálózatokat. Kimutattuk a kiterjedt statikus alhálózatok szerepét a célpont-ligandum komplexek stabilizálásában. E gráfok generálását szintén a MobyWat egyik modulja, a NetDraw segítségével végeztük el, amely gyakorlatilag bármekkora komplex esetében elő tudja állítani a gráfokat. A hálózati alapú megközelítésünket sikeresen alkalmaztuk a H3.3 és H4 hisztonok DAXX fehérjével képzett komplexe vad típusú és a H3.3:G90M mutáns közti stabilitáskülönbség magyarázatára és megmutattuk, hogy a mutáns esetében a korábbi statikus alhálózat helyett egy diffúz, dinamikus alhálózat jön létre, amely miatt lecsökken a mutáns verzió stabilitása, a kísérleti eredményekkel összhangban. A vízhálózatok ilyen teljeskörű, kvantitatív, matematikai pontosságú leírása, amellyel e tanulmányunkban [D17] foglalkoztunk, a jövőben lehetőséget nyújt még további hasonló magyarázatokra és predikciókra a biomolekuláris komplexek területén. A MobyWat módszerünket a szakmai közvélemény is elismerőleg befogadta[240–248].

A MobyWat alapú predikcióknak természetszerűleg fontos eleme a MD számítás, amelynek segítségével a klaszterezés alapjául szolgáló trajektóriát előállítjuk. Egy kapcsolódó tanulmányban [D18] szisztematikusan megvizsgáltuk a legfontosabb MD paraméterek hatását a vízszerkezet predikciójának sikerességére mind a felszíni mind az interfész esetekben, több rendszert is bevonva a számításokba. Az RKR adatgyűjtési és a számításnál alkalmazott hőmérséklet mellett a nyomás, az erőtér, az explicit vízmodell típusa, a számított sokaság megválasztása valamint a nehézvíz alkalmazása is vizsgálat tárgyát képezte. Az eredmények azt mutatták, hogy az NVT és NPT sokaságok, valamint a nehézvíz esetében egyformán jól teljesített a MobyWat, és a nyomás sem volt a sikerességre nagy befolyással, ha az a standard nyomásérték (0.1 MPa) körül volt. Az explicit vízmodellek (**2.2.2. fejezet**) kapcsán az eredmények azt mutatták, hogy a TIP3P és afeletti vízmodellek alkalmazása javasolható. Az az AMBER-ről az OPLS erőtérre történő áttérés nem okozott szignifikáns csökkenést a sikerességben. A legnagyobb eltéréseket a hőmérséklet változtatása esetén észleltük. Bár talán „életszerű” lenne a kísérletileg alkalmazott kriogenikus mérési hőmérsékleteken (**2.1. fejezet**) végezni a számításokat, azonban ez nem vezet célra. Ennek lehetséges okait a tanulmányban diszkutáltuk. Azt tapasztaltuk, hogy a hőmérséklet emelésével a sikeresség a 300 K körüli számítási hőmérsékletnél eléri a maximális értékét, így ennek alkalmazása javasolható.

A gyógyszertervezésben talán legnagyobb kihívást az a szituáció jelenti, amikor a ligandum-célpont komplexet a vízszerkezettel együtt szeretnénk előállítani. Ez praktikusan a korábbi szakaszokban ismertetett dokkoló eljárások alkalmazását igényelné a kísérletileg meghatározott vagy számított vízpozíciók figyelembe vételével. Ez egy igazi „tyúk vagy tojás” probléma: a ligandum és a vízmolekulák egymás kötődését befolyásolják, így nehéz – vagy inkább lehetetlen – eldönteni, melyikkel kezdjük a hidratált komplex szerkezetének összeállítását. Ha a vizekkel indítunk, akkor mennyit és melyik partnerre helyezzünk el? Ha a ligandum dokkolást végezzük előbb a száraz célpontra, akkor mi a garancia arra, hogy a vízmolekulák be tudnak diffundálni a ligandum által elzárt kavitásokba a célmolekula felszínén (lásd még a fenti [D17] közleménynél leírtakat)? Számos próbálkozás született e kérdések megválaszolására, amelyeket a fent már említett munkánkban [D15] külön fejezetben tárgyaltunk. Mivel végleges megoldást egyik ismert módszer sem hozott, ezért egy teljesen új stratégiát dolgoztunk ki a HydroDock [D19] protokollunkban. A HydroDock a dokkolási és a célmolekula (kötőzseb) felszínének vizezési lépéseit egymással párhuzamosan végzi el (**6. ábra**, 1. és 2. lépések), majd az előálló hidratált célpontot és dokkolt ligandum-célpont komplexet egyesíti (3. lépés) a ligandummal ütköző vízmolekulák eliminálásával.


CÉLPONT
+
LIGANDUM
+
VIZEK
1
SZÁRAZ KOMPLEX
2
HIDRATÁLT CÉLPONT
3
EGYESÍTETT ÉS FINOMÍTOTT KOMPLEX
4
A KOMPLEX MD PILLANATFELVÉ-TELEI
5
REPREZENTATÍV HIDRATÁLT KOMPLEX

6. ábra A HydroDock protokoll

Megjegyzendő, hogy a célpont felszínének hidratálásához a MobyWat-hoz hasonló módszerre van szükség, amely nem csupán az interfészre, de a teljes célpont felszínre is képes a vízszerkezetet kiszámítani. Ezután következik a hidratált komplex MD alapú számítása (4. lépés), majd az átlagos szerkezethez legközelebbi reprezentáns kiválasztása (5. lépés). A protokollt az influenza A vírus M2 transzmembrán ioncsatornáján próbáltuk ki és az amantadin sorozatba tartozó csatornablokkoló ligandumok esetében a kísérletileg kimért szerkezetekkel jó egyezést kaptunk. E virális ioncsatornán a kationok mellett vízmolekulák is közlekednek és utóbbiak az amantadinhoz hasonló kationos fejjel rendelkező ligandumok kötésében igen nagy szerepet játszanak. Ezt követően a HydroDock segítségével elvégeztük az amantadin kötési módjainak feltárását a SARS-CoV-2 boríték fehérjéje által kialakított ioncsatorna transzmembrán doménjére, amely egy bíztató gyógyszercélpont lett[249] a COVID-19 elleni küzdelemben. Ez esetben és az ioncsatorna mentén az amantadin több kötési módját is sikerült azonosítanunk és itt is jó egyezést kaptunk a korábbi NMR mérésekben azonosított kötődési mintázattal. Tekintettel arra, hogy az ioncsatornáknál és más célpontoknál is nagyon gyakran előfordul, hogy a belső felszínükhöz vízmolekulák kapcsolódnak, a HydroDock megoldást nyújt ez esetekben a ligandum tervezéshez és az alkalmazott új stratégiája miatt a fenti kérdésekre is választ ad.

3. Energia

A célpont-ligandum komplexek szerkezetének ismeretében (**2. fejezet**) és a megfelelő paraméterek birtokában elvégezhető a partnerek közötti kölcsönhatás erősségének számítása. A ligandum kötődési erősségének méréséről és számításáról több összegző mű is megjelent[250,251]. Természetesen itt sem célom, hogy ezeket kivonatoljam, csupán egy áttekintést nyújtok a lehetőségekről – az elméleti módszerek esetében jórészt fókuszálva az általunk leginkább alkalmazott és fejlesztett végpont módszerekre.

3.1. A célpont-ligandum kötődés termodinamikai optimalizálása

A célpont-ligandum kötéserősség egyensúlyi (**2. egyenlet,** másodlagos kölcsönhatásokkal létrejött, 1:1 sztöchiometriájú komplex képződésére felírva) jellemzésére a termodinamikai potenciálfüggvényeknek a reakciót kísérő változásai szolgálnak. A kötődési reakció szabadentalpiaváltozása (ΔG_b, ahol b a *binding* rövidítése, a standardizálás jelzésétől az egyszerűség kedvéért a továbbiakban eltekintek) a potenciálfüggvény definíciójából adódóan a jól ismert kapcsolatban (**3. egyenlet**) áll annak entalpiaváltozásával (ΔH_b) és entrópiaváltozásával (ΔS_b), ahol T a termodinamikai hőmérséklet. A mérések során sok esetben a kötődési reakció egyensúlyi állandója kerül meghatározásra, amelyből a ΔG_b értéke szintén számolható.

$$\text{Célpont}(H_2O)_C + \text{Ligandum}(H_2O)_L \rightleftharpoons \text{Célpont:Ligandum}(H_2O)_{C+L-w} + w\,H_2O \qquad (\mathbf{2})$$

$$\Delta G_b = \Delta H_b - T\Delta S_b \qquad (\mathbf{3})$$

Anyagszerkezeti szempontból a ΔH_b komponens a kialakuló és megszűnő célpont-ligandum kötések, valamint a dehidratáció ($w>0$) során az említett partnerek és a vízmolekulák egymással alkotott kötésrendszerének átalakulását írja le. A ΔS_b a reakcióban részt vevő molekulák transzlációs, rotációs, stb. szabadsági fokaihoz rendelhető molekuláris rendezetlenség változását számszerűsíti. A gyógyszertervezés során mindhárom termodinamikai potenciálfüggvény változását érdemes nyomonkövetni, ezek együttesen jelentik a ligandum termodinamikai profilját (ujjlenyomatát)[252–254]. A termodinamkai profil iránytűként szolgál a ligandum optimális (át)tervezéséhez. Érdekes módon a ligandum entrópikus optimalizálása, azaz szerkezetének átalakítása abból a célból, hogy a célponthoz kötődéskor a maximális ΔS_b álljon elő, könnyebbnek tűnik, mint az entalpikus optimalizálás[255]. A nehézségek mellett ugyanakkor érdemes a ΔH_b mérésével és számításával külön is foglalkozni, mert az entalpikus ligandum-optimalizálással[256,257] például a HIV-1 proteáz inhibició esetében nem csak erősebben kötődő, de szelektívebb és jobb rezisztencia profillal is rendelkező gyógyszerek születtek[258,259].

3.1.1. A kötődési mérések és korlátaik

A célpont-ligandum kötődés előző szakaszban bevezetett termodinamikai paramétereinek kísérleti meghatározásához elég széles módszertani paletta áll rendelkezésre. A különböző detektálást alkalmazó[260,261] kötési esszék[262,263] többsége az egyensúlyi állandó és ezáltal a ΔG_b meghatározására képes. Az oldatokkal dolgozó kötési esszéken túl lehetőség van több fázist is bevonni a vizsgálatokba, például a felületi plazmonrezonancia spektroszkópia (SPR)[264] eljárás képes a sejtmembránt mimikáló rögzített lipid kettősrétegbe vagy proteoliposzómába ágyazott fehérje célpontokkal is dolgozni. A ligandumnak az előző szakaszban említett teljes termodinamikai profilját (ΔG_b, ΔH_b és ΔS_b) a gyakorlatban az izotermális titrációs kalorimetria (ITC)[265] szolgáltatja, amelyet a kötődési mérések *gold standard*-jaként emlegetnek. Az ITC valóban nélkülözhetetlen eszköz lett a vezérvegyületek optimalizálásában[266,267]. A technika érzékenységének a fejlődésével az elmúlt évtizedekben a méréshez szükséges minta mennyisége is szignifikánsan lecsökkent, azaz a módszer áteresztőképessége megnőtt. Az összes előnye mellett az ITC is korlátokkal küzd a reprodukálhatóság terén[268] amely gyakran a felhasznált oldatok bemérési koncentrációjának hibájából származik[269].

A kötődési mérések hibái általában nagyon sokrétűek lehetnek[270] és sok esetben a célpont (fehérje) koncentrációjának meghatározási problémáira vezethetők vissza. A fehérjepreparátum sokszor igen szűkösen áll rendelkezésre és/vagy mátrix hatások is jelentkezhetnek az egyszerű és gyakran alkalmazott (kalibráció nélküli) spektrofotometriás koncentrációmérésekben. A robosztusabb és ezért referenciának tekinthető aminosav-analízist[271,272] még az ITC-s tanulmányoknak is sajnos csak a töredéke alkalmazza a fehérje törzsoldat koncentrációjának meghatározására.

3.1.2. A kötődési termodinamika számításának molekulamechanikai eljárásai

Az MM szintű szerkezeti számítások legfontosabb összetevői az erőtér és a kereső eljárások (**2.2.1. fejezet**), amelyekkel egy vagy több célpont-ligandum komplex szerkezetet előállíthatunk. A globális keresések során a valóságos szerkezeti sokaságból reprezentatív mintát nyerhetünk az E_{teljes} függvény (**1. egyenlet**) minimumainak alapos feltérképezésével, míg a lokális keresések esetében rendszerint megelégszünk egy optimalizált szerkezettel. A minta terjedelmét tekintve, a kötődési termodinamikai számítások esetében ezért statisztikus és végpont módszerekről beszélhetünk[273], aszerint hogy a kötődés során előforduló több állapotot vagy csak a végállapotokhoz tartozó szerkezeteket használjuk fel a számításokhoz.

A statisztikus módszerek[251,274] a legtöbb esetben az MD számítások során előállított trajektóriákra épülnek[275] és a statisztikus termodinamika[276] összefüggései segítségével teremtenek kapcsolatot a kötődési termodinamikai mennyiségek és a V függvény között. Elterjedten alkalmazzák ezek közül a szabadenergia-perturbációs (FEP)[277], a termodinamikai integrációs (TI)[278], lineáris kölcsönhatási energia (LIE)[279], az átlagos erő potenciálja (PMF)[278,280], valamint a nem-egyensúlyi munka (NEW)[281] módszereket. Bár a statisztikus módszerek a kiterjedt mintázás miatt elméletileg pontosabban írják le a kötődés termodinamikáját, a rutinszerű alkalmazásuk még mindig nem triviális[282] és korlátokkal küzd például a mintázás[283,284] és az oldószer kezelése[285] során.

A végpont módszerek a célpont-ligandum komplex (C-L), valamint a különálló célpont (C) és ligandum (L) szerkezeteket használják fel a számítás vépontjaként, így gyakran szolgáltatják a dokkoló programok pontozó (scoring) ΔG_b függvényeit[115]. A kötődés entalpikus és entrópikus részeit (**3. egyenlet**) rendszerint több tag összegeként írják le, hasonlóan a **4. egyenlet**hez, amelynek különböző változatait összefoglaló művek tárgyalják[115,286,287].

$$\Delta G_b = E_{\mathrm{inter}}(\mathrm{C\text{-}L}) + \Delta E_{\mathrm{intra}}(\mathrm{C}) + \Delta E_{\mathrm{intra}}(\mathrm{L}) + \Delta G_{\mathrm{szolv}} - T\Delta S_{\mathrm{tr\text{-}rot}} - T\Delta S_{\mathrm{konf}}(\mathrm{C}) - T\Delta S_{\mathrm{konf}}(\mathrm{L}) + \mathrm{konst.} \qquad \textbf{(4)}$$

A kifejezés első része a kötődés tisztán entalpikus részét írja le, a C és L molekulák közötti (C-L komplexen belüli) intermolekuláris-, valamint a partnerek intramolekuláris energiaváltozásai felhasználásával. A következő (de)szolvatációs szabadentalpiaváltozás tag értelemszerűen entalpikus és entrópikus hozzájárulást is tartalmaz. Végül a transzlációs és rotációs (tr-rot) szabadsági fokok valamint a konformációváltozás (konf) befagyásához köthető entrópikus tagokat vesszük számba. Rendszerint a különböző végpont módszerek a **4. egyenlet** egyes tagjait (például a $T\Delta S_{\mathrm{tr\text{-}rot}}$ tagot) állandónak veszik és ilyenkor ezek a többváltozós lineáris regresszió során a konstansba olvadnak össze. Az entalpikus tagokat az E_{teljes} függvénnyel (**1. egyenlet**) számítják sokszor elhanyagolva a $\Delta E_{\mathrm{intra}}$ tagokat és csak az $E_{\mathrm{inter}}(\mathrm{C\text{-}L})$-t hagyva meg.

A számítások egyszerűsítése végett előfordulnak pusztán ligandum alapú megközelítések is, például a $T\Delta S_{konf}$ hozzájárulást több ΔG_b függvény[161,288,289] egyszerűen csak a ligandum rotálható kötéseinek számával becsli. Különösen fehérje-fehérje kötődés ΔG_b számításánál gyakran alkalmazzák[290] a szintén MM alapú, molekulafelszín számolást is tartalmazó általánosított Born (MM/GBSA) vagy Poisson-Boltzmann (MM/PBSA) végpont módszereket[291] is. Ugyanakkor utóbbi módszerek az implicit vízmodellből (**2.2.2. fejezet**) fakadóan számos korláttal küszködnek[292], ami jórészt érvényes az explicit vízmodellt nem alkalmazó egyéb végpont módszerekre is.

3.1.3. Kvantummechanikai eljárások

Az MM megközelítések korlátai elsősorban az erőterek tökéletlenségéből és paraméterkészleteik hiányosságaiból fakadnak. A fent ismeretett MD alapú kötési termodinamikai számításoknál az erőtereken túl azonban egyéb tényezők, mint például a molekuláris mozgások alkalmazott eltérő kényszerfeltételei, a sokaságok eltérő mintázása is további reprodukálhatósági korlátokat jelentenek[293]. A mintázási problémák a végpont módszerek esetében értelemszerűen kevésbé jelentkeznek, azonban például az erőtér és a vízmodell problematikája ez esetben is fennáll.

Az erőtér problémára egyik megoldást a kvantummechanikai (QM) módszerek alkalmazása jelentheti. Az erőtér parametrizálás problémája a QM esetében nem lép fel[294], ugyanakkor a számítási idő jelentősen megnő, főleg az *ab initio* módszerek alkalmazása esetén, amit a szemiempírikus eljárások alkalmazásával jelentősen csökkenteni lehet[295]. A megfelelő vízmodell megtalálása a QM alapú számítások általános problémája és a mai napig kutatás tárgyát képezi[296]. A nyitott kérdések ellenére napjainkra a QM a gyógyszertervezésben igen sokrétűen alkalmazásra került[294,297]. Ezen belül a célpont-ligandum kölcsönhatások QM alapú számításának is igen kiterjedt irodalma van[298,299]. A számítási idő további csökkentésére alapvetően két megközelítés bontakozott ki az elmúlt évtizedekben[300]. A fragmentáláson alapuló megközelítéseknél a célpont kötőzsebének fragmenseit (fehérje esetében az aminosavakat, kisebb peptid szakaszokat) vágjuk ki és ezt a kivágott kötőzsebet számítjuk együtt a ligandummal[301,302]. A másik megközelítés során a teljes célpontot MM szinten kezeljük, de a ligandum, valamint a környező kötőzseb QM szinten kerül számításra. Utóbbi QM/MM[303–305] módszerek pontosságuk mellett szintén küszködnek korlátokkal[306,307].

3.2. Eredmények

A célpont-ligandum kötés szerkezetének számításánál (**2. fejezet**) láthattuk, hogy a hidratáció szerepe kulcsfontosságú. Ennek megfelelően a tisztán vákuumban történő számítások egyre inkább kiszorulnak az irodalomból. Akár MM, akár QM szinten végezzük a kötődési termodinamikai paraméterek becslését, az energiaszámítások során alapvetően implicit vagy explicit vízmodelleket alkalmazhatunk, illetve ezek hibridjeit. Ennek megfelelően tárgyalom a továbbiakban a területen kapott eredményeinket, kezdve az egyszerűbb, implicit megközelítéssel.

3.2.1. Energiaszámítások implicit vízmodellel [D20-D23]

A gyors dokkoló eljárások ΔG_b függvényei (**3.1.2. fejezet**) kettős szerepet látnak el. Egyrészt egy adott ligandum kötési módjának globális keresése során célfüggvényként alkalmazva elvezetnek az optimális (ideális esetben a valós) kötési módhoz. Másrészt az adott kötőzsebbe dokkolt különböző ligandumok rangsorolását is a számított ΔG_b értékeik alapján tehetjük meg első megközelítésben. Mi az utóbbi szerepben alkalmaztuk az AutoDock 3.0[161] program ΔG_b függvényének egy módosított verzióját egy enzimológiai tanulmányunkban [D20]. A munka kiindulópontja az volt, hogy a kötődési mérések során a sivatagi sáska kimotripszin inhibitor (SGCI) és az ízeltlábú tripszinek között öt nagyságrenddel erősebb kötést tapasztaltak, mint az SGCI és az emlős tripszinek között. E kiugróan nagy különbség szerkezeti háttere nem volt triviális, így kimérésre került az SGCI peptidnek a rák tripszinnel alkotott atomi felbontású komplexe. Ezen túl modellezéssel előállítottam a szarvasmarha tripszin-SGCI komplex szerkezetét is, ami lehetővé tette, hogy összehasonlítsam az SGCI kölcsönhatási mintázatát a két tripszin esetében. Már e munka során felmerült, hogy az AutoDock 3.0 eredetileg kisebb ligandumokra kalibrált ΔG_b függvénye az SGCI méretű peptidek esetén irreális, pozitív értékeket szolgáltat (ennek részleteiről a [D21] tanulmány kapcsán írok bővebben), így e munka során az entalpikus E_{inter}(C-L) valamint a deszolvatációs (ΔG_{szolv}, **4. egyenlet**) tagok összege került alkalmazásra. A ΔG_{szolv} tagot az AutoDock 3.0 egy atomi fragmens térfogatokon és térfogati betöltöttségen[308] alapuló implicit vízmodell segítségével számítja, emellett a víz árnyékoló hatását az elektrosztatikus kölcsönhatásokra az E_{inter}(C-L)-ben egy távolságfüggő relatív permittivitás függvénnyel közelíti (**2.2.2 fejezet**). A számítások eredményeképpen a kísérleti eredményekkel összhangban lévő, szignifikánsan erősebb kölcsönhatási energiát kaptunk a SGCI rák tripszinnel képzett komplexére, mint a szarvasmarha tripszin esetében. Ezen túlmenően, aminosavankénti lebontásban meg tudtuk mutatni, hogy e kölcsönhatási energia többlet a rák tripszin esetében egy kiterjedt kötési interfész régiót jelöl ki, amely az SGCI szelektivitásának alapját képezi.

A fentiekben ismertetett enzimológiai tanulmányunk [D20] során, a nagy méretű (35 aminosav hosszú) SGCI peptid ligandummal szerzett tapasztalatok rámutattak arra, hogy az AutoDock 3.0 kis molekulákra kalibrált ΔG_b függvénye változtatás nélkül nem alkalmazható a nagy méretű ligandumok számítására, így felmerült a ΔG_b függvény átalakításának és újrakalibrálásának a gondolata a következő [D21] tanulmányunkban. Mivel az entalpikus tagok használhatónak bizonyultak, már a kezdeti AutoDock 2.4-es verzió[309] óta és a [D20] tanulmányunkban is, valamint a probléma elsődlegesen a ligandum méretével tűnt kapcsolatosnak, így azokat a tagokat vizsgáltuk meg kritikusan, amelyek kizárólag a ligandum alapján kerültek számításra. Két ilyen „gyanús" tag volt szembeötlő az AutoDock 3.0 ΔG_b függvényében. Az egyik a ligandum belső torziós szabadsági fokai befagyásából számította a $T\Delta S_{konf}$(L) (**4. egyenlet**) értékét, egyszerűen azon torziós szögek számát véve, amelyek mentén a ligandum szabadon elforog (N_{tor}) beszorozva a regressziós koefficienssel. A másik tag a ligandumnak a bulk vízmolekulákkal alkotott H-hídjai összes energiáját fejezte ki, a ligandumban lévő, H-kötésben részt vehető atomok számát felszorozva konstansokkal. Mindkét tag értelemszerűen pozitív értékekkel járult hozzá a teljes ΔG_b-hez, ami oda vezetett, hogy egy SGCI méretű ligandum esetében a teljes ΔG_b-re hibás pozitív értéket számolt a program. Megjegyzendő, hogy e tagok más szerzőknek is „szemet szúrtak" szénhidrát ligandumok esetén[310]. Mi egy 50 komplexből álló, számos peptidet tartalmazó szettel végeztük el [D21] a ΔG_b függvény módosítását, amelyet az eredeti AutoDock 3.0 szetthez képest jelentősen nagyobb átlagos ligandum méret jellemzett. A többváltozós lineáris regresszió eredményei azt mutatták, hogy a fent említett, két problémás tag elhagyása már önmagában javított a megmaradt ΔG_b függvény statisztikai paraméterein. A függvényben így csak a bimolekulás, entalpikus és deszolvatációs tagok maradtak meg, amelyeket a következő lépésben kiegészítettünk ligandum-alapú deszkriptorokkal, és így új, hibrid ΔG_b függvényekhez jutottunk, igen jó statisztikai paraméterekkel. A munkát keresztvalidációk és a deszkriptorok robosztusságának vizsgálata egészítette ki. Ilyen módon nem csak az eredeti ΔG_b függvény kijavítását értük el, de a nagyobb, pl. peptid ligandum-célmolekula kötéserősség gyors számolására is rendelkezésre bocsátottunk egy új, hibrid kalkulátort és jó irodalmi visszhangot is kapott[311–317].

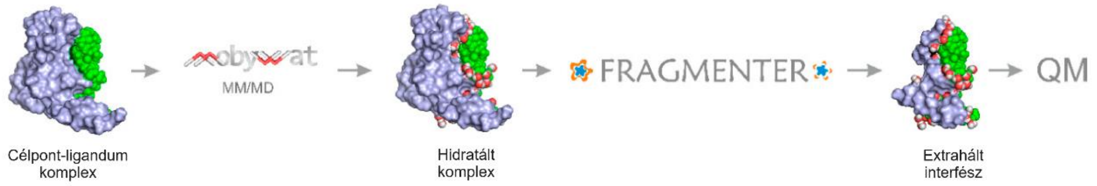
Az AutoDock 3.0 entalpikus tagjaival sikerült jó eredményeket kapnunk a humán α_{2A}-adrenoceptor agonistái esetében [D22] is. Ebben az esetben csupán kísérletileg meghatározott kötési adatok álltak rendelkezésünkre egy viszonylag kis agonista szettre. A célfehérje atomi felbontású szerkezetét homológiamodellezéssel kellett előállítani a munkához, majd ebbe dokkoltuk az agonista ligandumokat. Mivel ezek az agonisták kis molekulák, itt csupán a fent említett belső torziók befagyását leíró entrópikus tagot hagytuk el és már így is jó korrelációhoz jutottunk. Bevezettünk egy bináris deszkriptort is a ΔG_b kifejezésébe, amely egy kulcsfontosságú fragmens jelenlétét írta le a ligandumban. Tekintettel arra, hogy ez esetben egy adott célponton dolgoztunk, ennek a deszkriptornak az alkalmazása nem pusztán számszerűen tette jobbá a regressziós statisztikát, de lehetőségünk nyílt a fragmens jelenlétének értelmezésére a kötési termodinamika szempontjából. A jó korrelációk jelezték, hogy a célpont homológia modellje kvantitatív munkára is bevált és a kölcsönható, kötőzsebben lévő aminosavakat is jól be tudtuk azonosítani, összevetve más irodalmi eredményekkel.

Szintén kis méretű, aromás ligandumok dokkolására épült az *in silico* molekuláris kölcsönhatási (és affinitási) ujjlenyomatokat bevezető tanulmányunk [D23] is. Ez esetben 44 ligandum-célmolekula komplexet (39 különböző ligandum és 31 célpont) vettünk alapul és első lépésben elvégeztük az AutoDock 3.0 dokkoló eljárás validálását a kísérletileg kimért ligandum pozíciókkal történt összevetés segítségével. Mivel az AutoDock – hasonlóan az egyéb gyors dokkoló eljárásokhoz – az ilyen, gyógyszer méretű ligandumokra került bevezetésre, a várakozásnak megfelelően a ligadumok dokkolt kötési módjai jó egyezéseket mutattak a kísérletesekkel. Továbbá a dokkolt ligandum ΔG_b értékei rendre alacsonyabbak voltak, mint a referencia mért ligandum kötési módból számított ΔG_b értékek. Ilyen módon a pontozó függvény (a számított ΔG_b) is relevánsnak mutatkozott, tekintve, hogy mind szerkezetileg, mind energiáját tekintve jó kötési módokat eredményezett. Ezt követően a 31 × 39 kereszt-dokkolást elvégezve minden ligandum és célpont számított ΔG_b értéke a rendelkezésünkre állt és ebből az adott célfehérje eredeti liganduma ΔG_b-jének kivonásával megkaptuk az adott célfehérje kötőzsebére adott ligandum esetében vonatkozó ΔG_b többletet. Így eljutottunk a molekuláris kölcsönhatási ujjlenyomat (MIF) mátrixhoz, amely egy adott ligandum szelektivitását (vagy éppen promiszkuitását) írja le és jó alapot nyújt a mellékhatás-spektrumának előrejelzéséhez. E tanulmányunkat hivatkozta J.M. Thornton és csoportja is[318]. Később a MIF-et kombináltuk[319,320] a ligandumok farmakológiai profilját leíró mátrixokkal és így praktikus polifarmakológiás tervezési eszközökhöz is jutottunk.

3.2.2. Energiaszámítások explicit és hibrid vízmodellekkel [D24-D27]

Az előző fejezetben tárgyalt implicit vízmodelles megközelítések a mai napig igen fontos szerepet játszanak a gyógyszertervezésben, különösen ha sok ligandumot kell egyszerre vizsgálni. Ugyanakkor az explicit vízmolekulák hiányában nyilvánvaló korlátokkal is rendelkeznek, például a vízmolekulák irányított H-kötéseiből adódó termodinamikai hozzájárulások pontos számítása terén. Az explicit vízmodellel végrehajtott MD számítások viszont igen jó mintát szolgáltatnak a részletesebb és pontosabb kötődési termodinamikai számításokhoz. Ez esetben akár magának az oldószernek a célpont-ligandum komplex kötődéséhez történő hozzájárulása is pontosan számítható. Egy ilyen, dekompozíciós megközelítést írtunk le ciklodextrin-gyógyszer komplexeken [D24], amelynek alapját explicit vízmodelles, teljes MD mintázáson alapuló PMF (**3.1.2. fejezet**) számítások adták. A PMF profilok lefutásának (szimmetriájának) elemzése lehetővé tette a ciklodextrin esetében előforduló alternatív komplexképződési útvonalak nyomonkövetését. A ΔH_b és a ΔS_b tagok dekompozíciójára is mód nyílt, előbbi esetben külön kovalens és másodlagos, valamint a másodlagos esetben négy féle (ciklodextrin-víz, gyógyszer-víz, cikodextrin-gyógyszer, víz-víz) intermolekuláris, entalpikus hozzájárulást különítettünk el. Megállapítottuk, hogy a vizsgált komplexképződési folyamatok összességében entalpia-vezéreltek, ezen belül azonban igen tanulságos a dekompozíció során kapott egyes entalpiaváltozások szerepe. A fent említett négy intermolekuláris tag számértékeiben nagyságrendileg meghaladja a teljes komplexképződést leíró ΔG_b-t, valamint a négy tag nagyságrendileg összemérhető egymással. Így belátható, hogy a vizet tartalmazó három entalpikus tag számítása végeredményben kulcsfontosságú a ΔG_b számítása szempontjából. A vízhez köthető entrópikus hozzájárulás nagyságrendileg összemérhető a ΔG_b-vel, így ez is megerősíti az explicit vízmodell használatának szükségességét a ΔG_b számítások pontossága szempontjából. A fent ismertetett eredmények közvetlen alkalmazást nyertek a ciklodextrin-alapú nanoszerkezetek építésének szerkezeti és termodinamikai alapjairól szóló [D25] tanulmányunkban. A ciklodextrin dimerek alapvető építőkövei molekuláris nyakláncoknak illetve nanocsöveknek. E nanoszerkezetek felépülését templátként elősegíthetik a cikodextrinnel komplexet alkotó kisebb ligandumok, mint például a fent említett tanulmányban vizsgált gyógyszerek, olyan módon, hogy mindkét monomer ciklodextrin egységgel kölcsönhatnak és összekapcsolják azokat dimerré. Természetesen felmerül a kérdés, hogy ez a templát funkció milyen mértékű az egyes ligandumok esetében és a ciklodextrin monomerek kapcsolódása szerkezetileg hogy valósul meg a termodinamikailag legkedvezőbb módon. A felvetett kérdésekre azt a választ kaptuk, hogy a daidzein bot alakú molekulái esetében lesz termodinamikailag kedvező a cikodextrinek és a templátként alkalmazott molekula kölcsönhatása. Az ilyen templátokkal képzett komplexek esetében a cikodextrin molekulák a lehetséges három elrendeződés közül a szélesebb peremükkel egymás felé fordulva, fej-fej illeszkedésben alkotják a legstabilabb dimereket. A ΔG_b-nek az előző [D24] tanulmányhoz hasonló dekompozíciója itt tehát az említett nanoszerkezetek további tervezéséhez szolgáltat egy teljes termodinamikai leírást.

Míg az előző két esetben ismertetett eredményeink MM szintű megközelítést használtak viszonylag kisebb ligandumoknál (daidzein) és töltés nélküli célpontnál (ciklodextrin), a nagyobb méretű, többszörösen töltött célpont-ligandum komplexek esetében felmerül az elektronszerkezeti effektusok számításának szükségessége. Fokozottan érvényes ez, ha a környező vízmolekulák is kölcsönhatásba lépnek a töltésekkel, ami nyilván alaphelyzet például a peptid ligandumok esetében. Az elektronszerkezeti effektusok pontos leírását leginkább QM szinten lehet elvégezni, így mi is megvizsgáltuk a QM-alapú kötési termodinamika kalkulátorok kifejlesztésének lehetőségét. A munkához a felvezetésben (**3.1.3. fejezet**) említett két alternatíva közül a fragmentálásos utat (**7. ábra**) követtük és tanulmányunkban [D26] célul tűztük ki a ΔH_b számítására szolgáló eljárás kidolgozását a célpont-ligandum komplex szerkezetből, mint végpontból kiindulva. (Külön a ΔH_b számításának és az entalpikus optimalizációnak a fontosságáról a **3.1. fejezet** elején írtam.) Ehhez több praktikus és elméleti problémát kellett megoldanunk. Szerencsére a „száraz" célpont-ligandum komplexek hidratálását nem kellett kidolgoznunk, mert arra ismét alkalmazni tudtuk a MobyWat programunkat (**2.3.3. fejezet**).



7. ábra A célpont (szürke) – ligandum (zöld) komplex hidratálása után a célpont ligandumhoz kapcsolódó részeit az interfész vízmolekulákkal (piros-fehér) együtt kivágva a QM számítás céljaira előkészített szerkezethez jutunk.

Az első praktikus probléma a fragmentálás automatizálása volt. A peptid szakaszok célfehérjéből történő kivágása látszólag triviálisan megoldható a rendelkezésre álló modellező eszközökkel. A gyakorlati munka azonban azt mutatta, hogy a hidratált komplexből a kötési interfész kivágása csak részben és körülményesen volt megvalósítható, nem volt erre a feladatra praktikus célszerszám. Ki kellett ezért dolgoznunk a Fragmenter nevű eszközt, amely különböző (távolság, fragmens hossz) beállításokkal kivágja a célmolekula ligandumot környező peptid fragmenseit (**7. ábra**), valamint automatikusan blokkolja vagy éppen szabadon, ionos formában hagyja a peptid fragmensek láncvégeit. Ezen túlmenően egy előzetes energia-ellenőrzést is végez és az extrahált interfész régió szerkezetét és a partnereket is külön Mopac input fájlokban, futtatásra kész állapotban adja ki. A legnagyobb elméleti kérdés az alkalmazandó vízmodell volt. Bár a célpont-ligandum interfész teljes hidrátszerkezete a rendelkezésünkre állt a MobyWat jóvoltából, de kérdés volt, mely explicit hidratációs szférák használhatók a ΔH_b számításában, valamint hogy az implicit/explicit vagy a kettőt kombináló hibrid hidratációs megközelítés vezet-e eredményre. A különböző modellek összehasonlításába a kis ligandumos komplexeken túl bevontunk nagy méretű peptid ligadumokat tartalmazó rendszereket is, több mint 3000-es relatív molekulatömegig. Végül a hibrid vízmodell, egy kiterjedtebb explicit vízszférával igen jó korrelációt eredményezett a kísérleti ΔH_b értékekkel és csupán egy paraméterrel a mért és a szemiempírikus QM szinten, PM7 parametrizálással számított entalpiaváltozásokat össze lehetett skálázni, így egy használható, szerkezeti alapú kalkulátorhoz jutottunk.

Míg a fenti [D26] tanulmányban QM szerkezeti optimalizálást végeztünk, a kiindulási, MM-optimalizált komplex és a QM-optimalizált, extrahált interfész szerkezetek összehasonlításával azt tapasztaltuk, hogy maga a QM szerkezeti optimalizálás ritkán okoz nagy változást a szerkezeteinkben. Ebből logikusan következett, hogy az MM-optimalizált szerkezetet felhasználva egy QM végpontszámítással (*single point*, 1SCF) kombinálva kapjuk meg a ΔH_b számításához szükséges képződéshőket, a QM szerkezeti optimalizációs lépés elhagyásával. Az elképzelést a nemrég megjelent [D27] tanulmányunkban teszteltük le egy kiterjesztett adathalmazon, egyúttal a ΔG_b számítását is célul tűzve ki. *Nota bene*, a Mopac 1SCF számítások alkalmazása szerkezetek energetikai értékelésére másoknál is felmerült[295]. Mi tehát a MobyWat-hidratált, MM-minimalizált célpont-ligandum komplex szerkezetekből (és alkotóikból külön is) a Fragmenteres kivágás után a fentiekhez hasonlóan, szemiempírikus PM7 szinten 1SCF számítással kaptuk meg a képződéshő értékeket [D27]. Ez esetben is a hibrid vízmodell eredményezte a legjobb statisztikát a ΔH_b regressziónál. A számításokat QM optimalizációval valamint más parametrizálásokkal is megismételtük és nem kaptunk javulást az 1SCF (PM7, hibrid vízmodell) eredményekhez képest, így utóbbi megközelítést használtuk fel a ΔG_b kalkulátor kidolgozásához is. Ehhez alapul vettük a QM-számított entalpikus tagot, valamint egy többváltozós regresszióban kiegészítettük ligandum-alapú deszkriptorokkal. Számos deszkriptort végigpróbálva a ligandum nehéz atomjai és összes atomjai számának hányadosa bizonyult a legjobbnak. Az így kapott QMH-L kalkulátorral [D27] végül 1 kcal/mol (4,184 kJ/mol) átlagos hibával tudtuk a ΔG_b-t becsülni. Annak köszönhetően, hogy a QMH-L egyenletet úgy építettük meg, hogy az a ΔG_b-ben az entalpikus tagot külön tartalmazza és a ligandum alapú deszkriptornak sikerült alapos fizikai értelmezést is adnunk. Az így kapott kalkulátor gyors számítást tesz lehetővé, mivel a szerkezet előállítása MM szinten történik. Emellett az említett deszkriptor várhatóan más pontozó (ΔG_b) függvényekben is alkalmazható lesz, mivel nem függ a ligandum méretétől és így használatakor a **3.2.1. szakasz**ban az N_{tor} kapcsán említett problémák értelemszerűen nem jelentkeznek.

3.2.3. A kötéserősség méretfüggése és a hatékonysági indexek [D28-D30]

Több tanulmány kimutatta, hogy a ΔG_b függ a ligandum méretétől[256,321,322]. Ennek oka egyaránt lehet az entalpikus és az entrópikus hozzájárulások méretfüggése is, amelynek végső tisztázása még várat magára. A jelenség mindenesetre adott és igen logikus koncepcióként felmerült, hogy a ΔG_b-t függetlenítsük a ligandum méretétől, egyszerűen elosztva azt a ligandum méretével korreláló valamely mennyiséggel, mint például a ligandum nehéz atomjainak száma[323] (NHA, **5. egyenlet**), vagy a relatív molekulatömeg[324].

$$\mathrm{EI_{NHA}} = \frac{\Delta \mathrm{G_b}}{\mathrm{NHA}} \qquad (5)$$

Az így kapott mennyiségeket ligandum hatékonysági indexnek[325] (*efficiency index*, EI, **5. egyenlet**) nevezzük. A nevezőben szereplő, a ligandum méretétől függő mennyiségek ugyanakkor elég széleskörűen használatosak farmakokinetikai szűrőként is a gyógyszertervezés során. Az ilyen gyors szűrések során általában egy adott határérték alatti molekulatömeggel, lipofilicitási értékkel, stb. rendelkező vegyületeket szelektálnak ki a sok százezer vegyületet tartalmazó molekulakönyvtárakból, például a Lipinski-féle szabályok[326] alkalmazásával. Bár e szabályokat kiterjedten alkalmazták a tervezésben, egyre több jel utalt rá, hogy általános használatuk inkább káros, mint hasznos lehet sok esetben (erről e fejezet végén szólok még).

Az EI-k terén közölt első tanulmányunkban [D28] megvizsgáltuk azok szerkezeti alapú számításának lehetőségét és a korábbi [D21] tanulmányunkban alkalmazott 53 rendszer mellett még további 20 fehérje-célpont adatot gyűjtöttünk külső validáló szettként. Meglepően jó korrelációkat kaptunk a kísérleti és számított EI értékek között. Az EI-k kifejezésében (**5. egyenlet**) a nevezőben szereplő szokásos mennyiségeken túl további méretfüggő molekuláris deszkriptorokkal is definiáltunk új EI-ket és megvizsgáltuk a használhatóságukat is. Különösen a Wiener-index és az egyes főtengelyekere vett tehetetlenségi nyomatékok szorzata esetében kaptunk kiemelekedően jó korrelációkat. Értelmezésünk szerint az EI-kben a ligandum méretétől függő mennyiségekkel történő osztás összességében a ΔG_b entrópikus részét az illesztési konstansba olvassza és a maradék, mérettől függetlenített entalpikus részek között a korreláció erősebb lesz. Formálisan tekinthetőek az osztáskor alkalmazott méretfüggő deszkriptorok reciprokai illesztési súlyoknak is. Alkalmazhatóságukat tekintve az említett két legjobb, új EI esetében a vizsgált gyógyszerek és „nem gyógyszerek” csoportjai között jó szelekciót tapasztaltunk, míg pusztán a ΔG_b esetében a 2 csoport között nem láttunk elkülönülést. Ezen túlmenően a két új EI hatékonynak bizonyult egy 1760 molekulát tartalmazó könyvtár szűrésében is a progeszteron receptoron, az aktív noretindron referenciavegyületet a felső 0.5 %-ba sorolva (a ΔG_b önmagában csak a felső 10 %-ba sorolta a noretindront első körben).

Az előbbiekben tárgyalt [D28] tanulmányunkban az AutoDock 3 és 4 verzióinak a módosított ΔG_b függvényeit alkalmaztuk a számításokhoz. Felvetődött a kapott eredmények általánosíthatóságának a kérdése más, kereskedelemben hozzáférhető, szintén gyakran alkalmazott dokkoló eljárások (Gold, Glide) pontozó függvényeire (Goldscore, Chemscore) és ezeknek komponenseire is. Emellett még további ligandum deszkriptorokat is vizsgáltunk a következő tanulmányunkban [D29] , mint például a szénatomok száma és az oktanol-víz megoszlási együttható, amelyek a lipofilicitáson keresztül a (de)szolvatációs hozzájárulásra lehetnek jellemzőek. A vizsgált deszkriptorok közül az utóbbi kettő, valamint a Wiener-index mutatta itt is a legjobb eredményeket, a korrelációs együttható ez esetekben markánsan megnőtt. Tekintettel arra, hogy az említett ligandum-alapú deszkriptorok és a származtatott

EI-k számítása is nagyon gyors, így használatuk nem lassítja le a dokkolt ligandum szerkezetek rangsorolási folyamatát, ellenben a fenti eredmények alapján hatékonyabbá tehetik azt.

A ligandumot (méretét, flexibilitását, hidrofobicitását, H-kötési képességét) jellemző deszkriptorok és a belőlük származtatott EI-k (**5. egyenlet,** együttesen *property filter*-ként említve [D30]-ban) alkalmazhatóságát vizsgáltuk meg elemző tanulmányunkban [D30], ahol bevezetésre kerültek a szelektivitásukra és érzékenységükre vonatkozó statisztikai mérőszámok is. A tanulmány 1.-4. irodalmi áttekintő fejezeteiben a *drug-likeness* (gyógyszerszerűség, DL) koncepciót jártuk először körbe, megkülönböztetve annak általános és specifikus alkalmazásait. Részletesen tárgyaltuk az általános DL korlátait és a betegség, gyógyszeradminisztráció és célpont szerinti specifikus DL kategóriákat is. Ezt követően a ΔG_b-nek a ligandum méretétől való — fentiekben említett – függését tárgyaljuk részletesen, majd az EI koncepciót. Mindezek az előzmények vezettek el logikailag az 5. fejezethez, ahol a DL koncepció előbbiekben tárgyalt korlátainak okán megvizsgáltuk az említett deszkriptorok és EI-k alkalmazhatóságát egy olyan statisztikai modellben, ahol a ΔG_b-függést az elemzés során „kikapcsoltuk”. Ehhez olyan gyógyszer és „nem gyógyszer” szetteket állítottunk össze, amelyek ΔG_b eloszlásukat tekintve megegyeztek. Ezután megvizsgáltuk az egyes deszkriptorok és EI-k eloszlásának különbségét is a két szettben és ahol lehetőség volt rá, az eloszlásfüggvényeket analitikus formában illesztéssel meghatároztuk. Az eloszlásfüggvények ismeretében a szelektivitásra és érzékenységre mérőszámokat vezettünk be. Ezek alapján az adott szelektivitáshoz és érzékenységhez általános határértékeket tudtunk megadni az egyes deszkriptorokhoz és EI-khez, valamint betegség-specificitásukat szintén kvantitatíven jellemeztük. Összességében e vizsgálataink egyértelmű, kvantitatív képet adtak a DL koncepció felvezetésben említett korlátairól és alkalmazhatóságáról az egyes deszkriptorok és EI-k terén. A bevezetett mérőszámok általánosan alkalmazhatóak a tanulmányban nem vizsgált, vagy a jövőben definiálásra kerülő deszkriptorok és EI-k minősítésére és határértékük kalibrálásához, ami várhatóan a ligandumok szűrésének pontosságát növelni fogja.

4. Kitekintés

A célpont-ligandum kölcsönhatások számítása a racionális, szerkezetalapú gyógyszertervezés egyik legfontosabb lépése. A dolgozatban felvonultatott elméleti megközelítések az elmúlt évtizedekben igen nagy változásokon mentek át, sok esetben pozitív irányban. A számítástechnikai eszköztár mennyiségi és minőségi növekedésével e fejlődés világszerte egyre gyorsul. Ennek egyik következménye a korábban igen költségesnek tartott fizikai-kémiai alapú eljárások, például a hosszabb MD számítások mennyiségének az örvendetes növekedése a gyógyszertervezésben. Másrészt megfigyelhető az informatikai eszköztárat fokozottabban kihasználó, kevesebb (esetenként minimális vagy akár zérus) fizikai-kémiai alapot igénylő, gépi tanulásos és/vagy a nagy adatmennyiséget feldolgozó ún. ismeretalapú (*knowledge-based*) módszerek virágzása. A két megközelítés egyelőre meglehetősen párhuzamosan fejlődik, az igazán hatékony összekapcsolódásuk még várat magára a gyógyszertervezés területén. Nagy kérdés, hogy e kapcsolódás a közeljövőben meg tud-e valósulni – tekintettel a két irányzat filozófiájának különbségeire. Már jelenleg is érzékelhető, hogy a fizikai-kémiai alapokat nagy mértékben nélkülöző módszerek nehezen tudnak általánosan alkalmazható megoldásokat kínálni a molekulaszerkezeti számítások terén.

Eddigi kutatási projektjeim során én is tapasztaltam, hogy az ismeretalapú módszerek használhatósága korlátozott és nagy szakmai tapasztalatot igényel. Ilyen eset volt, amikor a citokróm CYP3A4 szerkezetének előállítását és gyors publikálását kérték tőlem olasz partnerek, mert az akkor még nem volt kimérve. Én ezt nem tettem meg, mert nem álltak rendelkezésre megfelelően (szekvenálisan) homológ templát fehérjék az ismeretalapú[327] építkezéshez. Ráadásul az említett citokróm várhatóan igen nagy kötőzsebbel rendelkezett a hem csoport feletti térrészben és a fehérjék belsejében az ilyen kiterjedt üregek jóslása szinte lehetetlen vállalkozás volt az akkor rendelkezésre álló homológiamodellezéses eljárásokkal. Mivel a CYP3A4 egy igen fontos metabolikus enzim, így mások több homológiamodellt is publikáltak. Végül sor került a kísérletes szerkezetmeghatározásra és az addigi tanulmányok homológiamodellezett szerkezeteit a valós szerkezet birtokában a kísérletes közlemény (negatívan) meghivatkozta[328], megjegyezve, hogy a korábban közölt homológiamodelleknek ez esetben korlátozott volt a használhatósága.

A fehérjeszerkezetek predikcióján túl a célpont-ligandum komplex szerkezetek előállítása (a számítógépes dokkolás) terén is erősen túlfűtöttnek bizonyulnak az ismeretalapú predikciós megközelítésekkel kapcsolatos várakozások. A kezdeti ígéretek után komoly tanulmányok jelennek meg a mesterséges intelligencia alapú módszerek korlátairól[329], amelyeket részben a fizikai-kémiai alapok mellőzése okoz. A dolgozatban tárgyalt (de)hidratációs problémák megoldására sem látszik[330] áttörést hozó ismeretalapú megközelítés. Ugyanakkor a meglévő, fizikai-kémiai alapú módszerek sokszor már most is kellő pontosságot adnak és korlátaik ismeretében a jövőbeli fejlődési irányaik jól kivehetőek az erőterek, a kereső módszerek és a vízmodellek terén is. A tervezésben az explicit vízmodellekre épülő technikák fejlesztése, valamint a QM-re épülő eljárások hatékonyságának és rutin alkalmazhatóságának növelése tűnnek a legfontosabb feladatoknak.

Az értekezés elején már idézett Nobel-díjas R. Henderson és társai véleménycikkének[2] összegző gondolatával zárnám e rövid kitekintésemet, amely szerint *„... solving the protein-folding problem means making accurate predictions of structures from amino acid sequences* ***starting from first principles*** *based on the underlying* ***physics and chemistry*** *...”*. Ez az iránymutatás jól jöhet a gyógyszertervezés említett, nyitott kérdéseinek megoldásában is, ahol a célpont-ligandum komplexek és a kapcsolt hidrátszerkezetek számításában a jövőben is építeni tudunk majd a fizikai-kémiai alaptörvényekre.

Megemlékezés

Ezúton kívánok megemlékezni szegedi tanáraimról, akik a kémia szeretetét megerősítették bennem és a matematikai, természetes szerves, komplex- és fizikai kémiai területeken inspiráltak. Huhn Péter, Vincze Irén, Burger Kálmán és Nagypál István professzorok előadásaira mindig szívesen gondolok vissza. Horváth István és Körtvélyesi Tamás egyetemi docensek kiváló témavezetőim voltak – fájdalmasan korán távoztak.

Köszönetnyilvánítás

Munkám során többször támaszkodhattam az MTA, az NKFIH, az ÚNKP és az EU által nyújtott anyagi támogatásokra. A szuperszámítógépes háttér folyamatos biztosításáért a Kormányzati Informatikai Fejlesztési Ügynökségnek (KIFÜ) tartozom köszönettel.

Penke Botond professzor úr témavezetőként segített elindulnom a kutatói pályán, a peptidek iránti fogékonyságomat és a területen átadott alapismereteket is neki köszönhetem, valamint, hogy mindvégig ösztönzött elméleti munkám folytatására.

Szegeden kiváló oktatóktól tanulhattam, igazi egyéniségektől, akik nagy tudással rendelkeznek. Nehéz lenne őket itt mind felsorolni és nagyon nem szeretnék senkit kihagyni – hálával gondolok mindannyiukra.

Az elmúlt két évtizedben számos kutatóhelyen megfordultam, itthon és külföldön. Az egyes állomásokon különböző, hasznos impulzusok értek, amelyek többnyire a gyakorlat, az alkalmazások oldaláról inspirálták a kutatómunkámat, gyümölcsöző együttműködéseket eredményezve. Ezúton szeretnék köszönetet mondani a Szegedi Tudományegyetem Orvosi Vegytani Intézete, az Uppsalai Egyetem Biokémiai, majd Sejt- és Molekuláris Biológiai Tanszéke, a Tartui Egyetem Kémiai Intézete, az Eötvös Loránd Tudományegyetem Biokémiai és Genetikai Tanszékei, a Semmelweis Egyetem Szerves Vegytani Intézete és – nem utolsó sorban – jelenlegi munkahelyem, a Pécsi Tudományegyetem Általános Orvostudományi Kar Farmakológiai és Farmakoterápiai Intézete kedves dolgozóinak.

Szüleimnek hálás vagyok támogató szeretetükért.

Irodalomjegyzék

(1) Szent-Györgyi Albert: *Az élet jellege*, 2. kiadás. Magvető Kiadó, Budapest, 1975.

(2) Moore, P. B.; Hendrickson, W. A.; Henderson, R.; Brunger, A. T. The Protein-Folding Problem: Not yet Solved. *Science* **2022**, *375* (6580), 507–507. https://doi.org/10.1126/science.abn9422.

(3) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(4) Nogales, E. Profile of Joachim Frank, Richard Henderson, and Jacques Dubochet, 2017 Nobel Laureates in Chemistry. *Proc. Natl. Acad. Sci.* **2018**, *115* (3), 441–444. https://doi.org/10.1073/pnas.1718898114.

(5) Frank, J.; Zhu, J.; Penczek, P.; Li, Y.; Srivastava, S.; Verschoor, A.; Radermacher, M.; Grassucci, R.; Lata, R. K.; Agrawal, R. K. A Model of Protein Synthesis Based on Cryo-Electron Microscopy of the E. Coli Ribosome. *Nature* **1995**, *376* (6539), 441–444. https://doi.org/10.1038/376441a0.

(6) Henderson, R. The Potential and Limitations of Neutrons, Electrons and X-Rays for Atomic Resolution Microscopy of Unstained Biological Molecules. *Q. Rev. Biophys.* **1995**, *28* (2), 171–193. https://doi.org/10.1017/S003358350000305X.

(7) Adrian, M.; Dubochet, J.; Lepault, J.; McDowall, A. W. Cryo-Electron Microscopy of Viruses. *Nature* **1984**, *308* (5954), 32–36. https://doi.org/10.1038/308032a0.

(8) Lepault, J.; Booy, F. P.; Dubochet, J. Electron Microscopy of Frozen Biological Suspensions. *J. Microsc.* **1983**, *129* (1), 89–102. https://doi.org/10.1111/j.1365-2818.1983.tb04163.x.

(9) Darby, J. F.; Hopkins, A. P.; Shimizu, S.; Roberts, S. M.; Brannigan, J. A.; Turkenburg, J. P.; Thomas, G. H.; Hubbard, R. E.; Fischer, M. Water Networks Can Determine the Affinity of Ligand Binding to Proteins. *J. Am. Chem. Soc.* **2019**, *141* (40), 15818–15826. https://doi.org/10.1021/jacs.9b06275.

(10) Pintilie, G.; Zhang, K.; Su, Z.; Li, S.; Schmid, M. F.; Chiu, W. Measurement of Atom Resolvability in Cryo-EM Maps with Q-Scores. *Nat. Methods* **2020**, *17* (3), 328–334. https://doi.org/10.1038/s41592-020-0731-1.

(11) Renaud, J.-P.; Chari, A.; Ciferri, C.; Liu, W.; Rémigy, H.-W.; Stark, H.; Wiesmann, C. Cryo-EM in Drug Discovery: Achievements, Limitations and Prospects. *Nat. Rev. Drug Discov.* **2018**, *17* (7), 471–492. https://doi.org/10.1038/nrd.2018.77.

(12) D'Imprima, E.; Kühlbrandt, W. Current Limitations to High-Resolution Structure Determination by Single-Particle cryoEM. *Q. Rev. Biophys.* **2021**, *54*, e4. https://doi.org/10.1017/S0033583521000020.

(13) Chari, A.; Stark, H. Prospects and Limitations of High-Resolution Single-Particle Cryo-Electron Microscopy. *Annu. Rev. Biophys.* **2023**, *52* (1), 391–411. https://doi.org/10.1146/annurev-biophys-111622-091300.

(14) Shi, Y. A Glimpse of Structural Biology through X-Ray Crystallography. *Cell* **2014**, *159* (5), 995–1014. https://doi.org/10.1016/j.cell.2014.10.051.

(15) Brooks-Bartlett, J. C.; Garman, E. F. The Nobel Science: One Hundred Years of Crystallography. *Interdiscip. Sci. Rev.* **2015**, *40* (3), 244–264. https://doi.org/10.1179/0308018815Z.000000000116.

(16) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Fayazi, M.; Feng, Z.; Flatt, J. W.; Ganesan, S.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Henry, J.; Hudson, B. P.; Khokhriakov, I.; Lawson, C. L.; Liang, Y.; Lowe, R.; Peisach, E.; Persikova, I.; Piehl, D. W.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Vallat, B.; Voigt, M.; Webb, B.; Westbrook, J. D.; Whetstone, S.; Young, J. Y.; Zalevsky, A.; Zardecki, C. RCSB Protein Data Bank (RCSB.Org): Delivery of Experimentally-Determined PDB Structures alongside One Million Computed Structure Models of Proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Res.* **2023**, *51* (D1), D488–D508. https://doi.org/10.1093/nar/gkac1077.

(17) Wang, H.-W.; Wang, J.-W. How Cryo-Electron Microscopy and X-Ray Crystallography Complement Each Other: Cryo-EM and X-Ray Crystallography Complement Each Other. *Protein Sci.* **2017**, *26* (1), 32–39. https://doi.org/10.1002/pro.3022.

(18) Zheng, H.; Handing, K. B.; Zimmerman, M. D.; Shabalin, I. G.; Almo, S. C.; Minor, W. X-Ray Crystallography over the Past Decade for Novel Drug Discovery – Where Are We Heading Next? *Expert Opin. Drug Discov.* **2015**, *10* (9), 975–989. https://doi.org/10.1517/17460441.2015.1061991.

(19) Zheng, H.; Hou, J.; Zimmerman, M. D.; Wlodawer, A.; Minor, W. The Future of Crystallography in Drug Discovery. *Expert Opin. Drug Discov.* **2014**, *9* (2), 125–137. https://doi.org/10.1517/17460441.2014.872623.

(20) Ezkurdia, I.; Juan, D.; Rodriguez, J. M.; Frankish, A.; Diekhans, M.; Harrow, J.; Vazquez, J.; Valencia, A.; Tress, M. L. Multiple Evidence Strands Suggest That There May Be as Few as 19 000 Human Protein-Coding Genes. *Hum. Mol. Genet.* **2014**, *23* (22), 5866–5878. https://doi.org/10.1093/hmg/ddu309.

(21) Pertea, M.; Salzberg, S. L. Between a Chicken and a Grape: Estimating the Number of Human Genes. *Genome Biol.* **2010**, *11* (5), 206. https://doi.org/10.1186/gb-2010-11-5-206.

(22) Hopkins, A. L.; Groom, C. R. The Druggable Genome. *Nat. Rev. Drug Discov.* **2002**, *1* (9), 727–730. https://doi.org/10.1038/nrd892.

(23) Kermani, A. A. A Guide to Membrane Protein X-ray Crystallography. *FEBS J.* **2021**, *288* (20), 5788–5804. https://doi.org/10.1111/febs.15676.

(24) Zhao, J.; Lin King, J. V.; Paulsen, C. E.; Cheng, Y.; Julius, D. Irritant-Evoked Activation and Calcium Modulation of the TRPA1 Receptor. *Nature* **2020**, *585* (7823), 141–145. https://doi.org/10.1038/s41586-020-2480-9.

(25) Niedzialkowska, E.; Gasiorowska, O.; Handing, K. B.; Majorek, K. A.; Porebski, P. J.; Shabalin, I. G.; Zasadzinska, E.; Cymborowski, M.; Minor, W. Protein Purification and Crystallization Artifacts: The Tale Usually Not Told: Protein Purification and Crystallization Artifacts. *Protein Sci.* **2016**, *25* (3), 720–733. https://doi.org/10.1002/pro.2861.

(26) Guo, Y. Be Cautious with Crystal Structures of Membrane Proteins or Complexes Prepared in Detergents. *Crystals* **2020**, *10* (2), 86. https://doi.org/10.3390/cryst10020086.

(27) Tsuchiya, Y. Discrimination between Biological Interfaces and Crystal-Packing Contacts. *Adv. Appl. Bioinforma. Chem.* **2008**, 99. https://doi.org/10.2147/AABC.S4255.

(28) Halle, B. Biomolecular Cryocrystallography: Structural Changes during Flash-Cooling. *Proc. Natl. Acad. Sci.* **2004**, *101* (14), 4793–4798. https://doi.org/10.1073/pnas.0308315101.

(29) Hajdu, J.; Neutze, R.; Sjögren, T.; Edman, K.; Szöke, A.; Wilmouth, R. C.; Wilmot, C. M. Analyzing Protein Functions in Four Dimensions. *Nat. Struct. Biol.* **2000**, *7* (11).

(30) Westenhoff, S.; Meszaros, P.; Schmidt, M. Protein Motions Visualized by Femtosecond Time-Resolved Crystallography: The Case of Photosensory vs Photosynthetic Proteins. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102481. https://doi.org/10.1016/j.sbi.2022.102481.

(31) Biedermannová, L.; Schneider, B. Hydration of Proteins and Nucleic Acids: Advances in Experiment and Theory. A Review. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2016**, *1860* (9), 1821–1835. https://doi.org/10.1016/j.bbagen.2016.05.036.

(32) Adams, P. D.; Grosse-Kunstleve, R. W.; Hung, L.-W.; Ioerger, T. R.; McCoy, A. J.; Moriarty, N. W.; Read, R. J.; Sacchettini, J. C.; Sauter, N. K.; Terwilliger, T. C. *PHENIX* : Building New Software for Automated Crystallographic Structure Determination. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58* (11), 1948–1954. https://doi.org/10.1107/S0907444902016657.

(33) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (4), 486–501. https://doi.org/10.1107/S0907444910007493.

(34) Brünger, A. T.; Kuriyan, J.; Karplus, M. Crystallographic *R* Factor Refinement by Molecular Dynamics. *Science* **1987**, *235* (4787), 458–460. https://doi.org/10.1126/science.235.4787.458.

(35) Brünger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (5), 905–921. https://doi.org/10.1107/S0907444998003254.

(36) DePristo, M. A.; De Bakker, P. I. W.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure* **2004**, *12* (5), 831–838. https://doi.org/10.1016/j.str.2004.02.031.

(37) Blundell, T. L.; Jhoti, H.; Abell, C. High-Throughput Crystallography for Lead Discovery in Drug Design. *Nat. Rev. Drug Discov.* **2002**, *1* (1), 45–54. https://doi.org/10.1038/nrd706.

(38) Blundell, T. L.; Patel, S. High-Throughput X-Ray Crystallography for Drug Discovery. *Curr. Opin. Pharmacol.* **2004**, *4* (5), 490–496. https://doi.org/10.1016/j.coph.2004.04.007.

(39) Müller, I. Guidelines for the Successful Generation of Protein–Ligand Complex Crystals. *Acta Crystallogr. Sect. Struct. Biol.* **2017**, *73* (2), 79–92. https://doi.org/10.1107/S2059798316020271.

(40) Savage, H.; Wlodawer, A. Determination of Water Structure around Biomolecules Using X-Ray and Neutron Diffraction Methods. *Methods Enzymol.* **1986**, *127*, 162–183. https://doi.org/10.1016/0076-6879(86)27014-7.

(41) Carugo, O. Correlation between Occupancy and B Factor of Water Molecules in Protein Crystal Structures. *Protein Eng. Des. Sel.* **1999**, *12* (12), 1021–1024. https://doi.org/10.1093/protein/12.12.1021.

(42) Finney, J. L.; Eley, D. D.; Richards, R. E.; Franks, F. The Organization and Function of Water in Protein Crystals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **1997**, *278* (959), 3–32. https://doi.org/10.1098/rstb.1977.0029.

(43) Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein-Ligand Binding Sites and Its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3* (12), 973–980. https://doi.org/10.1016/S1074-5521(96)90164-7.

(44) Afonine, P. V.; Grosse-Kunstleve, R. W.; Adams, P. D. A Robust Bulk-Solvent Correction and Anisotropic Scaling Procedure. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (7), 850–855. https://doi.org/10.1107/S0907444905007894.

(45) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. The Solvent Component of Macromolecular Crystals. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71* (5), 1023–1038. https://doi.org/10.1107/S1399004715006045.

(46) Badger, J. [17] Modeling and Refinement of Water Molecules and Disordered Solvent. In *Methods in Enzymology*; Academic Press, 1997; Vol. 277, pp 344–352. https://doi.org/10.1016/S0076-6879(97)77019-8.

(47) Halle, B. Protein Hydration Dynamics in Solution: A Critical Survey. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2004**, *359* (1448), 1207–1224. https://doi.org/10.1098/rstb.2004.1499.

(48) Islam, S. A.; Weaver, D. L. Molecular Interactions in Protein Crystals: Solvent Accessible Surface and Stability. *Proteins Struct. Funct. Bioinforma.* **1990**, *8* (1), 1–5. https://doi.org/10.1002/prot.340080103.

(49) Kossiakoff, A. A.; Sintchak, M. D.; Shpungin, J.; Presta, L. G. Analysis of Solvent Structure in Proteins Using Neutron D2O-H2O Solvent Maps: Pattern of Primary and Secondary Hydration of Trypsin. *Proteins Struct. Funct. Genet.* **1992**, *12* (3), 223–236. https://doi.org/10.1002/prot.340120303.

(50) Chatake, T.; Fujiwara, S. A Technique for Determining the Deuterium/Hydrogen Contrast Map in Neutron Macromolecular Crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **2016**, *72* (1), 71–82. https://doi.org/10.1107/S2059798315021269.

(51) Levitt, M.; Park, B. H. Water: Now You See It, Now You Don't. *Structure* **1993**, *1* (4), 223–226. https://doi.org/10.1016/0969-2126(93)90011-5.

(52) Tanaka, I.; Chatake, T.; Fujiwara, S.; Hosoya, T.; Kusaka, K.; Niimura, N.; Yamada, T.; Yano, N. Current Status and near Future Plan of Neutron Protein Crystallography at J-PARC. In *Methods in Enzymology*; Elsevier, 2020; Vol. 634, pp 101–123. https://doi.org/10.1016/bs.mie.2020.01.002.

(53) Kono, F.; Kurihara, K.; Tamada, T. Current Status of Neutron Crystallography in Structural Biology. *Biophys. Physicobiology* **2022**, *19* (0), n/a. https://doi.org/10.2142/biophysico.bppb-v19.0009.

(54) Wüthrich, K. The Way to NMR Structures of Proteins. *Nat. Struct. Biol.* **2001**, *8* (11).

(55) Wüthrich, K. Brownian Motion, Spin Diffusion and Protein Structure Determination in Solution. *J. Magn. Reson.* **2021**, *331*, 107031. https://doi.org/10.1016/j.jmr.2021.107031.

(56) Otting, G. NMR Studies of Water Bound to Biological Molecules. *Prog. Nucl. Magn. Reson. Spectrosc.* **1997**, *31* (2–3), 259–285. https://doi.org/10.1016/S0079-6565(97)00012-5.

(57) Armstrong, B. D.; Han, S. Overhauser Dynamic Nuclear Polarization To Study Local Water Dynamics. *J. Am. Chem. Soc.* **2009**, *131* (13), 4641–4647. https://doi.org/10.1021/ja809259q.

(58) Keserű György Miklós, Náray-Szabó Gábor. *Molekulamechanika*; A kémia újabb eredményei; Akadémiai Kiadó, Budapest, 1995.

(59) Andrew R. Leach. *Molecular Modeling*, 2nd ed.; Pearson Education Ltd, 2001.

(60) Poltev, V. Molecular Mechanics: Principles, History, and Current Status. In *Handbook of Computational Chemistry*; Leszczynski, J., Ed.; Springer Netherlands: Dordrecht, 2015; pp 1–48. https://doi.org/10.1007/978-94-007-6169-8_9-2.

(61) Andrews, D. H. The Relation Between the Raman Spectra and the Structure of Organic Molecules. *Phys. Rev.* **1930**, *36* (3), 544–554. https://doi.org/10.1103/PhysRev.36.544.

(62) Allinger, N. L. Calculation of Molecular Structure and Energy by Force-Field Methods. In *Advances in Physical Organic Chemistry*; Gold, V., Bethell, D., Eds.; Academic Press, 1976; Vol. 13, pp 1–82. https://doi.org/10.1016/S0065-3160(08)60212-9.

(63) Engler, E. M.; Andose, J. D.; Schleyer, P. V. R. Critical Evaluation of Molecular Mechanics. *J. Am. Chem. Soc.* **1973**, *95* (24), 8005–8025. https://doi.org/10.1021/ja00805a012.

(64) Hendrickson, J. B. Molecular Geometry. I. Machine Computation of the Common Rings. *J. Am. Chem. Soc.* **1961**, *83* (22), 4537–4547. https://doi.org/10.1021/ja01483a011.

(65) Wiberg, K. B. A Scheme for Strain Energy Minimization. Application to the Cycloalkanes. *J. Am. Chem. Soc.* **1965**, *87* (5), 1070–1078.

(66) Allinger, N. L. Molecular Mechanics. In *Theoretical and Computational Models for Organic Chemistry*; Formosinho, S. J., Csizmadia, I. G., Arnaut, L. G., Eds.; Springer Netherlands: Dordrecht, 1991; pp 125–135. https://doi.org/10.1007/978-94-011-3584-9_8.

(67) Karplus, M. Development of Multiscale Models for Complex Chemical Systems: From H+H $_2$ to Biomolecules (Nobel Lecture). *Angew. Chem. Int. Ed.* **2014**, *53* (38), 9992–10005. https://doi.org/10.1002/anie.201403924.

(68) Levitt, M. Birth and Future of Multiscale Modeling for Macromolecular Systems (Nobel Lecture). *Angew. Chem. Int. Ed.* **2014**, *53* (38), 10006–10018. https://doi.org/10.1002/anie.201403691.

(69) Warshel, A. Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture). *Angew. Chem. Int. Ed.* **2014**, *53* (38), 10020–10031. https://doi.org/10.1002/anie.201403689.

(70) Van Der Spoel, D. Systematic Design of Biomolecular Force Fields. *Curr. Opin. Struct. Biol.* **2021**, *67*, 18–24. https://doi.org/10.1016/j.sbi.2020.08.006.

(71) Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: *Ab Initio* Force Field Methods Derived from Quantum Mechanics. *J. Chem. Phys.* **2018**, *148* (9), 090901. https://doi.org/10.1063/1.5009551.

(72) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. In *Advances in Protein Chemistry*; Elsevier, 2003; Vol. 66, pp 27–85. https://doi.org/10.1016/S0065-3233(03)66002-X.

(73) Vanommeslaeghe, K.; Guvench, O.; Jr., D. M. A. Molecular Mechanics. *Curr. Pharm. Des.* **2014**, *20* (20), 3281–3292. https://doi.org/10.2174/13816128113199990600.

(74) Kollman, P. A.; Weiner, P. K.; Dearing, A. Studies of Nucleotide Conformations and Interactions. The Relative Stabilities of Double-Helical B-DNA Sequence Isomers. *Biopolymers* **1981**, *20* (12), 2583–2621. https://doi.org/10.1002/bip.1981.360201208.

(75) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197. https://doi.org/10.1021/ja00124a002.

(76) Weiner, P. K.; Kollman, P. A. AMBER: Assisted Model Building with Energy Refinement. A General Program for Modeling Molecules and Their Interactions. *J. Comput. Chem.* **1981**, *2* (3), 287–303. https://doi.org/10.1002/jcc.540020311.

(77) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106* (3), 765–784. https://doi.org/10.1021/ja00315a051.

(78) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids: An All Atom Force Field. *J. Comput. Chem.* **1986**, *7* (2), 230–252. https://doi.org/10.1002/jcc.540070216.

(79) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(80) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591.

(81) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field: Improved Protein Side-Chain Potentials. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958. https://doi.org/10.1002/prot.22711.

(82) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(83) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.

(84) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217. https://doi.org/10.1002/jcc.540040211.

(85) Nilsson, L.; Karplus, M. Empirical Energy Functions for Energy Minimization and Dynamics of Nucleic Acids. *J. Comput. Chem.* **1986**, *7* (5), 591–616. https://doi.org/10.1002/jcc.540070502.

(86) Jorgensen, W. L.; Swenson, C. J. Optimized Intermolecular Potential Functions for Amides and Peptides. Structure and Properties of Liquid Amides. *J. Am. Chem. Soc.* **1985**, *107* (3), 569–578. https://doi.org/10.1021/ja00289a008.

(87) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. https://doi.org/10.1021/ja00214a001.

(88) Lii, J.-H.; Allinger, N. L. The MM3 Force Field for Amides, Polypeptides and Proteins. *J. Comput. Chem.* **1991**, *12* (2), 186–199. https://doi.org/10.1002/jcc.540120208.

(89) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111* (23), 8551–8566. https://doi.org/10.1021/ja00205a001.

(90) Lii, J. H.; Allinger, N. L. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 2. Vibrational Frequencies and Thermodynamics. *J. Am. Chem. Soc.* **1989**, *111* (23), 8566–8575. https://doi.org/10.1021/ja00205a002.

(91) Lii, J. H.; Allinger, N. L. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 3. The van Der Waals' Potentials and Crystal Data for Aliphatic and Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111* (23), 8576–8582. https://doi.org/10.1021/ja00205a003.

(92) Allinger, N. L.; Chen, K.; Lii, J.-H. An Improved Force Field (MM4) for Saturated Hydrocarbons. *J. Comput. Chem.* **1996**, *17* (5–6), 642–668. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<642::AID-JCC6>3.0.CO;2-U.

(93) Hwang, M. J.; Stockfisch, T. P.; Hagler, A. T. Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *J. Am. Chem. Soc.* **1994**, *116* (6), 2515–2525. https://doi.org/10.1021/ja00085a036.

(94) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of Class II Force Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules. *J. Comput. Chem.* **1994**, *15* (2), 162–182. https://doi.org/10.1002/jcc.540150207.

(95) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

(96) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3* (6), 2034–2045. https://doi.org/10.1021/ct700127w.

(97) Halgren, T. A.; Damm, W. Polarizable Force Fields. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 236–242. https://doi.org/10.1016/S0959-440X(00)00196-2.

(98) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947. https://doi.org/10.1021/jp027815+.

(99) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4046–4063. https://doi.org/10.1021/ct4003702.

(100) Schlick, T. Optimization Methods in Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; pp 1–71. https://doi.org/10.1002/9780470125809.ch1.

(101) CAUCHY, A. Methode Generale Pour La Resolution Des Systemes d'equations Simultanees. *CR Acad Sci Paris* **1847**, *25*, 536–538.

(102) Lemaréchal, C. Cauchy and the Gradient Method. In *Optimization Stories*; Grötschel, M., Ed.; EMS Press, 2012; pp 251–254. https://doi.org/10.4171/dms/6/27.

(103) Hestenes, M. R.; Stiefel, E. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Natl. Bur. Stand.* **1952**, *49*, 409–435.

(104) Fletcher, R.; Reeves, C. M. Function Minimization by Conjugate Gradients. *Comput. J.* **1964**, *7* (2), 149–154. https://doi.org/10.1093/comjnl/7.2.149.

(105) Polak, E.; Ribiere, G. Note sur la convergence de méthodes de directions conjuguées. *Rev. Fr. Inform. Rech. Opérationnelle Sér. Rouge* **1969**, *3* (R1), 35–43.

(106) Raphson, J. *Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, & expedita...* , Editio secunda.; typis Tho. Braddyll, prostant venales apud Iohannem Taylor ...: Londini, 1697. https://doi.org/10.3931/e-rara-13516.

(107) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45* (1), 503–528. https://doi.org/10.1007/BF01589116.

(108) Deák István. *Véletlenszám generátorok és alkalmazásuk*; Az operációkutatás matematikai módszerei; Akadémiai Kiadó, 1986; Vol. 3.

(109) Michael T. Heath. *Scientific Computing*, 2nd ed.; McGraw-Hill, 2002.

(110) Holland, J. H. Outline for a Logical Theory of Adaptive Systems. *J. ACM* **1962**, *9* (3), 297–314. https://doi.org/10.1145/321127.321128.

(111) Goldberg, D. E.; Holland, J. H. Genetic Algorithms and Machine Learning. *Mach. Learn.* **1988**, *3* (2), 95–99. https://doi.org/10.1023/A:1022602019183.

(112) Salmaso, V.; Moro, S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Front. Pharmacol.* **2018**, *9*, 923. https://doi.org/10.3389/fphar.2018.00923.

(113) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949. https://doi.org/10.1038/nrd1549.

(114) Pinzi, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* **2019**, *20* (18), 4331. https://doi.org/10.3390/ijms20184331.

(115) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32* (1), 335–373. https://doi.org/10.1146/annurev.biophys.32.110601.142532.

(116) Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136* (2A), A405–A411. https://doi.org/10.1103/PhysRev.136.A405.

(117) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585–590. https://doi.org/10.1038/267585a0.

(118) Levitt, M. Protein Folding by Restrained Energy Minimization and Molecular Dynamics. *J. Mol. Biol.* **1983**, *170* (3), 723–764. https://doi.org/10.1016/S0022-2836(83)80129-6.

(119) Karplus, M.; Kuriyan, J. Molecular Dynamics and Protein Function. *Proc. Natl. Acad. Sci.* **2005**, *102* (19), 6679–6685. https://doi.org/10.1073/pnas.0408930102.

(120) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.

(121) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* **2019**, *151* (7), 070902. https://doi.org/10.1063/1.5109531.

(122) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220* (4598), 671–680. https://doi.org/10.1126/science.220.4598.671.

(123) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

(124) Monroe, J.; Barry, M.; DeStefano, A.; Aydogan Gokturk, P.; Jiao, S.; Robinson-Brown, D.; Webber, T.; Crumlin, E. J.; Han, S.; Shell, M. S. Water Structure and Properties at Hydrophilic and Hydrophobic Surfaces. *Annu. Rev. Chem. Biomol. Eng.* **2020**, *11* (1), 523–557. https://doi.org/10.1146/annurev-chembioeng-120919-114657.

(125) Hingerty, B. E.; Ritchie, R. H.; Ferrell, T. L.; Turner, J. E. Dielectric Effects in Biopolymers: The Theory of Ionic Saturation Revisited. *Biopolymers* **1985**, *24* (3), 427–439. https://doi.org/10.1002/bip.360240302.

(126) Ramstein, J.; Lavery, R. Energetic Coupling between DNA Bending and Base Pair Opening. *Proc. Natl. Acad. Sci.* **1988**, *85* (19), 7231–7235. https://doi.org/10.1073/pnas.85.19.7231.

(127) Smith, P. E.; Pettitt, B. M. Modeling Solvent in Biomolecular Systems. *J. Phys. Chem.* **1994**, *98* (39), 9700–9711. https://doi.org/10.1021/j100090a002.

(128) Mehler, E. L.; Solmajer, T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Eng. Des. Sel.* **1991**, *4* (8), 903–910. https://doi.org/10.1093/protein/4.8.903.

(129) Conway, B. E.; Bockris, J. O.; Ammar, I. A. The Dielectric Constant of the Solution in the Diffuse and Helmholtz Double Layers at a Charged Interface in Aqueous Solution. *Trans. Faraday Soc.* **1951**, *47*, 756. https://doi.org/10.1039/tf9514700756.

(130) Tucker, S. C.; Truhlar, D. G. Generalized Born Fragment Charge Model for Solvation Effects as a Function of Reaction Coordinate. *Chem. Phys. Lett.* **1989**, *157* (1–2), 164–170. https://doi.org/10.1016/0009-2614(89)87227-6.

(131) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127–6129. https://doi.org/10.1021/ja00172a038.

(132) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005–3014. https://doi.org/10.1021/jp961992r.

(133) Born, M. Volumen Und Hydratationswärme Der Ionen. *Z. Für Phys.* **1920**, *1* (1), 45–48. https://doi.org/10.1007/BF01881023.

(134) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J. Comput. Chem.* **2004**, *25* (2), 265–284. https://doi.org/10.1002/jcc.10378.

(135) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson–Boltzmann Equation for Biomolecular Electrostatics: A Tool for Structural Biology. *J. Mol. Recognit.* **2002**, *15* (6), 377–392. https://doi.org/10.1002/jmr.577.

(136) Lamm, G. The Poisson–Boltzmann Equation. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Larter, R., Cundari, T. R., Eds.; Wiley, 2003; Vol. 19, pp 147–365. https://doi.org/10.1002/0471466638.ch4.

(137) Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78* (1), 1–20. https://doi.org/10.1016/S0301-4622(98)00226-9.

(138) Onufriev, A. V.; Izadi, S. Water Models for Biomolecular Simulations. *WIREs Comput. Mol. Sci.* **2018**, *8* (2), e1347. https://doi.org/10.1002/wcms.1347.

(139) Levy, Y.; Onuchic, J. N. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35* (1), 389–415. https://doi.org/10.1146/annurev.biophys.35.040405.102134.

(140) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437* (7059), 640–647. https://doi.org/10.1038/nature04162.

(141) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2002**, *106* (3), 521–533. https://doi.org/10.1021/jp015514e.

(142) Schmid, R. Recent Advances in the Description of the Structure of Water, the Hydrophobic Effect, and the Like-Dissolves-Like Rule. *Monatsh. Chemie* **2001**, *132*, 1295–1326. https://doi.org/10.1007/s007060170019

(143) Dill, K. A.; Truskett, T. M.; Vlachy, V.; Hribar-Lee, B. Modeling water, the hydrophobic effect, and ion solvation. *Annu Rev Biophys Biomol Struct* **2005**, *34*, 173-99. https://doi.org/10.1146/annurev.biophys.34.040204.144517.

(144) Cheng, Y.-K.; Rossky, P. J. Surface Topography Dependence of Biomolecular Hydrophobic Hydration. *Nature* **1998**, *392* (6677), 696–699. https://doi.org/10.1038/33653.

(145) Bernal, J. D.; Fowler, R. H. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *J. Chem. Phys.* **1933**, *1* (8), 515–548. https://doi.org/10.1063/1.1749327.

(146) Jorgensen, W. L. Quantum and Statistical Mechanical Studies of Liquids. 10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103* (2), 335–340. https://doi.org/10.1021/ja00392a016.

(147) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–6271. https://doi.org/10.1021/j100308a038.

(148) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. A Systematic Study of Water Models for Molecular Simulation: Derivation of Water Models Optimized for Use with a Reaction Field. *J. Chem. Phys.* **1998**, *108* (24), 10220–10230. https://doi.org/10.1063/1.476482.

(149) Kiss, P. T.; Baranyai, A. Sources of the Deficiencies in the Popular SPC/E and TIP3P Models of Water. *J. Chem. Phys.* **2011**, *134* (5), 054106. https://doi.org/10.1063/1.3548869.

(150) Hetényi, C.; Van Der Spoel, D. Efficient Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Protein Sci.* **2002**, *11* (7), 1729–1737. https://doi.org/10.1110/ps.0202302.

(151) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand Binding: Functional Site Location, Similarity and Docking. *Curr. Opin. Struct. Biol.* **2003**, *13* (3), 389–395. https://doi.org/10.1016/S0959-440X(03)00075-7.

(152) Laurie ATR; Jackson RM. Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 395–406. https://doi.org/10.2174/138920306778559386.

(153) Hetényi, C.; Körtvélyesi, T.; Penke, B. Mapping of Possible Binding Sequences of Two Beta-Sheet Breaker Peptides on Beta Amyloid Peptide of Alzheimer's Disease. *Bioorg. Med. Chem.* **2002**, *10* (5), 1587–1593. https://doi.org/10.1016/S0968-0896(01)00424-2.

(154) Hetényi, C.; Szabó, Z.; Klement, É.; Datki, Z.; Körtvélyesi, T.; Zarándi, M.; Penke, B. Pentapeptide Amides Interfere with the Aggregation of β-Amyloid Peptide of Alzheimer's Disease. *Biochem. Biophys. Res. Commun.* **2002**, *292* (4), 931–936. https://doi.org/10.1006/bbrc.2002.6745.

(155) Söderhjelm, P.; Tribello, G. A.; Parrinello, M. Locating Binding Poses in Protein-Ligand Systems Using Reconnaissance Metadynamics. *Proc. Natl. Acad. Sci.* **2012**, *109* (14), 5170–5175. https://doi.org/10.1073/pnas.1201940109.

(156) Ghersi, D.; Sanchez, R. Improving Accuracy and Efficiency of Blind Protein-ligand Docking by Focusing on Predicted Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2009**, *74* (2), 417–424. https://doi.org/10.1002/prot.22154.

(157) Hassan, N. M.; Alhossary, A. A.; Mu, Y.; Kwoh, C.-K. Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio-Temporal Integration. *Sci. Rep.* **2017**, *7* (1), 15451. https://doi.org/10.1038/s41598-017-15571-7.

(158) Jofily, P.; Pascutti, P. G.; Torres, P. H. M. Improving Blind Docking in DOCK6 through an Automated Preliminary Fragment Probing Strategy. *Molecules* **2021**, *26* (5), 1224. https://doi.org/10.3390/molecules26051224.

(159) Morris, G. M.; Huey, R.; Olson, A. J. Using AutoDock for Ligand-Receptor Docking. *Curr. Protoc. Bioinforma.* **2008**, *24* (1). https://doi.org/10.1002/0471250953.bi0814s24.

(160) Scodeller, P.; Asciutto, E. K. Targeting Tumors Using Peptides. *Molecules* **2020**, *25* (4), 808. https://doi.org/10.3390/molecules25040808.

(161) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662. https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

(162) Egyed, A.; Kiss, D. J.; Keserű, G. M. The Impact of the Secondary Binding Pocket on the Pharmacology of Class A GPCRs. *Front. Pharmacol.* **2022**, *13*, 847788. https://doi.org/10.3389/fphar.2022.847788.

(163) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(164) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28* (6), 1145–1152. https://doi.org/10.1002/jcc.20634.

(165) Grosdidier, A.; Zoete, V.; Michielin, O. EADock: Docking of Small Molecules into Protein Active Sites with a Multiobjective Evolutionary Optimization. *Proteins Struct. Funct. Bioinforma.* **2007**, *67* (4), 1010–1025. https://doi.org/10.1002/prot.21367.

(166) Grosdidier, A.; Zoete, V.; Michielin, O. Blind Docking of 260 Protein–Ligand Complexes with EADock 2.0. *J. Comput. Chem.* **2009**, *30* (13), 2021–2030. https://doi.org/10.1002/jcc.21202.

(167) Hernandez, M.; Ghersi, D.; Sanchez, R. SITEHOUND-Web: A Server for Ligand Binding Site Identification in Protein Structures. *Nucleic Acids Res.* **2009**, *37* (Web Server), W413–W416. https://doi.org/10.1093/nar/gkp281.

(168) Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21* (9), 1908–1916. https://doi.org/10.1093/bioinformatics/bti315.

(169) Jackson, R. M. Q-FIt: A Probabilistic Method for Docking Molecular Fragments by Sampling Low Energy Conformational Space. *J Comput Aided Mol Des* **2002**, 16(1), 43-57. https://doi.org/10.1023/a:1016307520660

(170) Jr, G. P. B.; Stouten, P. F. W. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J Comput Aided Mol Des* **2000**, 14(4), 383-401. https://doi.org/10.1023/a:1008124202956

(171) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graph. Model.* **1997**, *15* (6), 359–363. https://doi.org/10.1016/S1093-3263(98)00002-3.

(172) Norman, A. W.; Mizwicki, M. T.; Norman, D. P. G. Steroid-Hormone Rapid Actions, Membrane Receptors and a Conformational Ensemble Model. *Nat. Rev. Drug Discov.* **2004**, *3* (1), 27–41. https://doi.org/10.1038/nrd1283.

(173) Van Hoorn, W. P. Identification of a Second Binding Site in the Estrogen Receptor. *J. Med. Chem.* **2002**, *45* (3), 584–589. https://doi.org/10.1021/jm0109661.

(174) Mizwicki, M. T.; Keidel, D.; Bula, C. M.; Bishop, J. E.; Zanello, L. P.; Wurtz, J.-M.; Moras, D.; Norman, A. W. Identification of an Alternative Ligand-Binding Pocket in the Nuclear Vitamin D Receptor and Its Functional Importance in 1α,25(OH) $_2$ -Vitamin D $_3$ Signaling. *Proc. Natl. Acad. Sci.* **2004**, *101* (35), 12876–12881. https://doi.org/10.1073/pnas.0403606101.

(175) Alonso, H.; Gillies, M. B.; Cummins, P. L.; Bliznyuk, A. A.; Gready, J. E. Multiple Ligand-Binding Modes in Bacterial R67 Dihydrofolate Reductase. *J. Comput. Aided Mol. Des.* **2005**, *19* (3), 165–187. https://doi.org/10.1007/s10822-005-3693-6.

(176) Barrera Guisasola, E. E.; Andujar, S. A.; Hubin, E.; Broersen, K.; Kraan, I. M.; Méndez, L.; Delpiccolo, C. M. L.; Masman, M. F.; Rodríguez, A. M.; Enriz, R. D. New Mimetic Peptides Inhibitors of Aβ Aggregation. Molecular Guidance for Rational Drug Design. *Eur. J. Med. Chem.* **2015**, *95*, 136–152. https://doi.org/10.1016/j.ejmech.2015.03.042.

(177) Brown, W. M.; Vander Jagt, D. L. Creating Artificial Binding Pocket Boundaries To Improve the Efficiency of Flexible Ligand Docking. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1412–1422. https://doi.org/10.1021/ci049853r.

(178) Bultum, L. E.; Tolossa, G. B.; Kim, G.; Kwon, O.; Lee, D. In Silico Activity and ADMET Profiling of Phytochemicals from Ethiopian Indigenous Aloes Using Pharmacophore Models. *Sci. Rep.* **2022**, *12* (1), 22221. https://doi.org/10.1038/s41598-022-26446-x.

(179) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand Binding: Functional Site Location, Similarity and Docking. *Curr. Opin. Struct. Biol.* **2003**, *13* (3), 389–395. https://doi.org/10.1016/S0959-440X(03)00075-7.

(180) Chee Wezen, X.; Chandran, A.; Eapen, R. S.; Waters, E.; Bricio-Moreno, L.; Tosi, T.; Dolan, S.; Millership, C.; Kadioglu, A.; Gründling, A.; Itzhaki, L. S.; Welch, M.; Rahman, T. Structure-Based Discovery of Lipoteichoic Acid Synthase Inhibitors. *J. Chem. Inf. Model.* **2022**, *62* (10), 2586–2599. https://doi.org/10.1021/acs.jcim.2c00300.

(181) Collins, T.; Young, G. T.; Millar, N. S. Competitive Binding at a Nicotinic Receptor Transmembrane Site of Two A7-Selective Positive Allosteric Modulators with Differing Effects on Agonist-Evoked Desensitization. *Neuropharmacology* **2011**, *61* (8), 1306–1313. https://doi.org/10.1016/j.neuropharm.2011.07.035.

(182) Daryanavard, M.; Jannesari, Z.; Javeri, M.; Abyar, F. A New Mononuclear Zinc(II) Complex: Crystal Structure, DNA- and BSA-Binding, and Molecular Modeling; in Vitro Cytotoxicity of the Zn(II) Complex and Its Nanocomplex. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2020**, *233*, 118175. https://doi.org/10.1016/j.saa.2020.118175.

(183) El-Wakil, M. H.; Meheissen, M. A.; Abu-Serie, M. M. Nitrofurazone Repurposing towards Design and Synthesis of Novel Apoptotic-Dependent Anticancer and Antimicrobial Agents: Biological Evaluation, Kinetic Studies and Molecular Modeling. *Bioorganic Chem.* **2021**, *113*, 104971. https://doi.org/10.1016/j.bioorg.2021.104971.

(184) Forrestall, K.; Pringle, E. S.; Sands, D.; Duguay, B. A.; Farewell, B.; Woldemariam, T.; Falzarano, D.; Pottie, I.; McCormick, C.; Darvesh, S. A Phenothiazine Urea Derivative Broadly Inhibits Coronavirus Replication via Viral Protease Inhibition. *Antiviral Res.* **2023**, *220*, 105758. https://doi.org/10.1016/j.antiviral.2023.105758.

(185) Ghersi, D.; Sanchez, R. Improving Accuracy and Efficiency of Blind Protein-ligand Docking by Focusing on Predicted Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2009**, *74* (2), 417–424. https://doi.org/10.1002/prot.22154.

(186) Karami, K.; Jamshidian, N.; Zakariazadeh, M.; Momtazi-Borojeni, A. A.; Abdollahi, E.; Amirghofran, Z.; Shahpiri, A.; Nasab, A. K. Experimental and Theoretical Studies of Palladium-Hydrazide Complexes' Interaction with DNA and BSA, in Vitro Cytotoxicity Activity and Plasmid Cleavage Ability. *Comput. Biol. Chem.* **2021**, *91*, 107435. https://doi.org/10.1016/j.compbiolchem.2021.107435.

(187) Lighvan, Z. M.; Khonakdar, H. A.; Heydari, A.; Rafiee, M.; Jahromi, M. D.; Derakhshani, A.; Momtazi-Borojeni, A. A. Spectral and Molecular Docking Studies of Nucleic Acids/Protein Binding Interactions of a Novel Organometallic Palladium (II) Complex Containing Bioactive PTA Ligands: Its Synthesis, Anticancer Effects and Encapsulation in Albumin Nanoparticles. *Appl. Organomet. Chem.* **2020**, *34* (10), e5839. https://doi.org/10.1002/aoc.5839.

(188) Lopata; Jójárt; Surányi; Takács; Bezúr; Leveles; Bendes; Viskolcz; Vértessy; Tóth. Beyond Chelation: EDTA Tightly Binds Taq DNA Polymerase, MutT and dUTPase and Directly Inhibits dNTPase Activity. *Biomolecules* **2019**, *9* (10), 621. https://doi.org/10.3390/biom9100621.

(189) Meli, M.; Colombo, G. Molecular Simulations of Peptides: A Useful Tool for the Development of New Drugs and for the Study of Molecular Recognition. In *Peptide Microarrays*; Cretich, M., Chiari, M., Eds.; Methods in Molecular Biology^{TM}; Humana Press: Totowa, NJ, 2009; Vol. 570, pp 77–153. https://doi.org/10.1007/978-1-60327-394-7_4.

(190) Melse, O.; Hecht, S.; Antes, I. DynaBiS : A Hierarchical Sampling Algorithm to Identify Flexible Binding Sites for Large Ligands and Peptides. *Proteins Struct. Funct. Bioinforma.* **2022**, *90* (1), 18–32. https://doi.org/10.1002/prot.26182.

(191) Morris, G. M.; Huey, R.; Olson, A. J. Using AutoDock for Ligand-Receptor Docking. *Curr. Protoc. Bioinforma.* **2008**, *24* (1). https://doi.org/10.1002/0471250953.bi0814s24.

(192) Pacholczyk, M.; Kimmel, M. Exploring the Landscape of Protein-Ligand Interaction Energy Using Probabilistic Approach. *J. Comput. Biol.* **2011**, *18* (6), 843–850. https://doi.org/10.1089/cmb.2010.0017.

(193) Pashameah, R. A.; Soltane, R.; Sayed, A. M. A Novel Inhibitor of SARS-CoV Infection: Lactulose Octasulfate Interferes with ACE2-Spike Protein Binding. *Heliyon* **2024**, *10* (1), e23222. https://doi.org/10.1016/j.heliyon.2023.e23222.

(194) Paskaleva, E. E.; Xue, J.; Lee, D. Y.-W.; Shekhtman, A.; Canki, M. Palmitic Acid Analogs Exhibit Nanomolar Binding Affinity for the HIV-1 CD4 Receptor and Nanomolar Inhibition of Gp120-to-CD4 Fusion. *PLoS ONE* **2010**, *5* (8), e12168. https://doi.org/10.1371/journal.pone.0012168.

(195) Paul, B. K.; Guchhait, N. Modulation of Prototropic Activity and Rotational Relaxation Dynamics of a Cationic Biological Photosensitizer within the Motionally Constrained Bio-Environment of a Protein. *J. Phys. Chem. B* **2011**, *115* (34), 10322–10334. https://doi.org/10.1021/jp2015275.

(196) Paul, B. K.; Ray, D.; Guchhait, N. Spectral Deciphering of the Interaction between an Intramolecular Hydrogen Bonded ESIPT Drug, 3,5-Dichlorosalicylic Acid, and a Model Transport Protein. *Phys. Chem. Chem. Phys.* **2012**, *14* (25), 8892. https://doi.org/10.1039/c2cp23496c.

(197) Raza, S.; Abbas, G.; Azam, S. S. Screening Pipeline for Flavivirus Based Inhibitors for Zika Virus NS1. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17* (5), 1751–1761. https://doi.org/10.1109/TCBB.2019.2911081.

(198) Rentzsch, R.; Renard, B. Y. Docking Small Peptides Remains a Great Challenge: An Assessment Using AutoDock Vina. *Brief. Bioinform.* **2015**, *16* (6), 1045–1056. https://doi.org/10.1093/bib/bbv008.

(199) Rogaski, B.; Lim, J. B.; Klauda, J. B. Sterol Binding and Membrane Lipid Attachment to the Osh4 Protein of Yeast. *J. Phys. Chem. B* **2010**, *114* (42), 13562–13573. https://doi.org/10.1021/jp106890e.

(200) Santos, L. H. S.; Ferreira, R. S.; Caffarena, E. R. Integrating Molecular Docking and Molecular Dynamics Simulations. In *Docking Screens for Drug Discovery*; De Azevedo, W. F., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2019; Vol. 2053, pp 13–34. https://doi.org/10.1007/978-1-4939-9752-7_2.

(201) Scodeller, P.; Asciutto, E. K. Targeting Tumors Using Peptides. *Molecules* **2020**, *25* (4), 808. https://doi.org/10.3390/molecules25040808.

(202) Segura-Cabrera, A.; Tripathi, R.; Zhang, X.; Gui, L.; Chou, T.-F.; Komurov, K. A Structure- and Chemical Genomics-Based Approach for Repositioning of Drugs against VCP/P97 ATPase. *Sci. Rep.* **2017**, *7* (1), 44912. https://doi.org/10.1038/srep44912.

(203) Sharafdini, R.; Mosaddeghi, H. Inhibition of Insulin Amyloid Fibrillation by Salvianolic Acids and Calix[*n*]Arenes: Molecular Docking Insight. *J. Comput. Biophys. Chem.* **2021**, *20* (05), 539–555. https://doi.org/10.1142/S2737416521500332.

(204) Sheik Amamuddy, O.; Veldman, W.; Manyumwa, C.; Khairallah, A.; Agajanian, S.; Oluyemi, O.; Verkhivker, G. M.; Tastan Bishop, Ö. Integrated Computational Approaches and Tools for Allosteric Drug Discovery. *Int. J. Mol. Sci.* **2020**, *21* (3), 847. https://doi.org/10.3390/ijms21030847.

(205) Skolnick, J.; Brylinski, M. FINDSITE: A Combined Evolution/Structure-Based Approach to Protein Function Prediction. *Brief. Bioinform.* **2009**, *10* (4), 378–391. https://doi.org/10.1093/bib/bbp017.

(206) Tasso, B.; Canu Boido, C.; Terranova, E.; Gotti, C.; Riganti, L.; Clementi, F.; Artali, R.; Bombieri, G.; Meneghetti, F.; Sparatore, F. Synthesis, Binding, and Modeling Studies of New Cytisine Derivatives, as Ligands for Neuronal Nicotinic Acetylcholine Receptor Subtypes. *J. Med. Chem.* **2009**, *52* (14), 4345–4357. https://doi.org/10.1021/jm900225j.

(207) Ugurlu, S. Y.; McDonald, D.; Lei, H.; Jones, A. M.; Li, S.; Tong, H. Y.; Butler, M. S.; He, S. Cobdock: An Accurate and Practical Machine Learning-Based Consensus Blind Docking Method. *J. Cheminformatics* **2024**, *16* (1), 5. https://doi.org/10.1186/s13321-023-00793-x.

(208) Xiang, Y.; Zhai, G.; Li, Y.; Wang, M.; Chen, X.; Wang, R.; Xie, H.; Zhang, W.; Ge, G.; Zhang, Q.; Xu, Y.; Caflisch, A.; Xu, J.; Chen, H.; Chen, L. Ginkgolic Acids Inhibit SARS-CoV-2 and Its Variants by Blocking the Spike Protein/ACE2 Interplay. *Int. J. Biol. Macromol.* **2023**, *226*, 780–792. https://doi.org/10.1016/j.ijbiomac.2022.12.057.

(209) Żołek, T.; Dömötör, O.; Rezler, M.; Enyedy, É. A.; Maciejewska, D. Deposition of Pentamidine Analogues in the Human Body – Spectroscopic and Computational Approaches. *Eur. J. Pharm. Sci.* **2021**, *161*, 105779. https://doi.org/10.1016/j.ejps.2021.105779.

(210) Zubrzycki, I. Z.; Borcz, A.; Wiacek, M.; Hagner, W. The Studies on Substrate, Product and Inhibitor Binding to a Wild-Type and Neuronopathic Form of Human Acid-β-Glucosidase. *J. Mol. Model.* **2007**, *13* (11), 1133–1139. https://doi.org/10.1007/s00894-007-0232-5.

(211) Hocker, H. J.; Rambahal, N.; Gorfe, A. A. LIBSA – A Method for the Determination of Ligand-Binding Preference to Allosteric Sites on Receptor Ensembles. *J. Chem. Inf. Model.* **2014**, *54* (2), 530–538. https://doi.org/10.1021/ci400474u.

(212) Whalen, K. L.; Tussey, K. B.; Blanke, S. R.; Spies, M. A. Nature of Allosteric Inhibition in Glutamate Racemase: Discovery and Characterization of a Cryptic Inhibitory Pocket Using Atomistic MD Simulations and p K_a Calculations. *J. Phys. Chem. B* **2011**, *115* (13), 3416–3424. https://doi.org/10.1021/jp201037t.

(213) García-Sosa, A. T.; Sild, S.; Maran, U. Design of Multi-Binding-Site Inhibitors, Ligand Efficiency, and Consensus Screening of Avian Influenza H5N1 Wild-Type Neuraminidase and of the Oseltamivir-Resistant H274Y Variant. *J. Chem. Inf. Model.* **2008**, *48* (10), 2074–2080. https://doi.org/10.1021/ci800242z.

(214) Bugatti, A.; Giagulli, C.; Urbinati, C.; Caccuri, F.; Chiodelli, P.; Oreste, P.; Fiorentini, S.; Orro, A.; Milanesi, L.; D'Ursi, P.; Caruso, A.; Rusnati, M. Molecular Interaction Studies of HIV-1 Matrix Protein P17 and Heparin. *J. Biol. Chem.* **2013**, *288* (2), 1150–1161. https://doi.org/10.1074/jbc.M112.400077.

(215) Roumenina, L.; Bureeva, S.; Kantardjiev, A.; Karlinsky, D.; Andia-Pravdivy, J. E.; Sim, R.; Kaplun, A.; Popov, M.; Kishore, U.; Atanasov, B. Complement C1q-Target Proteins Recognition Is Inhibited by Electric Moment Effectors. *J. Mol. Recognit.* **2007**, *20* (5), 405–415. https://doi.org/10.1002/jmr.853.

(216) Agarwal, T.; Annamalai, N.; Khursheed, A.; Maiti, T. K.; Arsad, H. B.; Siddiqui, M. H. Molecular Docking and Dynamic Simulation Evaluation of Rohinitib — Cantharidin Based Novel HSF1 Inhibitors for Cancer Therapy. *J. Mol. Graph. Model.* **2015**, *61*, 141–149. https://doi.org/10.1016/j.jmgm.2015.07.003.

(217) Kovács, M.; Tóth, J.; Hetényi, C.; Málnási-Csizmadia, A.; Sellers, J. R. Mechanism of Blebbistatin Inhibition of Myosin II. *J. Biol. Chem.* **2004**, *279* (34), 35557–35563. https://doi.org/10.1074/jbc.M405319200.

(218) Aguayo-Ortiz, R.; Dominguez, L. Unveiling the Possible Oryzalin-Binding Site in the α-Tubulin of *Toxoplasma Gondii*. *ACS Omega* **2022**, *7* (22), 18434–18442. https://doi.org/10.1021/acsomega.2c00729.

(219) Aguayo-Ortiz, R.; Guzmán-Ocampo, D. C.; Dominguez, L. Insights into the Binding of Morin to Human γD-Crystallin. *Biophys. Chem.* **2022**, *282*, 106750. https://doi.org/10.1016/j.bpc.2021.106750.

(220) Evans, D. J.; Yovanno, R. A.; Rahman, S.; Cao, D. W.; Beckett, M. Q.; Patel, M. H.; Bandak, A. F.; Lau, A. Y. Finding Druggable Sites in Proteins Using TACTICS. *J. Chem. Inf. Model.* **2021**, *61* (6), 2897–2910. https://doi.org/10.1021/acs.jcim.1c00204.

(221) Yepes-Molina, L.; Teruel, J. A.; Johanson, U.; Carvajal, M. Brassica Oleracea L. Var. Italica Aquaporin Reconstituted Proteoliposomes as Nanosystems for Resveratrol Encapsulation. *Int. J. Mol. Sci.* **2024**, *25* (4), 1987. https://doi.org/10.3390/ijms25041987.

(222) Paulsen, C. E.; Armache, J.-P.; Gao, Y.; Cheng, Y.; Julius, D. Structure of the TRPA1 Ion Channel Suggests Regulatory Mechanisms. *Nature* **2015**, *520* (7548), 511–517. https://doi.org/10.1038/nature14367.

(223) Liu, C.; Reese, R.; Vu, S.; Rougé, L.; Shields, S. D.; Kakiuchi-Kiyota, S.; Chen, H.; Johnson, K.; Shi, Y. P.; Chernov-Rogan, T.; Greiner, D. M. Z.; Kohli, P. B.; Hackos, D.; Brillantes, B.; Tam, C.; Li, T.; Wang, J.; Safina, B.; Magnuson, S.; Volgraf, M.; Payandeh, J.; Zheng, J.; Rohou, A.; Chen, J. A Non-Covalent Ligand Reveals Biased Agonism of the TRPA1 Ion Channel. *Neuron* **2021**, *109* (2), 273-284.e4. https://doi.org/10.1016/j.neuron.2020.10.014.

(224) De Logu, F.; Nassini, R.; Materazzi, S.; Carvalho Gonçalves, M.; Nosi, D.; Rossi Degl'Innocenti, D.; Marone, I. M.; Ferreira, J.; Li Puma, S.; Benemei, S.; Trevisan, G.; Souza Monteiro De Araújo, D.; Patacchini, R.; Bunnett, N. W.; Geppetti, P. Schwann Cell TRPA1 Mediates Neuroinflammation That Sustains Macrophage-Dependent Neuropathic Pain in Mice. *Nat. Commun.* **2017**, *8* (1), 1887. https://doi.org/10.1038/s41467-017-01739-2.

(225) Takaya, J.; Mio, K.; Shiraishi, T.; Kurokawa, T.; Otsuka, S.; Mori, Y.; Uesugi, M. A Potent and Site-Selective Agonist of TRPA1. *J. Am. Chem. Soc.* **2015**, *137* (50), 15859–15864. https://doi.org/10.1021/jacs.5b10162.

(226) Pozsgai, G.; Bátai, I. Z.; Pintér, E. Effects of Sulfide and Polysulfides Transmitted by Direct or Signal Transduction-mediated Activation of TRPA1 Channels. *Br. J. Pharmacol.* **2019**, *176* (4), 628–645. https://doi.org/10.1111/bph.14514.

(227) Latorre, R.; Díaz-Franulic, I. Profile of David Julius and Ardem Patapoutian: 2021 Nobel Laureates in Physiology or Medicine. *Proc. Natl. Acad. Sci.* **2022**, *119* (1), e2121015119. https://doi.org/10.1073/pnas.2121015119.

(228) Allingham, J. S.; Smith, R.; Rayment, I. The Structural Basis of Blebbistatin Inhibition and Specificity for Myosin II. *Nat. Struct. Mol. Biol.* **2005**, *12* (4), 378–379. https://doi.org/10.1038/nsmb908.

(229) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–Peptide Docking: Opportunities and Challenges. *Drug Discov. Today* **2018**, *23* (8), 1530–1537. https://doi.org/10.1016/j.drudis.2018.05.006.

(230) Rentzsch, R.; Renard, B. Y. Docking Small Peptides Remains a Great Challenge: An Assessment Using AutoDock Vina. *Brief. Bioinform.* **2015**, *16* (6), 1045–1056. https://doi.org/10.1093/bib/bbv008.

(231) Castro-Alvarez, A.; Costa, A.; Vilarrasa, J. The Performance of Several Docking Programs at Reproducing Protein–Macrolide-Like Crystal Structures. *Molecules* **2017**, *22* (1), 136. https://doi.org/10.3390/molecules22010136.

(232) Shvedunova, M.; Akhtar, A. Modulation of Cellular Processes by Histone and Non-Histone Protein Acetylation. *Nat. Rev. Mol. Cell Biol.* **2022**, *23* (5), 329–349. https://doi.org/10.1038/s41580-021-00441-y.

(233) Strahl, B. D.; Allis, C. D. The Language of Covalent Histone Modifications. *Nature* **2000**, *403* (6765), 41–45. https://doi.org/10.1038/47412.

(234) Shakespear, M. R.; Halili, M. A.; Irvine, K. M.; Fairlie, D. P.; Sweet, M. J. Histone Deacetylases as Regulators of Inflammation and Immunity. *Trends Immunol.* **2011**, *32* (7), 335–343. https://doi.org/10.1016/j.it.2011.04.001.

(235) Fraga, M. F.; Ballestar, E.; Villar-Garea, A.; Boix-Chornet, M.; Espada, J.; Schotta, G.; Bonaldi, T.; Haydon, C.; Ropero, S.; Petrie, K.; Iyer, N. G.; Pérez-Rosado, A.; Calvo, E.; Lopez, J. A.; Cano, A.; Calasanz, M. J.; Colomer, D.; Piris, M. Á.; Ahn, N.; Imhof, A.; Caldas, C.; Jenuwein, T.; Esteller, M. Loss of Acetylation at Lys16 and Trimethylation at Lys20 of Histone H4 Is a Common Hallmark of Human Cancer. *Nat. Genet.* **2005**, *37* (4), 391–400. https://doi.org/10.1038/ng1531.

(236) Peng, Y.; Li, S.; Landsman, D.; Panchenko, A. R. Histone Tails as Signaling Antennas of Chromatin. *Curr. Opin. Struct. Biol.* **2021**, *67*, 153–160. https://doi.org/10.1016/j.sbi.2020.10.018.

(237) Chignola, F.; Gaetani, M.; Rebane, A.; Org, T.; Mollica, L.; Zucchelli, C.; Spitaleri, A.; Mannella, V.; Peterson, P.; Musco, G. The Solution Structure of the First PHD Finger of Autoimmune Regulator in Complex with Non-Modified Histone H3 Tail Reveals the Antagonistic Role of H3R2 Methylation. *Nucleic Acids Res.* **2009**, *37* (9), 2951–2961. https://doi.org/10.1093/nar/gkp166.

(238) Robertson, M. J.; Meyerowitz, J. G.; Panova, O.; Borrelli, K.; Skiniotis, G. Plasticity in Ligand Recognition at Somatostatin Receptors. *Nat. Struct. Mol. Biol.* **2022**, *29* (3), 210–217. https://doi.org/10.1038/s41594-022-00727-5.

(239) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447. https://doi.org/10.1021/ct700301q.

(240) Abdulkareem, S. M.; Housaindokht, M. R.; Bozorgmehr, M. R. The Effect of PC20:0 and Di-C7-PC Amphiphilic Surfactants on the Aggregation of Aβ1–40 and Aβ1–42 Using Molecular Dynamics Simulation. *J. Iran. Chem. Soc.* **2023**, *20* (6), 1357–1370. https://doi.org/10.1007/s13738-023-02761-6.

(241) Chen, W.; He, H.; Wang, J.; Wang, J.; Chang, C. A. Uncovering Water Effects in Protein–Ligand Recognition: Importance in the Second Hydration Shell and Binding Kinetics. *Phys. Chem. Chem. Phys.* **2023**, *25* (3), 2098–2109. https://doi.org/10.1039/D2CP04584B.

(242) Fischer, N. M.; Polêto, M. D.; Steuer, J.; van der Spoel, D. Influence of Na+ and Mg2+ Ions on RNA Structures Studied with Molecular Dynamics Simulations. *Nucleic Acids Res.* **2018**, *46* (10), 4872–4882. https://doi.org/10.1093/nar/gky221.

(243) Kunstmann, S.; Engström, O.; Wehle, M.; Widmalm, G.; Santer, M.; Barbirz, S. Increasing the Affinity of an O-Antigen Polysaccharide Binding Site in *Shigella Flexneri* Bacteriophage Sf6 Tailspike Protein. *Chem. – Eur. J.* **2020**, *26* (32), 7263–7273. https://doi.org/10.1002/chem.202000495.

(244) Kunstmann, S.; Gohlke, U.; Broeker, N. K.; Roske, Y.; Heinemann, U.; Santer, M.; Barbirz, S. Solvent Networks Tune Thermodynamics of Oligosaccharide Complex Formation in an Extended Protein Binding Site. *J. Am. Chem. Soc.* **2018**, *140* (33), 10447–10455. https://doi.org/10.1021/jacs.8b03719.

(245) Pradhan, M. R.; Nguyen, M. N.; Kannan, S.; Fox, S. J.; Kwoh, C. K.; Lane, D. P.; Verma, C. S. Characterization of Hydration Properties in Structural Ensembles of Biomolecules. *J. Chem. Inf. Model.* **2019**, *59* (7), 3316–3329. https://doi.org/10.1021/acs.jcim.8b00453.

(246) Van Der Spoel, D.; Zhang, J.; Zhang, H. Quantitative Predictions from Molecular Simulations Using Explicit or Implicit Interactions. *WIREs Comput. Mol. Sci.* **2022**, *12* (1), e1560. https://doi.org/10.1002/wcms.1560.

(247) Yoon, H. R.; Park, G. J.; Balupuri, A.; Kang, N. S. TWN-FS Method: A Novel Fragment Screening Method for Drug Discovery. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 4683–4696. https://doi.org/10.1016/j.csbj.2023.09.037.

(248) Zhou, Y.; Chen, S.-J. Graph Deep Learning Locates Magnesium Ions in RNA. *QRB Discov.* **2022**, *3*, e20. https://doi.org/10.1017/qrd.2022.17.

(249) Mandala, V. S.; McKay, M. J.; Shcherbakov, A. A.; Dregni, A. J.; Kolocouris, A.; Hong, M. Structure and Drug Binding of the SARS-CoV-2 Envelope Protein Transmembrane Domain in Lipid Bilayers. *Nat. Struct. Mol. Biol.* **2020**, *27* (12), 1202–1208. https://doi.org/10.1038/s41594-020-00536-8.

(250) Holger Gohlke. *Protein-Ligand Interactions*; Methods and Principles in Medicinal Chemistry; Wiley, 2012; Vol. 53.

(251) Christophe Chipot; Andrew Pohorille. *Free Energy Calculations*; Springer Series in Chemical Physics; Springer, 2007; Vol. 86.

(252) Klebe, G. Applying Thermodynamic Profiling in Lead Finding and Optimization. *Nat. Rev. Drug Discov.* **2015**, *14* (2), 95–110. https://doi.org/10.1038/nrd4486.

(253) Geschwindner, S.; Ulander, J.; Johansson, P. Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip? *J. Med. Chem.* **2015**, *58* (16), 6321–6335. https://doi.org/10.1021/jm501511f.

(254) Ferenczy, G. G.; Keserű, G. M. Thermodynamics Guided Lead Discovery and Optimization. *Drug Discov. Today* **2010**, *15* (21–22), 919–932. https://doi.org/10.1016/j.drudis.2010.08.013.

(255) Freire, E. Do Enthalpy and Entropy Distinguish First in Class from Best in Class? *Drug Discov. Today* **2008**, *13* (19–20), 869–874. https://doi.org/10.1016/j.drudis.2008.07.005.

(256) Ferenczy, G. G.; Keseru, G. M. Enthalpic Efficiency of Ligand Binding. *J. Chem. Inf. Model.* **2010**, *50* (9), 1536–1541. https://doi.org/10.1021/ci100125a.

(257) Velazquez-Campoy, A.; Todd, M. J.; Freire, E. HIV-1 Protease Inhibitors: Enthalpic versus Entropic Optimization of the Binding Affinity. *Biochemistry* **2000**, *39* (9), 2201–2207. https://doi.org/10.1021/bi992399d.

(258) Muzammil, S.; Armstrong, A. A.; Kang, L. W.; Jakalian, A.; Bonneau, P. R.; Schmelmer, V.; Amzel, L. M.; Freire, E. Unique Thermodynamic Response of Tipranavir to Human Immunodeficiency Virus Type 1 Protease Drug Resistance Mutations. *J. Virol.* **2007**, *81* (10), 5144–5154. https://doi.org/10.1128/JVI.02706-06.

(259) Ohtaka, H.; Freire, E. Adaptive Inhibitors of the HIV-1 Protease. *Prog. Biophys. Mol. Biol.* **2005**, *88* (2), 193–208. https://doi.org/10.1016/j.pbiomolbio.2004.07.005.

(260) Zeilinger, M.; Pichler, F.; Nics, L.; Wadsak, W.; Spreitzer, H.; Hacker, M.; Mitterhauser, M. New Approaches for the Reliable in Vitro Assessment of Binding Affinity Based on High-Resolution Real-Time Data Acquisition of Radioligand-Receptor Binding Kinetics. *EJNMMI Res.* **2017**, *7* (1), 22. https://doi.org/10.1186/s13550-016-0249-9.

(261) Rinken, A.; Lavogina, D.; Kopanchuk, S. Assays with Detection of Fluorescence Anisotropy: Challenges and Possibilities for Characterizing Ligand Binding to GPCRs. *Trends Pharmacol. Sci.* **2018**, *39* (2), 187–199. https://doi.org/10.1016/j.tips.2017.10.004.

(262) Pollard, T. D. A Guide to Simple and Informative Binding Assays. *Mol. Biol. Cell* **2010**, *21* (23), 4061–4067. https://doi.org/10.1091/mbc.e10-08-0683.

(263) Tonge, P. J. Quantifying the Interactions between Biomolecules: Guidelines for Assay Design and Data Analysis. *ACS Infect. Dis.* **2019**, *5* (6), 796–808. https://doi.org/10.1021/acsinfecdis.9b00012.

(264) Cooper, M. A. Optical Biosensors in Drug Discovery. *Nat. Rev. Drug Discov.* **2002**, *1* (7), 515–528. https://doi.org/10.1038/nrd838.

(265) Bastos, M.; Abian, O.; Johnson, C. M.; Ferreira-da-Silva, F.; Vega, S.; Jimenez-Alesanco, A.; Ortega-Alarcon, D.; Velazquez-Campoy, A. Isothermal Titration Calorimetry. *Nat. Rev. Methods Primer* **2023**, *3* (1), 17. https://doi.org/10.1038/s43586-023-00199-x.

(266) Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The Thermodynamics of Protein–Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **2008**, *384* (4), 1002–1017. https://doi.org/10.1016/j.jmb.2008.09.073.

(267) Ladbury, J. E.; Klebe, G.; Freire, E. Adding Calorimetric Data to Decision Making in Lead Discovery: A Hot Tip. *Nat. Rev. Drug Discov.* **2010**, *9* (1), 23–27. https://doi.org/10.1038/nrd3054.

(268) Baranauskienė, L.; Petrikaitė, V.; Matulienė, J.; Matulis, D. Titration Calorimetry Standards and the Precision of Isothermal Titration Calorimetry Data. *Int. J. Mol. Sci.* **2009**, *10* (6), 2752–2762. https://doi.org/10.3390/ijms10062752.

(269) Tellinghuisen, J.; Chodera, J. D. Systematic Errors in Isothermal Titration Calorimetry: Concentrations and Baselines. *Anal. Biochem.* **2011**, *414* (2), 297–299. https://doi.org/10.1016/j.ab.2011.03.024.

(270) Jarmoskaite, I.; AlSadhan, I.; Vaidyanathan, P. P.; Herschlag, D. How to Measure and Evaluate Binding Affinities. *eLife* **2020**, *9*, e57264. https://doi.org/10.7554/eLife.57264.

(271) Fountoulakis, M.; Lahm, H.-W. Hydrolysis and Amino Acid Composition Analysis of Proteins. *J. Chromatogr. A* **1998**, *826* (2), 109–134. https://doi.org/10.1016/S0021-9673(98)00721-3.

(272) Reinmuth-Selzle, K.; Tchipilov, T.; Backes, A. T.; Tscheuschner, G.; Tang, K.; Ziegler, K.; Lucas, K.; Pöschl, U.; Fröhlich-Nowoisky, J.; Weller, M. G. Determination of the Protein Content of Complex Samples by Aromatic Amino Acid Analysis, Liquid Chromatography-UV Absorbance, and Colorimetry. *Anal. Bioanal. Chem.* **2022**, *414* (15), 4457–4470. https://doi.org/10.1007/s00216-022-03910-1.

(273) King, E.; Aitchison, E.; Li, H.; Luo, R. Recent Developments in Free Energy Calculations for Drug Discovery. *Front. Mol. Biosci.* **2021**, *8*, 712085. https://doi.org/10.3389/fmolb.2021.712085.

(274) Chipot, C. Frontiers in Free-Energy Calculations of Biological Systems. *WIREs Comput. Mol. Sci.* **2014**, *4* (1), 71–89. https://doi.org/10.1002/wcms.1157.

(275) Decherchi, S.; Cavalli, A. Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chem. Rev.* **2020**, *120* (23), 12788–12833. https://doi.org/10.1021/acs.chemrev.0c00534.

(276) Donald A. McQuarrie. *Statistical Mechanics*; University Science Books: California, 2000.

(277) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(278) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300–313. https://doi.org/10.1063/1.1749657.

(279) Aqvist, J.; Marelius, J. The Linear Interaction Energy Method for Predicting Ligand Binding Free Energies. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 613–626. https://doi.org/10.2174/1386207013330661.

(280) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, *91* (1–3), 275–282. https://doi.org/10.1016/0010-4655(95)00053-I.

(281) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690.

(282) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**, *114* (32), 10235–10253. https://doi.org/10.1021/jp102971x.

(283) Procacci, P. Methodological Uncertainties in Drug-Receptor Binding Free Energy Predictions Based on Classical Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2021**, *67*, 127–134. https://doi.org/10.1016/j.sbi.2020.08.001.

(284) Feng, M.; Heinzelmann, G.; Gilson, M. K. Absolute Binding Free Energy Calculations Improve Enrichment of Actives in Virtual Compound Screening. *Sci. Rep.* **2022**, *12* (1), 13640. https://doi.org/10.1038/s41598-022-17480-w.

(285) Wahl, J.; Smieško, M. Assessing the Predictive Power of Relative Binding Free Energy Calculations for Test Cases Involving Displacement of Binding Site Water Molecules. *J. Chem. Inf. Model.* **2019**, *59* (2), 754–765. https://doi.org/10.1021/acs.jcim.8b00826.

(286) Kim A. Sharp. Statistical Thermodynamics of Binding and Molecular Recognition Models. In *Protein-Ligand Interactions*; Methods and Principles in Medicinal Chemistry; Wiley, 2012; Vol. 53, pp 3–22.

(287) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem. Int. Ed.* **2002**, *41* (15), 2644–2676. https://doi.org/10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O.

(288) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489. https://doi.org/10.1006/jmbi.1996.0477.

(289) Bohm, H.-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8* (3), 243–256. https://doi.org/10.1007/BF00126743.

(290) Siebenmorgen, T.; Zacharias, M. Computational Prediction of Protein–Protein Binding Affinities. *WIREs Comput. Mol. Sci.* **2020**, *10* (3), e1448. https://doi.org/10.1002/wcms.1448.

(291) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119* (16), 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055.

(292) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.

(293) Childers, M. C.; Daggett, V. Validating Molecular Dynamics Simulations against Experimental Observables in Light of Underlying Conformational Ensembles. *J. Phys. Chem. B* **2018**, *122* (26), 6673–6689. https://doi.org/10.1021/acs.jpcb.8b02144.

(294) Par Soderhjelm; Samuel Genheden; Ulf Ryde. Quantum Mechanics in Structure-Based Ligand Design. In *Protein-ligand interactions*; Methods and Principles in Medicinal Chemistry; Wiley, 2012; Vol. 53, pp 121–143.

(295) Maia, J. D. C.; Urquiza Carvalho, G. A.; Mangueira, C. P.; Santana, S. R.; Cabral, L. A. F.; Rocha, G. B. GPU Linear Algebra Libraries and GPGPU Programming for Accelerating MOPAC Semiempirical Quantum Chemistry Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3072–3081. https://doi.org/10.1021/ct3004645.

(296) Liu, J.; Wan, J.; Ren, Y.; Shao, X.; Xu, X.; Rao, L. DOX_BDW: Incorporating Solvation and Desolvation Effects of Cavity Water into Nonfitting Protein–Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2023**, *63* (15), 4850–4863. https://doi.org/10.1021/acs.jcim.3c00776.

(297) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M. The Role of Quantum Mechanics in Structure-Based Drug Design. *Drug Discov. Today* **2007**, *12* (17–18), 725–731. https://doi.org/10.1016/j.drudis.2007.07.006.

(298) Cavalli, A.; Carloni, P.; Recanatini, M. Target-Related Applications of First Principles Quantum Chemical Methods in Drug Design. *Chem. Rev.* **2006**, *106* (9), 3497–3519. https://doi.org/10.1021/cr050579p.

(299) Ryde, U.; Söderhjelm, P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem. Rev.* **2016**, *116* (9), 5520–5566. https://doi.org/10.1021/acs.chemrev.5b00630.

(300) Hu, L.; Söderhjelm, P.; Ryde, U. On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011**, *7* (3), 761–777. https://doi.org/10.1021/ct100530r.

(301) Siegbahn, P. E. M.; Borowski, T. Modeling Enzymatic Reactions Involving Transition Metals. *Acc. Chem. Res.* **2006**, *39* (10), 729–738. https://doi.org/10.1021/ar050123u.

(302) Himo, F. Quantum Chemical Modeling of Enzyme Active Sites and Reaction Mechanisms. *Theor. Chem. Acc.* **2006**, *116* (1–3), 232–240. https://doi.org/10.1007/s00214-005-0012-1.

(303) Friesner, R. A.; Guallar, V. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (qm/mm) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* **2005**, *56* (1), 389–427. https://doi.org/10.1146/annurev.physchem.55.091602.094410.

(304) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48* (7), 1198–1229. https://doi.org/10.1002/anie.200802019.

(305) Lin, H.; Truhlar, D. G. QM/MM: What Have We Learned, Where Are We, and Where Do We Go from Here? *Theor. Chem. Acc.* **2007**, *117* (2), 185. https://doi.org/10.1007/s00214-006-0143-z.

(306) Söderhjelm, P.; Aquilante, F.; Ryde, U. Calculation of Protein−Ligand Interaction Energies by a Fragmentation Approach Combining High-Level Quantum Chemistry with Classical Many-Body Effects. *J. Phys. Chem. B* **2009**, *113* (32), 11085–11094. https://doi.org/10.1021/jp810551h.

(307) Klähn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. On Possible Pitfalls in Ab Initio QM/MM Minimization Approaches For Studies of Enzymatic Reactions. **2006**.

(308) Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simul.* **1993**, *10* (2–6), 97–120. https://doi.org/10.1080/08927029308022161.

(309) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* **1996**, *10* (4), 293–304. https://doi.org/10.1007/BF00124499.

(310) Laederach, A.; Reilly, P. J. Specific Empirical Free Energy Function for Automated Docking of Carbohydrates to Proteins. *J. Comput. Chem.* **2003**, *24* (14), 1748–1757. https://doi.org/10.1002/jcc.10288.

(311) Niño, H.; García-Pintos, I.; Rodríguez-Borges, J. E.; Escobar-Cubiella, M.; García-Mera, X.; Prado-Prado, F. Review of Synthesis, Biological Assay and QSAR Studies of -Secretase Inhibitors. *Curr. Comput. Aided Drug Des.* **2011**, *7*, 263–275.

(312) Niño, H.; Rodríguez-Borges, J. E.; García-Mera, X.; Prado-Prado, F. Review of Synthesis, Assay, and Prediction of   and -Secretase Inhibitors. *Curr. Top. Med. Chem.* **2012**, *12*, 828–844.

(313) Prado-Prado, F.; Garcia, I. Review of Theoretical Studies for Prediction of Neurodegenerative Inhibitors. *Mini-Rev. Med. Chem.* **2012**, *12* (6), 452–466. https://doi.org/10.2174/138955712800493780.

(314) Prado-Prado, F.; Escobar-Cubiella, M.; García-Mera, X. Review of Bioinformatics and QSAR Studies of -Secretase Inhibitors. *Curr. Bioinforma.* **2011**, *6*, 3–15.

(315) Rajamani, R.; Good, A. C. Ranking Poses in Structure-Based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discov. Devel.* **2007**, *10* (3), 308–315.

(316) Seifert, M. H. J. Optimizing the Signal-to-Noise Ratio of Scoring Functions for Protein−Ligand Docking. *J. Chem. Inf. Model.* **2008**, *48* (3), 602–612. https://doi.org/10.1021/ci700345n.

(317) Seifert, M. H. J. Targeted Scoring Functions for Virtual Screening. *Drug Discov. Today* **2009**, *14* (11–12), 562–569. https://doi.org/10.1016/j.drudis.2009.03.013.

(318) Macchiarulo, A.; Nobeli, I.; Thornton, J. M. Ligand Selectivity and Competition between Enzymes in Silico. *Nat. Biotechnol.* **2004**, *22* (8), 1039–1045. https://doi.org/10.1038/nbt999.

(319) Simon, Z.; Peragovics, A.; Vigh-Smeller, M.; Csukly, G.; Tombor, L.; Yang, Z.; Zahoranszky-Kohalmi, G.; Vegner, L.; Jelinek, B.; Hari, P.; Hetenyi, C.; Bitter, I.; Czobor, P.; Malnasi-Csizmadia, A. Drug Effect Prediction by Polypharmacology-Based Interaction Profiling. *J. Chem. Inf. Model.* **2012**, *52* (1), 134–145. https://doi.org/10.1021/ci2002022.

(320) Peragovics, A.; Simon, Z.; Brandhuber, I.; Jelinek, B.; Hari, P.; Hetenyi, C.; Czobor, P.; Malnasi-Csizmadia, A. Contribution of 2D and 3D Structural Features of Drug Molecules in the Prediction of Drug Profile Matching. *J. Chem. Inf. Model.* **2012**, *52* (7), 1733–1744. https://doi.org/10.1021/ci3001056.

(321) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The Maximal Affinity of Ligands. *Proc. Natl. Acad. Sci.* **1999**, *96* (18), 9997–10002. https://doi.org/10.1073/pnas.96.18.9997.

(322) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72* (3), 1047–1069. https://doi.org/10.1016/S0006-3495(97)78756-3.

(323) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discov. Today* **2004**, *9* (10), 430–431. https://doi.org/10.1016/S1359-6446(04)03069-7.

(324) Abad-Zapatero, C.; Metz, J. Ligand Efficiency Indices as Guideposts for Drug Discovery. *Drug Discov. Today* **2005**, *10* (7), 464–469. https://doi.org/10.1016/S1359-6446(05)03386-6.

(325) Abad-Zapatero, C. Ligand Efficiency Indices for Effective Drug Discovery. *Expert Opin. Drug Discov.* **2007**, *2* (4), 469–488. https://doi.org/10.1517/17460441.2.4.469.

(326) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432* (7019), 855–861. https://doi.org/10.1038/nature03193.

(327) Bajorath, J.; Stenkamp, R.; Aruffo, A. Knowledge-based Model Building of Proteins: Concepts and Examples. *Protein Sci.* **1993**, *2* (11), 1798–1810. https://doi.org/10.1002/pro.5560021103.

(328) Williams, P. A.; Cosme, J.; Vinković, D. M.; Ward, A.; Angove, H. C.; Day, P. J.; Vonrhein, C.; Tickle, I. J.; Jhoti, H. Crystal Structures of Human Cytochrome P450 3A4 Bound to Metyrapone and Progesterone. *Science* **2004**, *305* (5684), 683–686. https://doi.org/10.1126/science.1099736.

(329) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-Based Docking Methods Fail to Generate Physically Valid Poses or Generalise to Novel Sequences. *Chem. Sci.* **2024**, *15* (9), 3130–3139. https://doi.org/10.1039/D3SC04185A.

(330) Dahlström, K. M.; Salminen, T. A. Apprehensions and Emerging Solutions in ML-Based Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2024**, *86*, 102819. https://doi.org/10.1016/j.sbi.2024.102819.

MELLÉKLET

D1

FEBS Letters 580 (2006) 1447–1450

Blind docking of drug-sized compounds to proteins with up to a thousand residues

Csaba Hetényi[a,*], David van der Spoel[b]

[a] *Department of Biochemistry, Eötvös Loránd University, 1/C Pázmány P. sétány, 1117 Budapest, Hungary*
[b] *Molecular Biophysics Group, Department of Cell and Molecular Biology, Uppsala University, Box 596, 75124 Uppsala, Sweden*

Received 14 November 2005; revised 20 December 2005; accepted 24 January 2006

Available online 31 January 2006

Edited by Robert B. Russell

Abstract Blind docking was introduced for the detection of possible binding sites and modes of peptide ligands by scanning the entire surface of protein targets. In the present study, the method is tested on a group of drug-sized compounds and proteins with up to a thousand amino acid residues. Both proteins from complex structures and ligand-free proteins were used as targets. Robustness, limitations and future perspectives of the method are discussed. It is concluded that blind docking can be used for unbiased mapping of the binding patterns of drug candidates.

Keywords: Active site; Binding pocket; Rational design; Pharmacophore; AutoDock; Molecular interaction

1. Introduction

In silico molecular docking is one of the most powerful techniques of structure-based drug design [1]. Most applications of docking tools focus on the (supposed) primary binding region. However, there are cases in which the information on the binding region is missing. The AutoDock [2]-based blind docking (BD) approach [3] was introduced previously to search the entire surface of proteins for binding sites while simultaneously optimizing the conformations of the peptides. The results of BD were regarded as ‘‘very encouraging’’ in a recent review [4]. BD [5–7] and the recommended search parameters [8–12] have been used for solving various problems such as design of inhibitors [5], comparison of microtubule-stabilizing agents [7] and exploring substrate binding modes [8]. Because of the apparent success of the approach [4–12], we decided to perform further systematic tests on a set of 43 ligand–protein complexes which was previously used in a comprehensive study on the selectivity of binding of aromatic compounds [13]. In the previous study [13] searching was restricted to the surrounding of the primary binding site (for a list of the complexes, refer to Supplementary material, Table A), here we use BD on the entire protein surfaces. The set contains drug-sized aromatic ligands with relatively few free rotations. All of these have some (positive or negative) biological effects and some (e.g., cancer drugs tamoxifen in 3ert and methotrexate in 4dfr) of them are actual medicines (see Fig. 1).

2. Methods

In the present study, the original parameters [3] of BD (Supplementary material, Table C) were used in combination with an evaluation scheme based on binding free energy (ΔG) and root mean square deviation (RMSD) calculated between the crystallographic and the docked ligand conformations (RMSD, Supplementary material, Scheme A). By definition, the entire protein surfaces were subjected to the BD search. For every 5th complexes of Table A (starting with the 1st row) and for the system with the largest protein (1b70) coordinates of the ligand-free proteins were obtained from the protein databank (PDB). In two cases, where the unbound proteins were not available (complexes 1a0q and 1gaf) the next systems (1a53 and 1guh, respectively) were involved in the study. The selected ten ligand-free proteins (Supplementary material, Table B) were superimposed on the corresponding protein–ligand complex structures and used for BD as described previously.

3. Results and discussion

3.1. Test of BD search on large protein targets

The results of the BD calculations on the 43 proteins (ligand-bound conformations, marked with the corresponding PDB codes, Table A) are summarized in Table 1 (for a detailed list of results, refer to Supplementary material, Table D). For 34 of 43 systems the BD search identified the crystallographic binding site and mode of ligands as the energy minimum of the whole BD job, i.e., all 100 docking trials (runs). In terms of averages and standard deviations (Table 1) the corresponding ranks contain energetically uniform members, with a significant population in most of the cases. In six of the remaining nine cases (1dy4, 1e7a, 1eqg, 1ivb, 1ngp, 3pcn) the native ligand position was ranked in the best 2nd–7th ranks and in three cases (1hz4, 1ju4, 1pth) an additional 1–3 accumulative BD jobs were necessary to locate the native binding mode (for details of accumulative BD refer to Supplementary material, Scheme A). In two out of the nine cases (1dy4, 1ngp) the native binding mode was also placed in the 2nd rank in the restricted docking study [13] indicating

[*] Corresponding author. Fax: +36 1 3812172.
E-mail address: csabahete@yahoo.com (C. Hetényi).

Abbreviations: BD, blind docking; PDB, protein databank; RMSD, root mean square deviation

0014-5793/$32.00 © 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2006.01.074

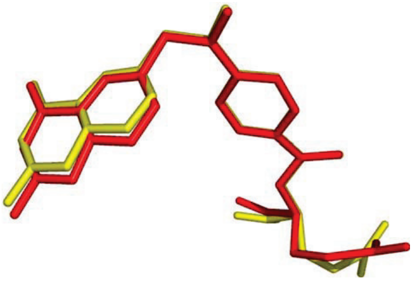


Fig. 1. The match of the crystallographic (red) and the minimum energy blind docked (yellow) conformations of methotrexate, the largest ligand molecule investigated (system 4dfr). The size of the ligand molecules did not affect the results of BD in the present study.

that the reason of these results is not the insufficient BD search. The average RMSD_{min} (RMSD corresponding to the energy minimum of the rank) of the 43 systems (1.0 ± 0.7 Å) is similar to the value calculated from the results of restricted docking [13] for the same set (1.2 ± 0.7 Å). This comparison shows, that in the case of drug-sized compounds, both the AutoDock scoring function and the Lamarckian genetic algorithm with the pseudo-Solis and Wets local search method can be applied to the large BD search space, i.e., the whole target surface solely by tuning the search parameters (Supplementary material, Table C). In the original BD study [3] the largest protein was 316 AA. In the present study, proteins with up to 1040 residues were involved in the calculations and 16 of 43 systems have more than 316 AAs. In seven of these 16 cases, including the largest protein investigated (Fig. 2) the native binding conformation was in the 1st rank, i.e., as the energy minimum of 100 trials. In the other nine cases the binding mode was correctly reproduced in terms of RMSD, but placed in higher ranks due to higher binding energy (for explanation, refer to Section 3.4).

3.2. Protein flexibility: robustness and limitations

BD to the 10 ligand-free protein structures (marked with U in Table 1 and Table D) provides additional information on the sensitivity of BD on protein flexibility. Such information may be useful for the situations, where only the unbound protein is available for the calculations, as expected for most real applications. In eight of the selected 10 cases the ranking of docked conformations with the best RMSD-s were identical or lower (better) compared to the results obtained for the corresponding proteins from complexes (previous section, Table 1) which demonstrates the robustness of BD. In two cases ($1\text{b}70^{\text{U}}$ and 1ivb^{U}) the best-RMSD-solution moved to higher ranks (rank serial numbers increased with 2 and 1, respectively). At $1\text{b}70^{\text{U}}$, a turn of 180° (respective to the 1b70 complex) of the amide group of a central glutamine residue spoiled the favorable H-bonding pattern with the ligand at the binding site (Fig. 2). This resulted in higher ΔG-s and higher ranking if compared with 1b70 (Table 1). However, the corresponding RMSD has not increased dramatically, due to the remaining (e.g., hydrophobic) interactions at the site. It should be remarked, that in these systems only moderate changes can be observed between the bound and ligand-free protein structures (see C_α-RMSD-s in Table B, Supplementary material). For these systems with moderate flexibility in the active site BD proved to be robust, but obviously BD alone may prove insufficient for systems with a higher degree of induced fit upon ligand binding. To overcome this problem, methods which handle structural flexibility [14] could be used in post-docking mode with the (prerequisite) binding positions and conformations of ligands found by BD as input.

3.3. Ligand flexibility

Neither the number of flexible torsions in the ligands (tabulated in Table A, Supplementary material), nor the size of the ligands affects the accuracy of the results of BD for the investigated systems (Fig. 1). The computational cost (efficiency) of the BD runs does depend on ligand flexibility. For systems with the smallest (1mpj) and largest (4dfr) ligand molecules BD runs took 5 and 22 min (Opteron 2 GHz), respectively.

3.4. Competition for the binding sites between the ligand and solvent molecules. Multiple binding sites

It should be remarked, that docking calculations generally use 'dry' protein molecules for the search, i.e., all ions, water molecules etc., are removed from the coordinate files before docking. In six out of nine cases where the native binding mode did not belong to the 1st rank, inspection of the original PBD files showed, that the low-energy binding sites of the first ranks found for the ligand during BD are occupied by water molecules (or other solvent) in the PDB structure. This can be due to the energetically favorable protein–solvent interactions at those sites, but it is also possible that the crystallographic complexes do not include all binding sites/modes of the ligands. In the systems (1ev3, 1mpj, 1qiz, 1tym) where insulin oligomers were used as targets in this study, multiple crystallographic binding sites at the protein interfaces were reproduced, showing the applicability of BD for multiple binding site search. Although some methods have been proposed for the modeling of ligand–solvent competition 'on-line', i.e., during docking simulations [15], or 'off-line' with mixed maps for the restricted search space [16], there is no trivial solution for BD yet. However, there is no alternative to using a dry target if multiple sites are searched for since water molecules covering the putative sites may hinder entrance of the ligand molecules.

3.5. Recommendations for BD of drugs

(1) In 3 cases (1hz4, 1ju4, 1pth) additional, accumulative BD jobs were necessary to find the native ligand conformation. In these cases the previously found representative ligand conformations (one per rank) were merged with the protein structure and these molecular complexes were used as docking targets in the next job. This procedure can be useful in BD calculations aimed at mapping all possible binding sites and can be automated by setting a limit criterion in terms of, e.g., binding free energy (Supplementary material, Scheme A). (2) In general, 0.55 Å grid spacing (Supplementary material, Table B) was adequate for the BD search of the drug-sized compounds in the present study to obtain acceptable RMSD-s. However, in one case (1ju4) a re-docking was performed for the located binding site with 0.375 Å grid spacing and the fit was refined from 4.136 to 0.629 Å (Supplementary material, Table D).

Table 1
Results of the blind docking calculations (abridged)

PDB	Job #	Rank #	ΔG_{min}	$\mathrm{RMSD}_{\mathrm{min}}$	Population	ΔG_{avg}	ΔG_{sdev}
1a0q	1	1	−9.16	2.212	16	−8.96	0.21
1a53	1	1	−10.03	0.646	50	−9.61	0.29
1a53$^{\mathrm{U}}$	1	1	−10.55	1.223	54	−9.91	0.45
1a8u	1	1	−6.48	0.439	100	−6.48	0.00
1alw	1	1	−6.41	2.658[a]	86	−6.23	0.10
1az8	1	1	−11.99	0.544	83	−11.48	0.23
1az8$^{\mathrm{U}}$	1	1	−11.26	0.987	82	−10.59	0.33
1b70	1	1	−8.72	0.891	42	−8.64	0.05
1b70$^{\mathrm{U}}$	1	3	−7.15	1.023	22	−7.06	0.06
1bzj	1	1	−12.88	0.567	100	−12.80	0.04
1c83	1	1	−11.30	0.541	100	−11.13	0.07
1c84	1	1	−10.45	0.862	93	−10.11	0.24
1c85	1	1	−9.96	0.741	100	−9.86	0.05
1c85$^{\mathrm{U}}$	1	1	−8.95	1.557	70	−8.76	0.20
1ca7	1	1	−7.89	0.854	91	−7.84	0.04
1d1q	1	1	−10.85	0.545	99	−10.75	0.08
1dy4	1	2	−8.79	0.777	13	−8.37	0.36
1e7a	1	2	−6.07	1.023	73	−6.03	0.03
1ecv	1	1	−11.24	0.674	100	−10.85	0.30
1ecv$^{\mathrm{U}}$	1	1	−8.64	1.166	25	−8.07	0.48
1eqg	1	4	−7.64	0.727	56	−7.59	0.02
1ev3	1	1	−4.95	1.075	10	−4.95	0.01
1f5k	1	1	−7.45	0.432	57	−7.45	0.00
1fiw	1	1	−9.00	0.832	100	−8.99	0.01
1gaf	1	1	−10.00	0.409	62	−9.46	0.39
1guh	1	1	−11.50	0.792	17	−10.58	0.66
1guh$^{\mathrm{U}}$	1	1	−11.01	1.180	22	−10.06	0.90
1hd2	1	1	−5.32	0.739	100	−5.31	0.01
1hdu	1	1	−8.60	0.525	68	−8.42	0.15
1hz4	2	3	−5.42	0.490	1	−5.42	–
1ivb	1	3	−6.46	0.200	26	−6.31	0.10
1ivb$^{\mathrm{U}}$	1	4	−5.47	2.717	28	−5.35	0.11
1ju4	3[b]	1	−5.09	0.629	7	−5.09	0.00
1kel	1	1	−12.25	1.932	55	−11.32	0.65
1mpj	1	1	−3.88	0.465	54	−3.87	0.01
1ngp	1	2	−7.35	0.691	36	−7.23	0.11
1pth	4	8	−3.95	2.450	3	−3.95	0.01
1pth$^{\mathrm{U}}$	3	4	−4.60	2.660	5	−4.59	0.01
1qiz	1	1	−4.63	2.481	25	−4.61	0.01
1rfn	1	1	−8.61	0.573	100	−8.60	0.00
1sri	1	1	−9.06	1.006	47	−8.63	0.27
1tnj	1	1	−7.47	1.964	84	−7.27	0.07
1tym	1	1	−5.89	1.830	71	−5.79	0.06
1tym$^{\mathrm{U}}$	1	1	−5.06	1.919	12	−5.01	0.06
2ay5	1	1	−9.50	2.085	22	−9.18	0.15
3cpa	1	1	−8.74	0.757	44	−8.23	0.22
3ert	1	1	−9.84	1.646	58	−9.38	0.23
3pax	1	1	−6.14	1.208	100	−6.02	0.05
3pcn	1	7	−5.11	2.568	12	−5.00	0.06
3pcn$^{\mathrm{U}}$	1	2	−5.24	2.658	3	−4.94	0.27
43ca	1	1	−5.17	0.419	100	−5.16	0.01
4dfr	1	1	−13.35	1.086	19	−12.54	0.93
4ts1	1	1	−6.94	0.504	76	−6.68	0.13

PDB, protein databank code; U, unbound (ligand-free) protein; Job #, number of the accumulative jobs; Rank #, serial number of the Rank; ΔG_{min}, the minimum of AutoDock free energy of binding (kcal/mol) values of the members of Rank; Population, population of the Rank (the maximum value is 100 corresponding to a docking job, i.e., 100 docking runs); $\mathrm{RMSD}_{\mathrm{min}}$, root mean square deviation (Å) of the conformation conjugated to ΔG_{min}. Averages (ΔG_{avg}) and standard deviations (ΔG_{sdev}) are calculated for the rank.

[a] The crystallographic ligand used for comparison has erroneous structure.

[b] In case of 1ju4, Job 3 was a re-docking with 0.375 Å grid spacing focused on the previously located (Job 2:Rank 2) binding site.

Such re-dockings are of limited computational cost (10 docking runs usually suffice) and can be recommended for all BD studies. (3) In general, post-docking refinement with, e.g., normal mode methods [14] accounting for protein flexibility at the docked complexes may be advantageous to increase precision of ranking.

3.6. Future applications of BD

In combination with experimental techniques such as site-directed mutagenesis, BD can be a useful tool for mapping of binding modes of drug candidates on protein targets and even the selection of new protein targets (protein screening [13]) for existing drugs.

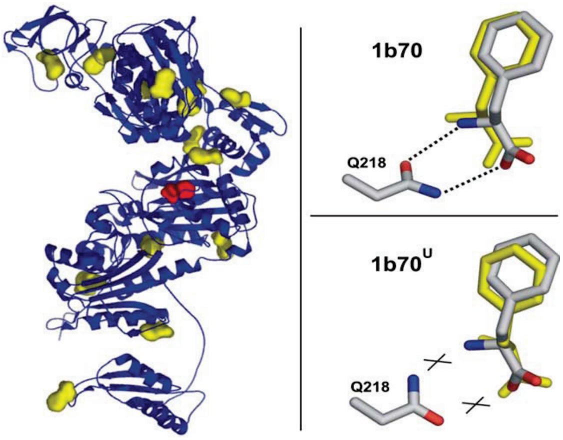


Fig. 2. The result of blind docking for phenylalanyl-tRNA synthase (blue cartoon, 1b70), a protein with more than a thousand amino acids. Representative ligand conformations of each rank and the crystallographic one are depicted as yellow and red surfaces, respectively (on the left). Due to the large protein surface, numerous putative sites can be found among which the crystallographic site was identified in the 1st or 3rd best ranks using the bound or ligand-free proteins as targets, respectively. The blind docked ligand conformations (yellow sticks) have good match with the crystallographic ligand conformation (sticks colored by atom type) if using either the bound (1b70, top on the right) or the ligand-free (1b70^U, bottom on the right) protein structures. In case of 1b70^U the amide group of the key H-bonding glutamine (Q218) residue is turned with ca. 180° hindering formation of the H-bonds (dotted lines) which exist in the complex form (1b70) and cause a higher ΔG value when docking to 1b70^U. Figures were prepared using PyMol [17].

Acknowledgment: This work was financed by the Eötvös Fellowship of the Hungarian State.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2006.01.074.

References

[1] Brooijmans, N. and Kuntz, I.D. (2003) Molecular recognition and docking algorithms. Annu. Rev. Biophys. Biomol. Struct. 32, 335–373.

[2] Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19, 1639–1662.

[3] Hetényi, C. and van der Spoel, D. (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci. 11, 1729–1737.

[4] Campbell, S.J., Gold, N.D., Jackson, R.M. and Westhead, D.R. (2003) Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. 13, 389–395.

[5] Jining, L., Makagiansar, I., Yusuf-Makagiansar, H., Chow, V.T.K., Siahaan, T.J. and Jois, S.D.S. (2004) Design, structure and biological activity of β-turn peptides of CD2 protein for inhibition of T-cell adhesion. Eur. J. Biochem. 271, 2873–2886.

[6] Brown, W.M. and Vander Jagt, D.L. (2004) Creating artificial binding pocket boundaries to improve the efficiency of flexible ligand docking. J. Chem. Inf. Comput. Sci. 44, 1412–1422.

[7] Pineda, O., Farràs, J., Maccari, L., Manetti, F., Botta, M. and Vilarrasa, J. (2004) Computational comparison of microtubule-stabilising agents laulimalide and peloruside with taxol and colchicine. Bioorg. Med. Chem. Lett. 14, 4825–4829.

[8] Bjelic, S. and Åqvist, J. (2004) Computational prediction of structure, substrate binding mode, mechanism, and rate for a malaria protease with a novel type of active site. Biochemistry 43, 14521–14528.

[9] Österberg, F. and Åqvist, J. (2005) Exploring blocker binding to a homology model of the open hERG K$^+$ channel using docking and molecular dynamics methods. FEBS Lett. 579, 2939–2944.

[10] Li, L., Geng, X., Yonkunas, M., Su, A., Densmore, E., Tang, P. and Drain, P. (2005) Ligand-dependent linkage of the ATP site to inhibition gate closure in the K$_{ATP}$ channel. J. Gen. Physiol. 126, 285–299.

[11] Alonso, H., Gillies, M.B., Cummins, P.L., Bliznyuk, A.A. and Gready, J.E. (2005) Multiple ligand-binding modes in bacterial R67 dihydrofolate reductase. J. Comput.-Aided Mol. Des. 19, 165–187.

[12] Yonkunas, M.J., Xu, Y. and Tang, P. (2005) Anesthetic interaction with ketosteroid isomerase: insights from molecular dynamics simulations. Biophys. J. 89, 2350–2356.

[13] Hetényi, C., Maran, U. and Karelson, M. (2003) A comprehensive docking study on the selectivity of binding of aromatic compounds to proteins. J. Chem. Inf. Comput. Sci. 43, 1576–1583.

[14] Lindahl, E. and Delarue, M. (2005) Refinement of docked protein–ligand and protein–DNA structures using low frequency normal mode amplitude optimization. Nucleic Acids Res. 33, 4496–4506.

[15] Rarey, M., Kramer, B. and Lengauer, T. (1999) The particle concept: placing discrete water molecules during protein–ligand docking predictions. Proteins 34, 17–28.

[16] Österberg, F., Morris, G.M., Sanner, M.F., Olson, A.J. and Goodsell, D.S. (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 46, 34–40.

[17] DeLano, W.L. (2002) PyMol Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA.

D2

Toward prediction of functional protein pockets using blind docking and pocket search algorithms

Csaba Hetényi[1,2]* and David van der Spoel[1]

[1]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

[2]Department of Genetics, Eötvös University, Budapest, Hungary

Received 11 November 2010; Revised 6 March 2011; Accepted 7 March 2011
DOI: 10.1002/pro.618
Published online 16 March 2011 proteinscience.org

Abstract: Location of functional binding pockets of bioactive ligands on protein molecules is essential in structural genomics and drug design projects. If the experimental determination of ligand-protein complex structures is complicated, blind docking (BD) and pocket search (PS) calculations can help in the prediction of atomic resolution binding mode and the location of the pocket of a ligand on the entire protein surface. Whereas the number of successful predictions by these methods is increasing even for the complicated cases of exosites or allosteric binding sites, their reliability has not been fully established. For a critical assessment of reliability, we use a set of ligand-protein complexes, which were found to be problematic in previous studies. The robustness of BD and PS methods is addressed in terms of success of the selection of truly functional pockets from among the many putative ones identified on the surfaces of ligand-bound and ligand-free (holo and apo) protein forms. Issues related to BD such as effect of hydration, existence of multiple pockets, and competition of subsidiary ligands are considered. Practical cases of PS are discussed, categorized and strategies are recommended for handling the different situations. PS can be used in conjunction with BD, as we find that a consensus approach combining the techniques improves predictive power.

Keywords: peptide; binding site; drug; complex; solvent; co-factor; free energy; scoring; equilibrium

Introduction

Location of functional binding pockets on protein molecules is a cornerstone of structural genomics[1–4] and targeted drug design projects. The advancement of experimental techniques, such as high throughput crystallography[5,6] allows the atomic level determination of ligand structures bound to their protein pockets at an increasing rate. However, there are still many cases where the determination of the structure of the complex of a protein with its known ligand fails (1) or even the ligand is unknown (2), and still the knowledge of location of the functional pocket(s) is necessary.

The blind docking (BD) method has been introduced[7,8] as an extension of the use of the very powerful docking engine AutoDock[9,10] for the abovementioned case 1 where the chemical identity of the ligand is known. During BD, the entire surface of the protein target is scanned for putative binding pockets of the ligand, and an atomic resolution complex structure is resulted. It was shown[7,8] that in many cases, the primary, functional binding pocket of the ligand can be selected from among the identified pockets according to the binding free energy (ΔG) values corresponding to the interactions of the ligand with the different pockets. Notably, ΔG is produced on-the-fly by a scoring function during the docking procedure. Besides the location of the pocket of primary ligands, numerous studies[11–15]

Additional Supporting Information may be found in the online version of this article.

Grant sponsors: János Bolyai Research Scholarship of the Hungarian Academy of Sciences. European Social Fund (Grant number: TAMOP-4.2.1/B-09/1/KMR-2010-0003) Eötvös Fellowship of the Hungarian State

*Correspondence to: Dr. Csaba Hetényi, Box 596, 75124 Uppsala, Sweden. E-mail: csabahete@yahoo.com

Table I. *Overview of Blind Docking (BD) and Pocket Search (PS) Methods involved in this Study*

Name	Class	Search method	Scoring	References
AutoDock4	BD without PS	Genetic algorithm	$\Delta G = E_{\mathrm{vdW}} + E_{\mathrm{H\text{-}bond}} + E_{\mathrm{elec}} + \Delta G_{\mathrm{solv}} \text{-} T\Delta S_{\mathrm{tors}}$	18,19
EADock$_{\mathrm{SF}}$ EADock$_{\mathrm{FF}}$	BD with PS	LIGSITE-based PS and a subsequent local search using an evolutionary algorithm	$\Delta G_{\mathrm{SF}} = E_{\mathrm{intra,L}} + E_{\mathrm{intra,P}} + E_{\mathrm{vdW}} + E_{\mathrm{elec}}$ $\Delta G_{\mathrm{FF}} = \Delta G_{\mathrm{SF}} + \Delta G_{\mathrm{solv}}$	20–22
SITEHOUND$_{\mathrm{X}}$	PS (chemical)	Chemical probes are placed on evenly spaced grid points, their IEs are calculated and binding pockets are defined as clustered grid points with highest TIE.	$\mathrm{IE_X} = E_{\mathrm{vdW}} + E_{\mathrm{elec}}$ $\mathrm{TIE_X} = \Sigma E_{\mathrm{X,cluster}}$ (X=C or OP)	23
Q-SiteFinder	PS (chemical)		$\mathrm{IE_C} = E_{\mathrm{vdW}} + E_{\mathrm{H\text{-}bond}} + E_{\mathrm{elec}} \mathrm{TIE_C} = \Sigma \mathrm{IE_C}$	24,25
Pocket-Finder	PS (geometrical)	Probe spheres are placed on evenly spaced grid points, and clustered into putative pockets.	The count of grid points which are well-buried in the protein (exceeding a pre-defined threshold).	25
PASS	PS (geometrical)	Protein surface is covered with layers of probe spheres. Pockets are predicted as active site points (ASP) having the largest weight among all probe spheres.	The weight of an ASP is proportional to the count of probe spheres in the vicinity and the extent to which they are buried.	26

ΔG: free energy of binding. E: interaction energy between all ligand (L) and protein (P) atoms except cases where "intra" refers to intra-molecular interactions inside L or P. vdW: van der Waals-interactions. H-bond: hydrogen bonding interactions. Elec: electrostatic interactions. ΔG_{solv}: change of solvational free energy during ligand binding. T: thermodynamic temperature. ΔS_{tors}: change of entropy of internal rotations during ligand binding. SF: SimpleFitness scoring. FF: FullFitness scoring. $\mathrm{IE_X}$: Interaction Energy of a probe X with the protein target. TIE: Total Interaction Energy for probes in a cluster. C: probe mimicking a methyl group. OP: probe mimicking a phosphate group.

have shown that the BD approach is useful in the solution of delicate problems such as the detection of subsidiary binding pockets containing e.g. exosites or allosteric binding sites.

In case 2, where the ligand is not known, only the protein sequence and/or structure can be used as input information. There are various site detection[16] and pocket search[17] (PS) methods available to accomplish this task. In Table I, a short summary is given on some PS methods used in this study. These methods are citation-classics (Q-SiteFinder[24] and Pass[26]), and a novel, promising program Sitehound[23] is also included. The PS algorithms use either geometrical or simplified, chemical grid-based search routines, and represent the putative binding pockets as a cluster of probe spheres. Since a PS does not use ligand information, it cannot provide the atomic resolution ligand-protein complex and the corresponding ΔG. Instead of ΔG, PS methods calculate other type of scores for ranking and selection of the most probable pockets. Such scores are based on the depth of the pocket or a sum of interaction energy values of the clustered probes with the protein.

In BD calculations based on AutoDock, the docking of the ligand structure can be performed in parallel in, for example, 100 trials starting from 100 different random positions around the entire protein surface and this global search results in 100 putative binding modes (pockets) and the corresponding ΔG values. Thus, preliminary PS is not necessary in principle, as the numerous global search trials scan the entire protein surface at atomic resolution. However, in other docking packages such as EADock[20,21] or GOLD, the PS is a necessary prerequisite of BD as the atomic level docking calculations are focused only on the pockets previously identified by PS. A recent study[27] also suggests that a preliminary PS can improve BD by AutoDock, as well.

Despite the above-mentioned increasing knowledge on the application of BD, PS, and their combinations, detection of functional pockets and atomic level binding modes is still challenging for the following reasons. Generally, BD and PS methods identify many putative binding modes and pockets including the real one(s), but the scoring schemes cannot select the real, functional pockets in all cases. Ideally, the aim of BD and PS is the location of the primary pocket. However, in reality, there are subsidiary ligands (co-factors, solvent additives, ions, etc.) available for the same protein target. Together with the hydrating water molecules, the primary and subsidiary ligands compete with each other for the available pockets and can interfere with the equilibrium binding process of each other. Similarly, one ligand can bind to subsidiary, e.g. allosteric pockets besides its primary pocket on the protein.

To address the above problems and formulate some rules on the applicability of the BD and PS methodology, a comparative analysis was conducted using different search engines and scoring schemes (Table I) as follows.

1. The entire surface of the ligand-bound (holo) and primary-ligand-free (where available, apo) conformations of protein targets (Table II) were

Table II. *Protein-Ligand Complexes used for Evaluation*

Protein Code[b]	Name	AA Count[c]	Water Count[d]	Apo structure[e]	RMS (Å)[f]	Ligand Code (j)	Name[g]	Volume (Å^3)	Category[a]
PDB codes									
1b70	phenylalanyl tRNA synthetase	1039	134	1pys	0.42	1	phenylalanine	203	BD-passed
1cea	recombinant kringle 1 domain of human plasminogen	79	148	1pkr	0.50	1	aminocaproic acid	181	Drug complex
1dy4	cellobiohydrolase I	434	342	1cel	0.30	1	s-Propranolol	338	BD-failed
						2a	NAG 435	240	
						2b	NAG 436	240	
1e7a	human serum albumin	577	120	1ao6	0.89	1a	propofol 4001	255	BD-failed
						1b	propofol 4002	255	
1eqg	prostaglandin h2 synthase-1	550	251	1prh	0.38	1	Ibuprofen	294	BD-failed
						2a	NAG 661	240	
						2b	NAG 681	240	
						2c	NAG 662	240	
						2d	NAG 671	240	
						2e	NAG 1672	240	
						3a	BOG 802	383	
						3b	BOG 801	383	
						4	HEME	732	
1h61	pentaerythritol tetranitrate reductase	364	545	1h50	0.21	1	prednisone	427	Drug complex
						2	FMN	456	
1hvy	human thymidylate synthase	287	596	1hw3	0.77	1	Raltitrexed (Tomudex)	521	Drug complex
						2	UMP (covalently bound)	293	
1hz4	transcription factor malt domain iii	366	408			1	benzoic acid	147	BD-failed
						2	GOL	114	
1ivb	influenza virus b/lee/40 neuraminidase	390	0			1	4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid	245	BD-failed
						2	NAG	240	
1ju4	cocaine esterase	569	436	3i2j	0.22	1	benzoic acid	147	BD-failed
1lna	thermolysin	316	158	1l3f	0.62	1	Val-Lys	322	BD-failed
1m2z	human glucocorticoid receptor ligand-binding domain	254	205			1	Dexamethasone	459	Drug complex
						2a	BOG 501	383	
						2b	BOG 778	383	
						2c	BOG 779	383	
1ngp	n1g9 (igg1-lambda) fab fragment	431	131	1ngq	0.29	1	2-(4-hydroxy-3-nitrophenyl) acetic acid	204	BD-failed
1pth	prostaglandin h2 synthase-1	550	1	1prh	0.36	1	salicylic acid	158	BD-failed
						2a	NAG 661	240	
						2b	NAG 671	240	
						2c	NAG 672	240	
						2d	NAG 681	240	
						3	BOG	383	
						4	HEME	732	
3pcn	protocatechuate 3,4-dioxygenase	436	1374	2pcd	0.39	1	2-(3,4-dihydroxyphenyl) acetic acid	194	BD-failed
3tpi	trypsinogen-bpti complex	287	152			1	Ile-Val	314	BD-passed

[a] Categories of complexes according to previous investigations. BD-passed/BD-failed: BD of the ligand to the protein was successful/failed with AutoDock3 in the previous studies[7,8]. Drug complex: complexes with drug as ligand molecule bound to protein.

[b] Protein code used as a reference in this study for both the holo and apo target forms (identical to the PDB ID of the holo conformation of the protein).

[c] Number of amino acid residues in the protein target.

[d] Number of crystallographic water molecules found in the holo PDB file and used for evaluation.

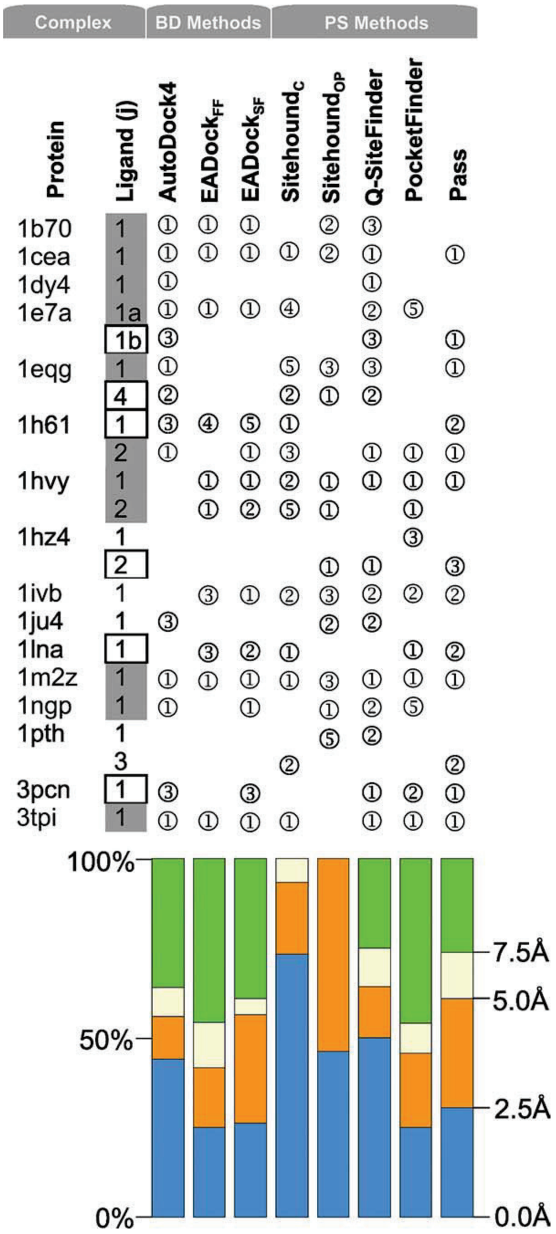
[e] PDB ID of the primary-ligand-free (apo) conformation of the protein target (not used as a reference code in the text).

[f] Root Mean Square Deviation between the C_α atoms of the holo and apo conformations of the target protein.

[g] Abbreviated names of ligand molecules. NAG: *N*-acetyl-D-glucosamine. BOG: β-octylglucoside. FMN: flavin mononucleotide. GOL: glycerol. UMP: 2′-deoxyuridine 5′-monophosphate.

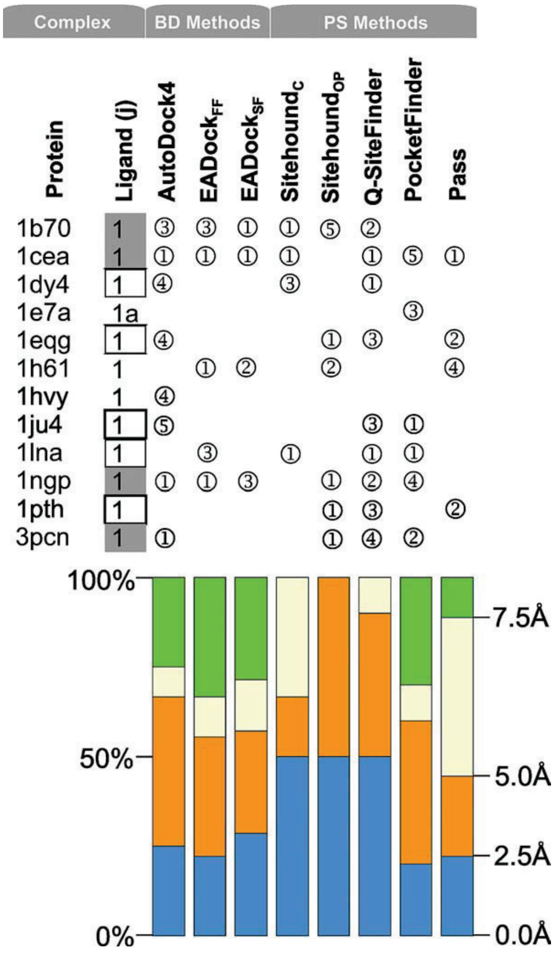
subjected to all BD and PS methods studied, and the results with the closest hits are summarized in the Supporting Information. From among the closest hits, Figures 1 and 2 list the top five rank numbers where the root mean squared deviation (RMSD) or the distance measured from the crystallographic ligand is smaller than 5 Å.

2. Only amino acid residues of the proteins were involved as target structures, that is, waters, ions, and all ligands were removed from the target during BD and PS. Importantly, even the modifications of native amino acids were removed in all cases to mimic the situation when a protein is built using only sequence data by means of structural genomics (homology modeling).
3. The most important part of our test set was composed of 10 protein targets which had been found problematic in previous studies[7,8] using BD driven



Complex		BD Methods			PS Methods				
Protein	Ligand (j)	AutoDock4	EADock$_{FF}$	EADock$_{SF}$	Sitehound$_C$	Sitehound$_{OP}$	Q-SiteFinder	PocketFinder	Pass
1b70	1	①	①	①		②	③		
1cea	1	①	①	①	①	②	①		①
1dy4	1	①					①		
1e7a	1a	①	①	①	④		②	⑤	
	$\boxed{1b}$	③					③		①
1eqg	1	①			⑤	③	③		①
	$\boxed{4}$	②			②	①	②		
1h61	$\boxed{1}$	③	④	⑤	①				②
	2	①		①	③		①	①	①
1hvy	1		①	①	②	①	①	①	①
	2		①	②	⑤	①		①	
1hz4	1							③	
	$\boxed{2}$					①	①		③
1ivb	1		③	①	②	③	②	②	②
1ju4	1	③				②	②		
1lna	$\boxed{1}$		③	②	①			①	②
1m2z	1	①	①	①	①	③	①	①	①
1ngp	1	①		①		①	②	⑤	
1pth	1					⑤	②		
	3				②				②
3pcn	$\boxed{1}$	③		③			①	②	①
3tpi	1	①	①	①	①		①	①	①

Figure 1. Successful predictions using the ligand-bound conformation of proteins as targets. Rank serial numbers of the top five Ranks with an RMSD/distance <5Å (compared with the crystallographic ligand pose) are listed in circles. Grey-filled boxes mark ligands with Category 1 predictions. Empty boxes denote Category 2 predictions (see Section Discussion for categories.) Color bars represent the precision of the methods in terms of distribution of the above RMSD/distance for complexes where the closest solution was found in the top five Ranks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]



Complex		BD Methods			PS Methods				
Protein	Ligand (j)	AutoDock4	EADock$_{FF}$	EADock$_{SF}$	Sitehound$_C$	Sitehound$_{OP}$	Q-SiteFinder	PocketFinder	Pass
1b70	1	③	③	①	①	⑤	②		
1cea	1	①	①	①	①		①	⑤	①
1dy4	$\boxed{1}$	④			③		①		
1e7a	1a							③	
1eqg	$\boxed{1}$	④				①	③		②
1h61	1		①	②		②			④
1hvy	1	④							
1ju4	$\boxed{1}$	⑤					③	①	
1lna	$\boxed{1}$		③		①		①	①	
1ngp	1	①	①	③		①	②	④	
1pth	$\boxed{1}$					①	③		②
3pcn	1	①				①	④	②	

Figure 2. Successful predictions using the primary-ligand-free conformation of proteins as targets. Rank serial numbers of the top five Ranks with an RMSD/distance <5Å (compared with the crystallographic ligand pose) are listed in circles. Grey-filled boxes mark ligands with Category 1 predictions. Empty boxes denote Category 2 predictions (see Section Discussion for categories.) Color bars represent the precision of the methods in terms of distribution of the above RMSD/distance for complexes where the closest solution was found in the top five Ranks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

by AutoDock3.[9] In these cases, BD had not been able to reproduce the correct crystallographic pose of the primary ligand (Ligand 1 or L_1), and/or the scoring scheme had not been able to distinguish the accurately reproduced pose of L_1 from the other, nonrelevant poses. In the latter cases, the correct pose had been incorrectly sorted into a higher (>1), energetically less favorable rank and, therefore, the identification of the appropriate pose as a 1st rank had failed. In this study, the above problematic complexes were considered as negative test cases and marked as "BD-failed" in Table II. There were also other two complexes for which BD had been successful (BD-passed) previously, and four proteins with drug ligands included as untested cases.

4. The methods were tested not only for finding the 17 primary ligand pockets (L_1) but also for the rather difficult detection of 23 pockets of subsidiary ligands ($L_{j\geq 2}$), which turned out to be a real challenge for the methods. Thus, altogether 40 different structural complexes were considered. In cases where two or more binding pockets were available for the same ligand in the Protein Databank (PDB) complex, the binding poses in the different pockets were distinguished by small letters (Table II) after the numeric code of the ligand and used as separate references in comparisons.
5. Interference of ligands and hydrating water molecules on BD search was also explored to analyze the results of failed predictions.

The aim of this study is to give an estimate on the reliability of the methods in the case of problematic complexes. We are particularly interested if the PS methods of different background can help in the verification of binding pockets found by BD. We discuss whether a consensus (BD+PS) approach may show the way ahead toward functional protein pockets.

Results

Easy cases with primary ligands

Among the positive (BD-passed) test cases, the 3tpi protein is relatively small (Table II), its ligand is a dipeptide and, therefore, the holo protein form of this complex is an easy job for BD or PS. The RMSD of Rank 1 (Supporting Information) was an excellent 0.9 Å with AutoDock4[18,19] and 2.6 Å using EADock. The PS methods also found the pocket centrums as Rank 1 (Fig. 1) with good accuracy except Sitehound$_{OP}$ (Sitehound with phosphate probe), and the best algorithm in this case was Q-SiteFinder with a 0.8 Å distance. Notably, Sitehound$_{OP}$ has been recommended[23] for phosphate-containing compounds originally. BD with both AutoDock4 and EADock successfully passed the other positive test case of holo form of 1b70, a very large target and also of two small drug-binding proteins 1cea and 1m2z. However, PS methods except Sitehound$_{OP}$ and Q-SiteFinder failed for 1b70 and Pocket-Finder failed in the case of 1cea. BD and PS were also successful in finding the L_1 pocket on the apo form of 1cea and partly of 1b70 (Fig. 2, notably the PDB ID of the holo form of the protein is used as a code also for the apo form in this study). It was somewhat unexpected, that the two BD methods had only partial success in the cases of the holo target forms of two additional drug complexes 1h61-L_1 and 1hvy-L_1 in both ranking and precision. The only top 1 BD result was found with EADock, the 1hvy-L_1, but the corresponding precision was still moderate with a 3.5 Å RMSD. For the apo forms, the success was also limited. The PS methods provided good hints for the holo target forms: Sitehound$_C$, identified the L_1 pocket as 1st rank for target 1h61, as well as Sitehound$_{OP}$ did for 1hvy. The other PS methods (Q-SiteFinder, Pocket-Finder and Pass) identified the L_1 pockets correctly as a 1st rank at 1hvy and less correctly for 1h61 and for the apo target forms where the distance between the pocket center identified by PS and that of the real pocket was either above 5 Å or it was ranked too high (>5).

Problematic cases with primary ligands

Figure 1 shows that two BD-failed groups can be distinguished according to the performance of AutoDock4 on L_1 complexes on the holo target form. In the first group, there are L_1-complexes with 1dy4, 1e7a, 1eqg, and 1ngp. In these cases, the present AutoDock4 ranking of the correct pose improved to the 1st rank compared with our previous studies[7,8] with AutoDock3, where they had not been found in the best rank. For example, the primary L_{1a} pocket of propofol on 1e7a had been identified[7] as Rank 2, whereas now it is located in Rank 1 [Fig. 2(a)] with a nice structural match. AutoDock4 produced a correct RMSD for all four cases. EADock reproduced the crystallographic L_1 structures at targets 1e7a and 1ngp (the other two complexes were misranked). In general, the PS methods were not successful in these four cases as they provided only some isolated good hits for the 1st rank at 1dy4 (Q-SiteFinder), 1eqg (Pass), and 1ngp (Sitehound$_{OP}$). There were also some cases with the correct pose located in the 2nd and 3rd ranks by Q-SiteFinder and once by Pocket-Finder. For the apo target forms, BD generated a top 1 rank only for 1ngp, whereas for 1eqg and 1dy4 only Rank 4 was produced. The above-mentioned second group includes six targets (1hz4, 1ivb, 1ju4, 1lna, 1pth, 3pcn), where AutoDock4 could not improve the ranking/RMSD precision for L_1 on the holo target forms compared with previous studies[7,8] and failed. EADock$_{SF}$ found the correct pose in only one of six cases (1ivb) as Rank 1. At 1lna, EADock$_{SF}$ found the crystallographic ligand structure as Rank 2, which is remarkable as the Co^{2+}-ion important in ligand binding was not used in this study due to our strict criteria of BD (Introduction). Unexpectedly, the PS methods performed fairly well for two-third (1ivb, 1ju4, 1lna, 3pcn) of this challenging group placing the real pocket into the first three ranks (Fig. 1). In the case of the apo target forms, BD failed to predict the pocket for this group with an exception of 3pcn. The PS methods provided good hints for the apo forms too (Fig. 2).

Subsidiary ligands and pockets

In the case of 1e7a, the same primary ligand (L_1, Propofol) binds at two different pockets in the crystallographic structure [Fig. 3(a)]. The second binding pocket (L_{1b}) of Propofol was identified as a 3rd rank by AutoDock4, and two PS methods (Q-SiteFinder and Pass) also placed it in the top three ranks using the holo target form (Fig. 1). No method found L_{1b} on the apo protein structure. Besides the primary ligands, for eight (1dy4, 1eqg, 1h61, 1hvy, 1hz4, 1ivb, 1m2z, 1pth) of the sixteen protein targets of this study, there are also subsidiary ligands ($L_{j\geq2}$) some of them having multiple binding pockets. The subsidiary ligands (Table II) can be divided into two groups: functional or structural partners (1) and small molecules or solvent additives (2).

The first group contains strong binders such as the nucleotides (FMN, UMP) and the HEME. In the complexes of FMN and UMP (1h61-L_2 and 1hvy-L_2), the co-factors are located close to L_1 discussed above. Thus, in both complexes, L_1 and L_2 interact with each other and their binding pockets are not separated influencing the docking results (see also next section). In the case of FMN as L_2, the BD methods performed better for the L_2 than for the L_1 at target 1h61 (only holo protein structures were used as targets for subsidiary ligands). However, the only acceptable solution was produced for 1h61-L_2 by AutoDock4. Sitehound$_{OP}$ identified the 1st rank for the phosphate containing L_2 (UMP) of 1hvy similar to its L_1 earlier and Sitehound$_C$ ranked the L_2-s into worse ranks than the L_1-s (Fig. 1). The other PS methods (Q-SiteFinder, Pocket-Finder and Pass) at most identified the L_2 pockets for both complexes correctly. HEME is the largest ligand investigated with a 732 Å^3 molecular volume (Table II). It is part of two target-ligand complexes, the 1eqg-L_6 and the 1pth-L_7. Although these complexes were not re-produced perfectly by BD, a 2nd rank at 1.4 Å RMSD (1eqg) and a 1st rank with a 7 Å RMSD (1pth) were obtained with AutoDock4. PS methods Sitehound and Q-SiteFinder worked well in the case of 1eqg but none of them were really successful for 1pth.

Molecules of the second group are loose binders (BOG, GOL, NAG), and/or they sit in a shallow surface pocket [Fig. 3(b)] of the protein in question. Ligands of this group can be found in various

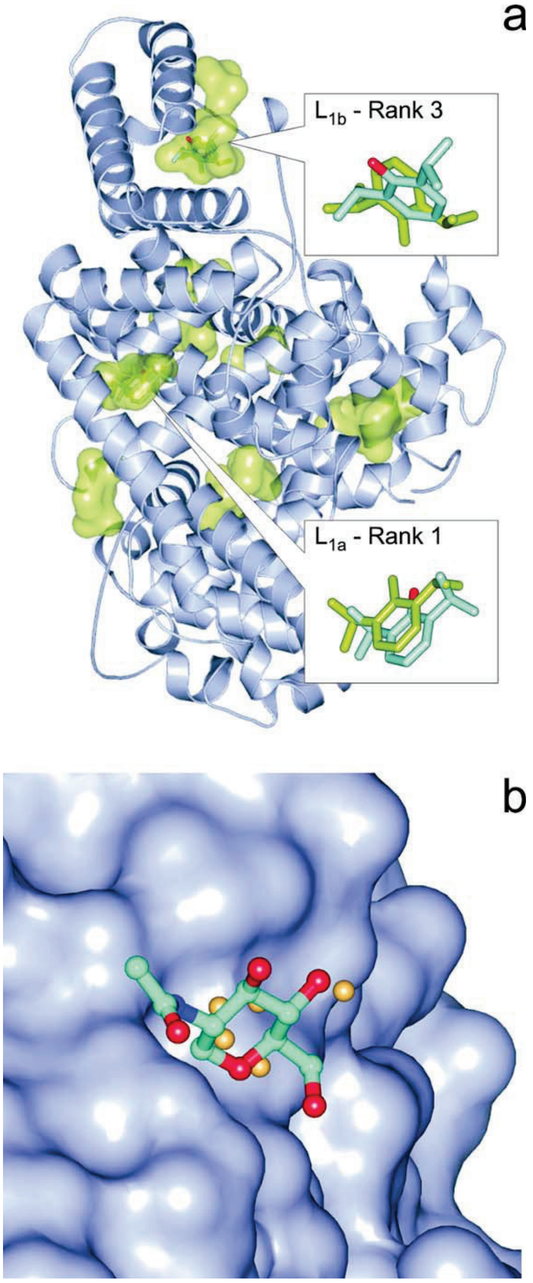


Figure 3. (a) In PDB structure 1e7a two binding pockets of the primary ligand (L_1) propofol had been detected by crystallography. Whereas the L_{1a} pocket and the binding mode was identified precisely by BD (shown in inset as green sticks) as Rank 1 and Q-SiteFinder, the L_{1b} pocket was located by PS methods and BD found it as Rank 3 (green sticks) with a rather high deviation from the crystallographic position (sticks colored by atom type). Other pockets found by BD are also shown as green surfaces. **(b)** Sitehound identified the shallow pocket (protein shown as surface) of NAG in the complex 1eqg-L_{2d} in a 2.9 Å distance from the crystallographic ligand position (balls and sticks colored by atom type) by placing a few Carbon probes (beige spheres) into the proposed pocket. The small number of probes resulted in a low TIE value and a mis-ranking of this real binding pocket into the 88th of 112 ranks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

complexes (Table II). Neither BD nor PS methods were successful in correct identification of their binding poses or pockets. There were few isolated cases where the methods produced good hints, such as Pass (1m2z-L_{2c}, 1pth-L_3, 1hz4-L_2), Sitehound$_C$ (1pth-L_3), Sitehound$_{OP,}$ and Q-SiteFinder (1hz4-L_2).

The influence of subsidiary ligands

Subsidiary ligands of a target protein are not just test cases of BD and PS (previous section), but they also introduce difficulties due to their inherent competition for the pockets in real life (Introduction). Thus, the scoring function of BD has to be selective enough for a ligand j (L_j) to distinguish its primary crystallographic pocket not only from its subsidiary pockets but also from pockets occupied by other, competing ligands n ($L_{n\neq j}$) on the same target protein. To gain information on this kind of selectivity of BD, it is useful to check whether the crystallographic binding poses of L_ns on the same target indeed interfere with the (mis)docked L_j poses. Ideally, if the BD method is precise and selective enough, then no interference should be measured between the docked L_j pose in the 1st rank and the crystallographic poses of other L_ns. In other words, L_j should occupy its crystallographic conformation with the lowest ΔG (Rank #1) and bind to crystallographic pockets of other L_ns (or waters, see next section) at Ranks>1 of higher ΔG values. Such interferences were checked at both BD methods and ranks on all holo target forms, by the measurement of the distances between the docked L_j and the crystallographic L_n poses (see Methods for details). The L_j ranks with significant L_n interferences (<5 Å distance) found are listed in the Supporting Information.

As it was expected, a comparison with the results (Fig. 1, Supporting Information) shows no interference at Rank 1 in the cases where the BD method identified the crystallographic pocket of L_j correctly. Some interferences can still be found for these positive cases but only at Ranks >1. To illustrate one of the examples mentioned, Figure 4(a) depicts prostaglandin h2 synthase-1 (1eqg) with 10 binding poses of L_1 (Ibuprofen) representing the 10 ranks found by AutoDock4 (Supporting Information). There is only one point of interference between the 10 rank-representatives of L_1 and the other ligands (L_n): L_4 (HEME) sits at the same place as Rank 8 of L_1 [Fig. 4(a)]. Naturally, as this interference occurs at Rank 8, it corresponds to a higher ΔG of L_1, and therefore, it can be concluded that BD could discriminate between the binding of L_1 to its real site and to the HEME-site by assigning a lower ΔG for the real one.

Besides analyses of the above-mentioned positive examples, checking ligand-interferences may provide even more important information in the negative cases where the crystallographic pocket of L_j was not identified correctly as Rank 1. For example,

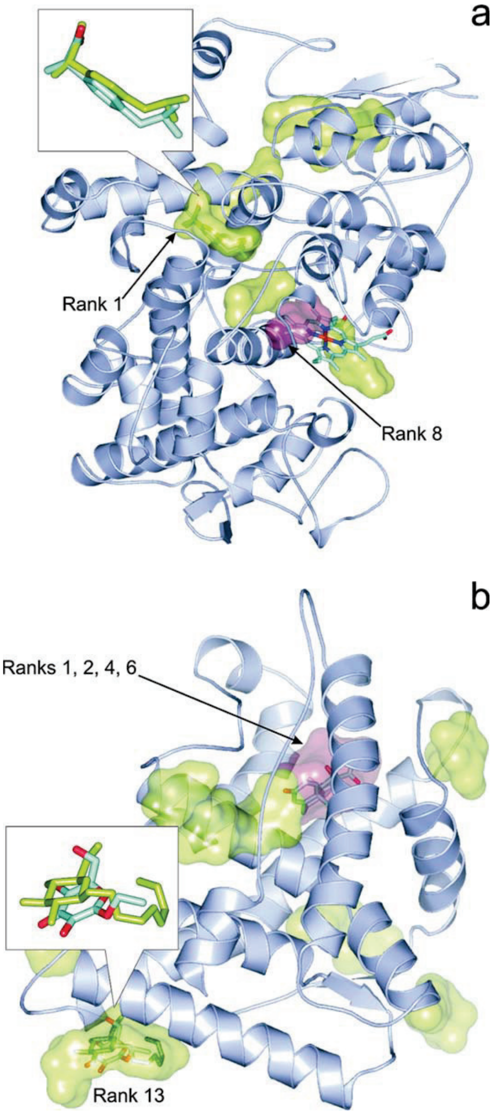


Figure 4. (a) Prostaglandin H2-synthase-1 (1eqg, cartoon) and the binding pockets (green surface) of the primary ligand Ibuprofen corresponding to the 10 ranks found by AutoDock4. In the inset, the overlap between crystallographic (sticks colored by atom type) and docked (sticks in green) Rank 1 conformations of Ibuprofen is featured. Interference of Rank 8 pocket (pink surface) with HEME (sticks colored by atom type) did not affect the results of BD. **(b)** The binding pockets of BOG (green) identified by AutoDock4 on the surface of human glucocorticoid receptor (1m2z, cartoon). In the inset, the overlap between the crystallographic (sticks colored by atom type) and docked (sticks in green) Rank 13 conformations of BOG is shown. Notably, the octyl group of BOG was not assigned in the crystal structure and the B-factors of the assigned atoms are relatively high (76-90). Predicted pockets of BOG corresponding to Ranks 1, 2, 4, and 6 (pink surface) interfere with the pocket of the primary ligand dexamethasone (sticks colored by atom type), which is partly responsible for the mis-ranking of the correct pocket only as Rank 13. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

at complexes 1dy4-L_2, 1eqg-L_{3a}(n = 4), 1h61-L_1, 1m2z-L_{2c} (AutoDock4), and at 1h61-$L_{1,2}$, 1m2z-L_{2c}, 1pth-L_3 (EADock), there are interferences (Supporting Information) with L_ns at ranks with serial numbers equal to/smaller than the best ranks listed in Figure 1. The 1m2z-L_2 complex is such an example [Fig. 4(b)], where the binding site of BOG was identified on the surface of human glucocorticoid receptor (1m2z) at the L_{2c} pocket with a reasonably good fit to the crystallographic conformation at 2.2 Å RMSD. However, predicted pockets of BOG corresponding to Ranks 1, 2, 4, and 6 overlap with the pocket of the competing primary ligand dexamethasone (L_1), and this interference is partly responsible for the mis-ranking of the correct pocket only as Rank 13 of 14 (Supporting Information).

Another example is 1h61-L_1, where the solution for prednisone (L_1) with the best RMSD (1.0 Å) was placed to the 3rd rank of the total of 10 ranks found by AutoDock4 as a consequence of L_1 interference with FMN [L_2, Fig. 5(a)] in the cases of Ranks 1 and 2. As the mis-docked L_1 poses at Ranks 1 and 2 adopted a lower ΔG at the L_2 binding site (not shown in the figure), the crystallographic pocket was identified only as Rank 3 suggesting that BD was not energetically selective enough to favor the real pocket of L_1 over the actual pockets of another ligand (L_2).

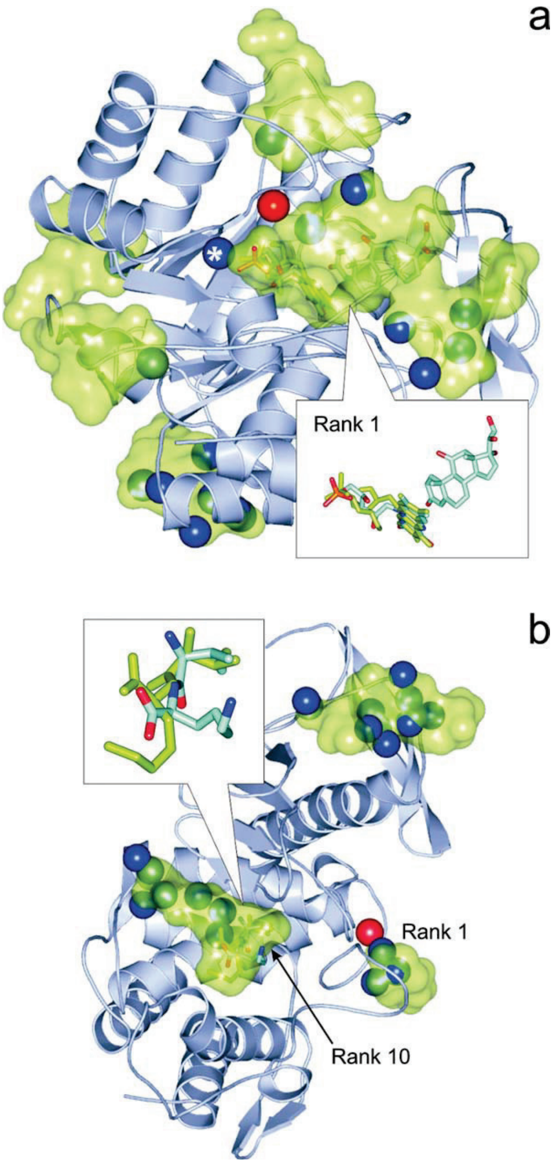


Figure 5. **(a)** Pentaerythritol tetranitrate reductase (1h61, cartoon) and the binding pockets (green surface) of subsidiary ligand FMN corresponding to the 10 ranks found by AutoDock4. In the inset, the overlap between crystallographic (sticks colored by atom type) and docked (green sticks) Rank 1 conformations of FMN is featured. Notably, the primary ligand prednisone (sticks colored by atom type, in the right corner of inset) is located in an adjacent pocket very close to FMN resulting in ligand-interference and mis-ranking during the docking of prednisone. Blue and red spheres depict the positions of crystallographic water oxygen atoms inside (Type 1) and the bottom (Type 2) of the binding pockets, respectively. The oxygen atom marked with an asterisk represents the only type 1 water interfering with Rank 1 (see main text for details). **(b)** In the case of docking of ValLys to thermolysin (1lna, cartoon), the binding pocket of the ligand was found as a 10th rank, whereas, for example, in the 1st rank identified by docking two crystallographic water molecules (blue spheres) are sitting in reality. In pockets of Ranks 2. . .9 (green surfaces), there are 18 water molecules (blue spheres) occupying the binding positions instead of the ligand. In the inset, the overlap between crystallographic (sticks colored by atom type) and docked (green sticks) Rank 10 conformations of ValLys is featured showing a large deviation in the position of the charged side-chain. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The influence of hydration

Like the subsidiary ligands discussed above, solvent molecules can also compete with the binding of a ligand in question. Due to their different locations related to the (docked) ligand and the target protein, there are two types of water molecules distinguished in the present investigation. As BD is usually performed for “dry” protein target, it is possible that predicted ligand binding sites in fact are occupied by solvent molecules. These water molecules sit inside the pocket and are classified as Type 1 water in this study (see also Methods). There are also other water molecules located at the interface between the (docked) ligand and the protein target, at the bottom of the pocket (Type 2) not occupying the docked ligand position. Whereas the Type 1 waters surely compete with the ligand for the pocket, and in the real situation hinder its binding, Type 2 water molecules are not obviously expected to block ligand binding to the actual pocket as they can also assist the ligand-protein interaction via bridging or as spacers. Water-interferences with the representative docked ligand structures at all ranks for both BD methods are tabulated in the Supporting Information.

Similar to the previous section in most of the positive cases of Figure 1, there are no Type 1 water molecules interfering with the docked L_j positions at Rank 1 except three cases (1h61-L_2, 1ngp-L_1 at AutoDock4 and 3tpi-L_1 at EADock) each of them with only one water molecule inside the pocket. At the same time, in Ranks=2, there are several waters located at the same place as the docked ligand. This finding shows that the scoring function was able to distinguish between the real (Rank 1) pocket of an L_j ligand and the others filled with Type 1 water molecules assigning higher ΔG a for the latter pockets. For example, in the case of 1h61-L_2 169 Type 1 water molecules can be found (Supporting Information) at different Ranks=2 of higher ΔG values calculated by AutoDock4 and only one Type 1 water molecule [Fig. 5(a)] is located in the middle of the pocket corresponding to Rank 1. The excellent 1.5 Å RMSD between the Rank 1 and the crystallographic L_2 conformation also shows that interference of only one Type 1 water molecule may not destroy the correct ranking of the pocket. In these cases, the absence of one Type 2 water molecule also did not influence the correct ranking order during docking (robustness).

Further comparison of the docking results and the table on water interferences (Supporting Information) shows that there are several negative cases, where a ligand was mis-docked into pockets corresponding to lower ranks filled with Type 1 water, in reality. Such examples are 1h61-L_1, 1hvy-$L_{1,2}$, 1hz4-L_2, 1ju4-L_1, 1lna-L_1, 1m2z-2c, 3pcn-L_1 (AutoDock4) and 1h61-L_2, 1hvy-L_1, 1lna-L_1, 3pcn-L_1 (EADock). For example, at 1lna-L_1 [Fig. 5(b)] Rank 10 holds the closest docked ligand with an RMSD of 2.8 Å. Interferences of the docked L_1 poses of all higher Ranks (1...9) with Type 1 waters are shown among the AutoDock4 results. Remarkably, in Rank 4, there are four water molecules sitting in the place of L_1. In the case of, for example, 3pcn-L_1 (AutoDock4) the lack of three Type 2 water molecules (Rank 3) also contribute to the mis-docking of the corresponding pose.

Discussion

The results of this study are categorized according to the success of BD and PS methods in the following sections.

Category 1: at least two different BD or at least one BD and one PS method provides a successful, consensus prediction in Rank 1

Figure 1 shows that in 11 of all 40 complexes (9 of 17 L_1-complexes) a valid Category 1 prediction could be produced in case of the holo target forms. Remarking that our test set contained mainly problematic complexes, it can be concluded that in the case of primary ligands, for half of the complexes a good consensus prediction can be achieved. This ratio is much less (4 of 17) for apo targets (Fig. 2). In four BD-failed cases (1dy4, 1e7a, 1eqg, 1ngp), the crystallographic position of L_1 was identified correctly using the holo targets as Rank 1 in this study, whereas in the earlier papers[7,8] using AutoDock3 these complexes were listed in higher ranks. This improvement may be due to the modified solvation term[19] of the scoring function (Table I) of AutoDock4. However, in case of the apo targets BD was successful at 1ngp only (Fig. 2).

In general, the ratio of the top five ranks at <2.5 Å with AutoDock4 dropped to almost the half at the apo forms (Fig. 2) compared with the holo targets (Fig. 1). At the same time, the ratio of the top five ranks with lower precision (2.5...5 Å) increased for the apo targets with AutoDock4 resulting that >50% of the top five ranks is below a 5 Å precision limit. The precision of EADock is low for both target forms. Notably, for subsidiary ligands only one consensus prediction was achieved 1h61-L_2 (Fig. 1). The reason of the low success rate at these ligands can be attributed to their disturbing interference with the primary ligand and their higher (less specific) binding energy.

In practice, consensus pockets of Category 1 are the most reliable BD predictions as they are supported by a different BD and/or PS methods.

Category 2: only PS methods provide successful predictions in Rank 1, and BD methods fail

The precision of PS methods is fairly independent on the target form. The ratio of the best (<2.5 Å) solutions dropped with about 20% (Fig. 2) at the apo forms in the case of Sitehound$_C$, but the precision of the other four PS methods remained practically the same if compared with the holo forms (Fig. 1). Considering that PS methods are generally based on a simplified search and scoring scheme (Table I) it may be somewhat surprising that they outperform the atomic-level BD calculations producing Rank 1 hits for an additional 6/40 (holo form) and 5/17 (apo form) complexes of this category (Figs. 1 and 2).

We have previously demonstrated[7,8] that AutoDock reproduces many protein-ligand complexes faithfully using holo forms of the proteins. Here we find that, despite the above-mentioned (Section Category 1) improvements in the energy function of AutoDock version 4, some targets remain difficult especially their apo forms. Since the BD protocol tries to dock the entire ligand, a somewhat smaller/closed cavity in an apo structure may preclude insertion of the ligand. In some cases, the decline of ranking precision of BD methods on the apo forms (Fig. 2 vs. Fig. 1) can be attributed to the large change in overall protein conformation as indicated by the

RMS C_α distances (1e7a, 1hvy) or by local conformational change of the binding pocket residues (1cel, 1h61). PS algorithms might still detect the cavity however.

Similar to the docking methods, the scoring schemes of the PS methods are also generally cumulative (Table I), that is, the total interaction energy (TIE) score is the sum of individual interaction energy values (Sitehound, Q-SiteFinder) of probes or weighted count of probes (Pocket-Finder, Pass). Therefore, it can be expected, that large pockets corresponding to large ligands (with many interactions) will be found easily by PS at it was the case for 1eqg-L_4, with the largest ligand HEME (732 Å^3, Table I) or the still considerable 1hvy-L_1 (raltitrexed, 521 Å^3). However, in the other (1pth) complex of HEME its pocket was not found, and small pockets of, for example, ligands 3pcn-L_1 (194 Å^3), 1hz4-L_2 (114 Å^3), and 1ju4-L_1 (147 Å^3) were identified correctly as Rank 1 or 2 showing that pocket size is not the only factor which contributes to the good performance of PS in this category.

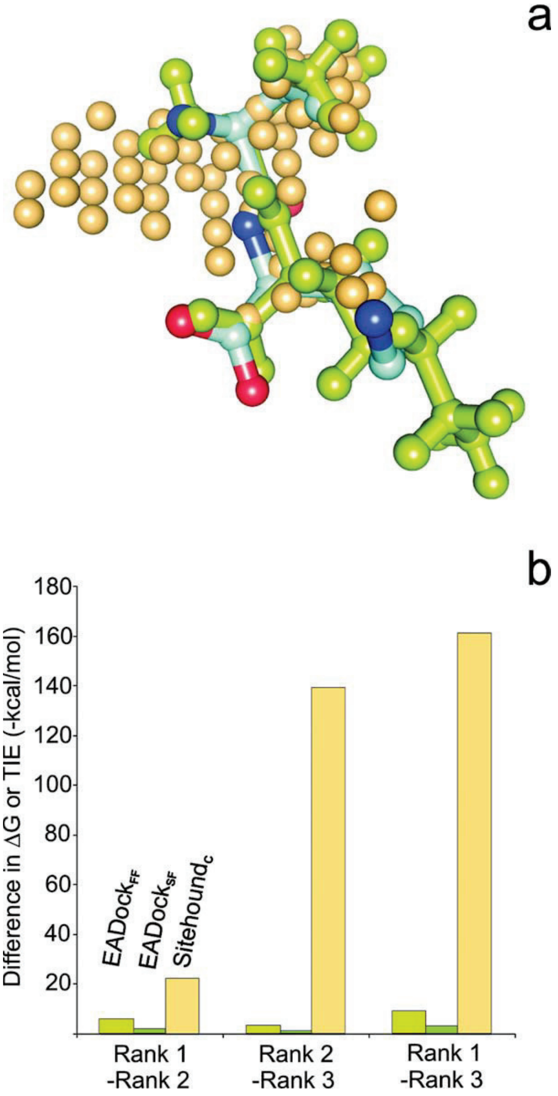


Figure 6. (a) The number of evenly and tightly spaced probes (beige spheres) representing the pocket according to Sitehound$_C$ is significantly larger than the number of atoms of the ValLys primary (L_1) ligand molecule docked by EADock (green balls and sticks). Protein thermolysin (1lna) is not shown and the crystallographic ligand conformation is represented by balls and sticks colored by atom type. Although the docked and crystallographic ligand conformations match with each other, this correct solution was placed to only Rank 3 according to EADock$_{FF}$ scoring. **(b)** Pairwise differences of the first three ranks in terms of binding free energy ΔG values calculated by EADock scoring schemes and the TIE values obtained by Sitehound$_C$ for the 1lna-L_1 complex. According to summation of interaction energy values corresponding to the relatively large number of probes shown in part (a) the differences in TIE values is larger than the differences between the ΔG values obtained for the few atoms of the docked ligand. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

A possible reason of the success of PS methods may arise from the nature of the grid of the probes used for the above cumulative scoring. Whereas in BD only the few atoms of a ligand molecule are used for calculation of interaction energies, in PS the count of probes/grid points of the identified pocket can exceed the number of atoms of the actual ligand [Fig. 6(a)]. In contrast with the connected atoms in the molecules, the location of grid points is determined by their even spacing and all of them are retained within a cluster without any concern on their possible connections. Thus, if the clustering algorithm of a PS works accurately, a functional pocket with large number of probe interactions/grid points will be ranked with a large energy difference [Fig. 6(b)] into the first rank rather than into the lower ranks by PS. In the case of PS, the TIE scores the general functionality of the pocket and not only its suitability/availability for the possible binding conformations of a ligand as is done by ΔG in the case of BD. A recent study suggests[28] that these kinds of robust scores and approaches representing multiple binding modes at a pocket may reflect the inherently dynamic nature of ligand binding. In this sense, the TIE may be considered as a score of pocket functionality in certain situations [Fig. 6(a,b)], where the small ΔG differences lead to misranking of pockets in BD even if its ΔG scoring is more sophisticated (multiple atom types, 3D geometry, connectivity, etc. considered) than the PS scoring like TIE, which is generally based on a single grid/probe type. Notably, this benefit is highly dependent on the robustness of the clustering scheme (the selection procedure of relevant probes/grid points for the proposed pocket) of PS and has mostly geometrical and no physical background. Obviously, the success of the PS scoring is not guaranteed. For the tricky situation of 1e7a-L_1 where the same ligand occupies two different pockets (a and b), two PS methods (Q-SiteFinder and Pass) with different background (Table I) provided a good hint (Fig. 1) for the second pocket L_{1b} [Fig. 3(a)] using the holo target, and at the same time L_{1a} was not identified by Pass (only by Q-SiteFinder). On the apo target, no PS methods ranked pocket L_{1b} in the top five. This example of complex 1e7a-L_1 shows the limitations of PS methods and the necessity of consensus pocket identification by at least two PS methods for a reasonable prediction.

In summary, Figures 1 and 2 suggest that independent top (1…3) rankings of the same pocket by 2-3 PS methods with different scoring and search schemes serve as an indication of a functional pocket even if BD fails to identify the pocket in question.

Category 3: all BD and PS methods fail to identify the pocket in Rank 1

Whereas pockets of the main ligands (L_1) were identified at least by one BD and/or PS method (Fig. 1), this was not the case for the co-factors and weak binders, and therefore, most of the $L_{j\geq 2}$ ligands fall into Category 3. A possible reason of the low performance of the BD methods in this category could be that these ligands are mostly weak binders and, therefore, the afore-mentioned energy difference between the Ranks is even smaller than it was in the cases of Category 2. Furthermore, as these ligands bind generally small and/or shallow pockets on the protein surface the above advantage of the cumulative grid scoring of the PS methods also cannot prevail due to the small number of selected (clustered) grid points/probes [Fig. 3(b)], defining the pocket, if any. Interferences with other ligands and hydrating water molecules detailed in section Results provide additional explanation for the failure of BD. The results on the influence of subsidiary ligands show that mis-docked $L_{j\geq 2}$ ligands may find the well-defined binding pocket of, for example, the main ligand L_1 [Fig. 4(b)] or of a large ligand. BD methods similarly failed, in the cases where the ligand was mis-docked in the pocket of two or more hydrating water molecules. Both types of interferences are results of the inadequate scoring function, which cannot distinguish energetically between the fine differences occurring at binding to different pockets. In the particular case of interferences of hydrating water molecules, the inappropriate solvation term (Table I), whereas for ligand interferences the whole scoring scheme is responsible for the lack of energy differences between the Ranks.

Category 4: only BD methods provide successful predictions in Rank 1 and all PS methods fail

This is a pseudo-category as the above situation did not occur in any of the 40 complexes. Once a BD method produced a successful (Rank 1) prediction, there was always at least one PS method providing the same pocket in Rank 1 and in one case in Rank 2 (Figs. 1 and 2). This finding is very important for the verification of Rank 1 hits in future BD studies showing that a true Rank 1 prediction by BD should be accompanied with at least one positive Rank 1…3 PS prediction for the same pocket. If it is not the case then it is suspicious that indeed the pocket was mis-found by BD and we face a Category 3 situation.

Conclusions

Recommendations

To estimate the reliability of a future search for the main (lowest energy) functional binding pocket on the entire protein surface, some rules can be concluded. (1) A Category 1 consensus pocket can be considered as a reliable prediction in most of the cases as it is based on positive Rank 1 results obtained by at least two different methods (sufficient and satisfactory conditions). (2) It is a suspicious situation if the pocket found by BD is not verified by at least one of the PS methods (pseudo-Category 4). This case may easily be indeed a Category 3, where the binding pocket is mis-found by inappropriate modeling or interferences of other binding partners. (3) If at least 2-3 different PS methods result in the same pocket which does not match the BD prediction (Category 2), then further investigations are necessary. For example, a local re-docking may be necessary including, for example, molecular dynamics calculations with explicit water surrounding to obtain an improved complex structure. As PS methods are very fast (some seconds), their above use for verification of BD results or prediction will not slow down the work.

Methodological aspects

The main problem with the BD and PS methods is that they produce a large number (=10) of possible ranks and corresponding pockets (Supporting Information) with very small differences (BD) in their scores in many cases. In this study, it was investigated whether among the many ranks, the consensus top ranks from BD and PS contain the real, functional pocket represented in the crystallographic structure or not. For this, a test set containing many problematic and/or weakly bound complex structures was used. It was also shown how other factors such as the interference of subsidiary ligands and/or a group of two or more Type 1 hydrating water molecules can negatively influence the BD results. The comparison of the quality of the methods for a non-BD (focused docking) problem was not the aim of the study, as there are numerous thorough analyses available, for focused (restricted) docking search. Similar to other studies,[7,28] the problem of protein flexibility was addressed in this study by the involvement of apo structures with ligand-free pocket conformations.

Multiple functionality

Beyond the crystallographic pocket (ideally corresponding to Rank 1, the global energy minimum), there may be others with equal or even more important function (e.g., allosteric binding sites) located in BD Ranks>1. Moreover, it can happen that the same ligand has more than one experimentally determined binding pockets [e.g., propofol on 1e7a, Fig. 3(a)]. The above recommendations were not meant for these pockets. In the case if the detection of these pockets is necessary, then (consensus) BD and PS Ranks>1 should be also considered and the corresponding sites checked by, for example, experiments.

Methods

Preparation of protein and ligand molecules

All protein-ligand complexes including the holo protein form and primary ligand–free (apo) protein structures (where available) were obtained from the PDB. All apo structures were superimposed on the holo structures and the respective RMSD measured between the C_α atoms of the holo and the superimposed apo protein structures are listed in Table II. This superimposition step allowed a comparison of the results on the apo structures with the holo-bound crystallographic ligand position. A list of the PDB codes is provided in Table II and used for identification of the protein in this study. All chains available in the PDB file were processed except the following cases where only the first copy of identical chains was used (chain identifiers listed in brackets): 1cea, 1e7a, 1eqg, 1hvy, 1m2z, 1pth (chain A), 1ngp (chains L and H), 3pcn (chains A and M). All nonamino acid (AA) residues and ligands were removed from the target proteins. AA side-chains containing post-translational modifications and non-AA (HETATM) groups were changed to contain only AA parts by deletion of the HETATM parts. That is, the first residue of 1dy4 was deleted, the Cme43 residue of 1hvy was mutated to Cys, and the Oah530 residue of 1pth was mutated to Ser. The use of only AA-containing targets allowed the study of real situations where no information on post-translational modifications is available. For BD with AutoDock4 and EADock the protein molecules were equipped with H-atoms using AutoDock Tools.[18] The ligand molecules (Table II) including co-factors and solvent additives were equipped with H-atoms and energy-minimized using Mopac 6[29] with a PM3 Hamiltonian and eigenvector following routine for energy minimization (except of HEME for where the crystal structure was used). In all cases, the force constant matrices were positive definite. For comparison, ligand volumes (Table II) were calculated by an analytical algorithm.[30]

AutoDock 4 calculations

BD jobs including 100 runs each were set up as described previously.[7] Briefly, the target and ligand molecules were equipped with Gasteiger charges using AutoDock Tools. Grid maps were calculated at 0.55 Å spacing and covered the entire surface of the target proteins. Docking runs were started with a random ligand position and orientation. The Lamarckian genetic algorithm and the pseudo-Solis and Wets local search with a maximum number of 20 million energy evaluations, 250 population size, 2 Å translation and 50–50° rotation and quaternion steps were applied. All sigma bonds of the ligand except rings and amide bonds were released during the flexible docking. Protein target was kept rigid, that is, protein flexibility was not considered during the calculation.

EADock calculations

The EADock calculations were performed using the SwissDock server (http://swissdock.vital-it.ch/). The ligand molecules were converted to Sybyl mol2 format using UCSF Chimera software as required by the server. The target molecules were provided as PDB files. Docking type was set to "accurate" in BD mode. The DOCK4-type outputs of the server containing 250 docked ligand conformations in each were used for subsequent ranking evaluations.

Ranking of BD results

A uniform procedure[7] was applied to rank the 100 and 250 docked ligand structures of each complex produced by AutoDock 4 and EADock jobs, respectively. Briefly, in consecutive cycles, the structure of lowest ΔG (AutoDock 4) or "FullFitness" (EADock) was selected and the neighboring docked ligand structures within 5 Å RMSD were collected in the rank, then a new rank was opened with the lowest energy of the remaining docked structures, etc. The ranking was continued until all 100 (AutoDock 4) or 250 (EADock) docked ligand structures were used up in a rank. RMSDs from the crystallographic ligand structures were calculated for the lowest energy (representative) members of each rank. The ranks of the lowest RMSD values are listed in the tables of the Supporting Information. A full list including all ranks is also provided in the Supporting Information. For comparison, the ranking was performed with the "SimpleFitness" scoring of EADock, as well.

Pocket search

The heavy atoms of the protein structures were used as inputs in all cases. The off-line version of Sitehound was applied. In Sitehound and Q-SiteFinder, grid maps are calculated for the probes covering the entire proteins with 1 and 0.9 Å spacing, respectively. Sitehound was tested with both carbon and phosphate probes, whereas Q-SiteFinder applies a methyl probe. In case of Q-SiteFinder and PocketFinder the server produced the top 10 binding sites. Sitehound and Q-SiteFinder ranks the results according to the TIE, which is the sum of nonbonded interaction energy of all probe points with the protein atoms in the detected binding site. PocketFinder and Pass use probe spheres, measuring how much the spheres are geometrically buried in protein pockets and ranks the pockets according to the number of well-buried probe spheres. For all methods, the default parameters were used. For the present evaluations, distances between the centers of predicted pockets and the crystallographic ligand structures were measured for all ranks and methods and the smallest distances and the corresponding ranks are listed in the tables of the Supporting Information. In cases of Q-SiteFinder and PocketFinder precision is calculated as the percentage of the probes of a site that are within 1.6 Å of an atom of a particular ligand, that is, it is a measurement of how well the predicted site maps onto the ligand coordinates (ideally at least 25%).

Water and ligand interferences

The interferences of docked ligands with hydrating water molecules were investigated as follows. All hydrating crystallographic water molecules (Table II) were classified as sitting inside (1), at the bottom of (2) or outside the pocket corresponding to a BD rank. For this, the distances between the crystallographic water oxygen atom and all heavy atoms of the representative ligand structure of the BD rank were measured and the shortest distance was selected. If the shortest distance was smaller than 2.5 Å, that is, the oxygen atom practically overlapped the ligand structure then this water molecule was considered to sit middle-inside the pocket (Type 1). Similarly, the shortest distance of the crystallographic water oxygen atom was measured to the protein and if the distance was smaller than 3.5 Å and the distance from the representative ligand was larger than/equal to 2.5 Å, then the water molecule was considered to sit at the bottom of the pocket, that is, on the protein surface, below the ligand (Type 2). The thresholds 2.5 and 3.5 Å were selected as typical covalent and H-bond lengths between heavy atoms, respectively, with some tolerance. This selection procedure was repeated for all crystallographic water molecules and BD ranks at each complex of Table II. The number of water molecules were counted by rank and summarized by type and complex (Supporting Information). The usefulness of the distinction between the two types of in-pocket water molecules is discussed in the main text. All coordinates of Type 1 and 2 waters are provided as Supporting Information. The interferences of docked ligands (L_j) with other known interfering ligand molecules ($L_{n\neq j}$) were also studied. For this the distance between the centrum of the representative L_j structure of the BD rank and the centrum of a L_n crystallographic ligand structure was measured. If the distance was smaller than 5 Å, then the rank number and distance of interference was tabulated for both BD methods (Supporting Information).

References

1. van Voorhis WC, Hol WGJ, Myler PJ, Stewart LJ (2009) The role of medical structural genomics in discovering new drugs for infectious diseases. PLoS Comput Biol 5:e1000530.
2. Weigelt J, McBroom-Cerajewski LDB, Matthieu Schapira M, Zhao Y, Arrowmsmith CH (2008) Structural genomics and drug discovery: all in the family. Curr Opin Chem Biol 12:32–39.
3. Mirkovic N, Li Z, Parnassa A, Murray D (2007) Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. Proteins 66:766–777.
4. Goldsmith-Fischman S, Honig B (2003) Structural genomics: computational methods for structure analysis. Protein Sci 12:1813–1821.
5. Joachimiak A (2009) High-throughput crystallography for structural genomics. Curr Opin Struct Biol 19: 573–584.
6. Blundell TL, Jhoti H, Abell C (2002) High-throughput crystallography for lead discovery in drug design. Nat Rev Drug Discov 1:45–54.
7. Hetényi C, van der Spoel D (2006) Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett 580:1447–1450.
8. Hetényi C, van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11:1729–1737.
9. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19: 1639–1662.
10. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 10:293–304.
11. Gutierrez LJ, Enriz RD, Baldoni HA (2010) Structural and thermodynamic characteristics of the exosite binding pocket on the human bace1: a molecular modeling approach. J Phys Chem A 114:10261–10269.
12. Young GT, Zwart R, Walker AS, Sher E, Millar NS (2008) Potentiation of α7 nicotinic acetylcholine receptors via an allosteric transmembrane site. Proc Natl Acad Sci USA 105:14686–14691.
13. Othman R, Kiat TS, Khalid N, Yusof R, Newhouse EI, Newhouse JS, Alam M, Rahman NA (2008) Docking of noncompetitive inhibitors into dengue virus type 2 protease: understanding the interactions with allosteric binding sites. J Chem Inf Model 48:1582–1591.
14. Iorga B, Herlem D, Barré E, Guillou C (2006) Acetylcholine nicotinic receptors: finding the putative binding site of allosteric modulators using the "blind docking" approach. J Mol Model 12:366–372.

15. Espinoza-Fonseca LM, Trujillo-Ferrara JG (2006) The existence of a second allosteric site on the M1 muscarinic acetylcholine receptor and its implications for drug design. Bioorg Med Chem Lett 16:1217–1220.
16. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand binding site prediction and functional annotation. Proc Natl Acad Sci USA 105:129–134.
17. Laurie ATR, Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. Curr Prot Pept Sci 7:395–406.
18. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791.
19. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28:1145–1152.
20. Grosdidier A, Zoete V, Michielin O (2009) Blind docking of 260 protein–ligand complexes with EADock 2.0. J Comput Chem 30:2021–2030.
21. Grosdidier A, Zoete V, Michielin O (2007) EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins 67: 1010–1025.
22. Hendlich M, Rippmann F, Barnickel G (1997) Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15:359–363.
23. Hernandez M, Ghersi D, Sanchez R (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res 37:W413–W416.
24. Laurie ATR, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916.
25. Jackson RM (2002) Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. J Comput Aided Mol Des 16:43–57.
26. Brady GP, Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 14:383–401.
27. Ghersi D, Sanchez R (2009) Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins 74:417–424.
28. Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 10:378–391.
29. Stewart JJP (1990) Mopac: a general molecular orbital package. Quant Chem Prog Exch 10:86.
30. Connolly ML (1985) Computation of molecular volume. J Am Chem Soc 107:1118–1124.

D3

hetenyi.csaba_83_23

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 13 June 2017
Accepted: 17 October 2017
Published online: 01 November 2017

Dynamic changes in binding interaction networks of sex steroids establish their non-classical effects

Mónika Bálint[1,2], Norbert Jeszenői[3], István Horváth[4], István M. Ábrahám[3] & Csaba Hetényi[1]

Non-classical signaling in the intracellular second messenger system plays a pivotal role in the cytoprotective effect of estradiol. Estrogen receptor is a common target of sex steroids and important in mediating estradiol-induced neuroprotection. Whereas the mechanism of genomic effects of sex steroids is fairly understood, their non-classical effects have not been elucidated completely. We use real time molecular dynamics calculations to uncover the interaction network of estradiol and activator estren. Besides steroid interactions, we also investigate the co-activation of the receptor. We show how steroid binding to the alternative binding site of the non-classical action is facilitated by the presence of a steroid in the classical binding site and the absence of the co-activator peptide. Uncovering such dynamic mechanisms behind steroid action will help the structure-based design of new drugs with non-classical responses and cytoprotective potential.

Estrogens are responsible for a wide range of biological actions from the regulation of fertility to cytoprotection[1–3]. Gonadal 17β-estradiol (E2) has a remarkable neuroprotective potential[4]. Besides slow, classical, genomic effects[5,6] (Fig. 1) E2 also exerts rapid, non-classical effects on intracellular second messenger molecules[7–10], via estrogen receptors (ERs, Fig. 1).

Importantly, neuroprotection of E2 is attributed to such rapid actions[11–14] and its binding to estrogen receptor alpha (ERα)[15]. Previously we have shown that a single dose of E2 as well as Activators of Non-Classical Estrogen-Like Signaling (ANCELS) such as estren-3α,17β-diol (EN)[16] induce ERα-dependent neuroprotection via intracellular signaling pathways in neurodegenerative animal model[17,18]. The protective effect was also observed after traumatic brain injuries[4] in rodents. Clinical studies showed that hormone replacement therapy with estrogen and progestin[1] decreases the incidence of neurodegenerative diseases such as Alzheimer's disease, but it also increases risks of stroke and breast cancer. However, structural dynamics of biding events establishing non-classical E2 action on ERs has not been fully elucidated. The lack of such details of molecular mechanisms of neuroprotective actions of estrogens hinders the exploitation of their therapeutic potential.

Estrogen binding to the classical binding site (CBS) of human estrogen receptor alpha (hERα) is well-explained by atomic resolution structures of the Protein Databank (PDB)[5,19]. The CBS is located between helices H3, H4, H6, H8 and H11[20] (Supplementary Video S1) of the ligand-binding domain (LBD) of hERα, and it is known to mediate the slow, genomic actions of ligands, such as the native agonist E2 and antagonist 4-OH-tamoxifen selectively modulate gene expression[21].

Besides slow, genomic actions (Fig. 1 top) ANCELS such as EN[22], substance A and substance B[23] exhibit weak transcriptional activity, selectively activating the non-classical E2 signaling as validated by functional assays[22,23]. Such non-classical actions of E2 on the signaling system have been known for more than forty years[24]. However, the underlying mechanism has not been understood due to the lack of atomic resolution structures of the complexes of effector ligands and ERs. An interesting study[25] proposed an alternative binding site (ABS) of E2 and EN on hERα, further discussed by Norman and co-workers[26], conveying the non-classical actions, analogously to vitamin D receptor[25]. The proposed ABS is located at the C terminus of H1 and N terminus of H3 helices, with

[1]Department of Pharmacology and Pharmacotherapy, University of Pécs, Szigeti út 12, 7624, Pécs, Hungary. [2]Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117, Budapest, Hungary. [3]MTA NAP-B Molecular Neuroendocrinology Group, Institute of Physiology, Szentágothai Research Center, Center for Neuroscience, University of Pécs, Szigeti út 12, 7624, Pécs, Hungary. [4]Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720, Szeged, Hungary. Mónika Bálint and Norbert Jeszenői contributed equally to this work. Correspondence and requests for materials should be addressed to I.M.Á. (email: istvan.abraham@aok.pte.hu) or C.H. (email: csabahete@yahoo.com)

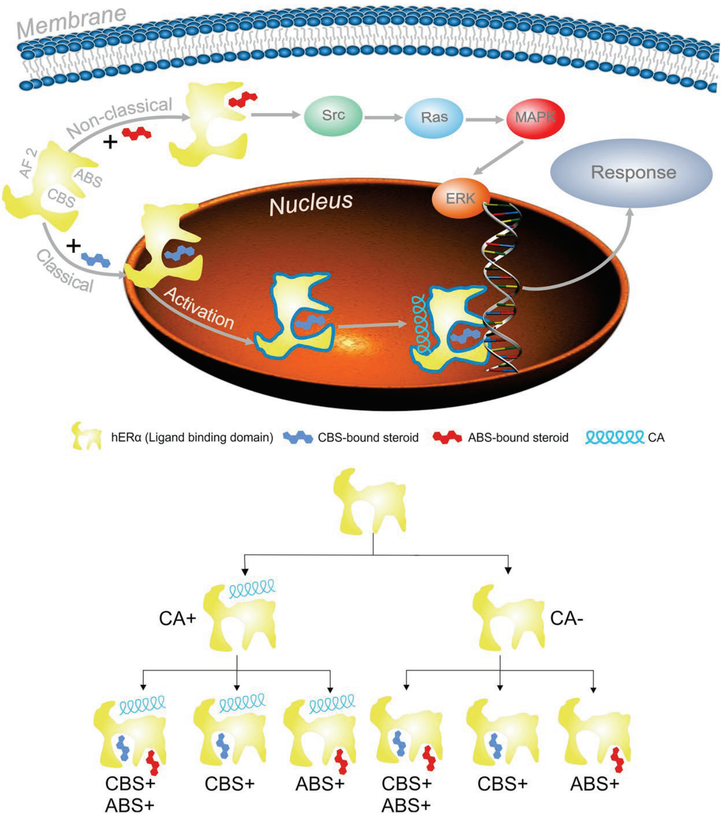


Figure 1. Effects of sex steroids in the cell (top) and their possible binding scenarios to human estrogen receptor alpha experimented in our study (hERα, bottom). Binding site of the co-activator (CA) is marked as AF2. Sex steroids can bind to classical (CBS) and alternative binding sites (ABS) as marked on the schematic representation of the ligand binding domain of hERα. In the classical pathway, activation of hERα by steroid binding to CBS is necessary for subsequent CA binding to AF2. In the non-classical pathway, steroid binding to ABS initiates signal transduction via Src, Ras proteins of the mitogen-activated protein kinase (MAPK) pathway.

a conserved R residue (R274 in vitamin D receptor and R394 in hERα) in the site. E2 binding to ABS[26] does not directly alter gene expression, but rapidly activates the mitogen-activated protein kinase/extracellular-signal regulated kinase (MAPK/ERK) signaling pathway instead (Fig. 1, top)[8,9].

Previous studies[25,27] identified R394 and E353 as key E2-binding residues of ABS, located at the proximity of 3-hydroxyl group of E2, while the other, 17-hydroxyl group is oriented to R335[25]. From these results, a conformational ensemble model was constructed[26] to explain the different behaviour of the nuclear and membrane associated forms of hERα. In this model, a "concurrent occupancy" was also proposed, when both ABS and CBS sites are simultaneously occupied by two copies of E2. However, the dynamics of simultaneous occupancy has not been investigated yet.

Besides ABS and CBS, there is a binding site for different transcriptional co-activator proteins. A conserved, LXXLL binding motif can be found in the amino acid sequences of these proteins[28]. Receptors are often co-crystallized with a peptide fragment of the co-activator (CA) protein containing the above conserved sequence bound to the activation function site 2 (AF2 site, Fig. 1 top part)[5,29,30]. In these structures, CA bridges between helices H3 and H12[20,31] via hydrogen bonding at residues K362 on the H3 side and E542 on the H12 side. Furthermore, if E2 binds, and hER is activated (Fig. 1 top), the CA bridge fixes H12 in a position covering the E2-bound CBS[20,26] and shielding it from the bulk solvent. Y537 plays an important role in the activation, and it was demonstrated that it is very prone to mutations (Y537S) which make the receptor resistant to estrogen antagonist drugs[30]. H3 residues E353, H356, M357 and W360 are proposed to form the ABS, and therefore, any perturbation of the conformation of H3 at these residues by CA can influence the binding of ligands to ABS, as well. Despite the importance of the above effects of the CA-bridge on E2 binding, the dynamics of the underlying mechanism, and the route of structural communication between the proposed ABS[25,26] and CA has not been elucidated at atomic level.

Although the current cutting edge super-resolution imaging techniques such as single molecule fluorescence resonance energy transfer or stimulated emission depletion microscopy are capable to produce sequence of images in given time frame they have limited temporal (5 µs) and spatial (1 nm)[32,33] resolution. Due to the limitations of current structure determination techniques[34,35] investigation of the above questions is fairly challenging and "new techniques may be required to study the formation of such transient, though potentially biologically meaningful complexes" of sex steroids with hERα[16]. At present, molecular dynamics (MD) calculation is the only approach available for investigation of such real time binding events in a receptor-ligand system at atomic resolution. Consequently, several research groups apply MD calculations and present their results on conformational changes of various proteins[36–38] and binding events of ligands[20,39–43] at atomic resolution.

Accordingly, the present study also applies up-to-date, extensive MD calculations to investigate the real time changes of interaction networks of hERα and its ligands at atomic level. The structural dynamics of steroid binding was investigated at both ABS and CBS, taking into account the role of the CA, as well. For this, blind docking of E2 and EN to hERα was performed for an unbiased mapping of available sites. Subsequent MD of the docked complexes surrounded by several thousand of explicit water molecules was applied mimicking the natural dissociation route of the sexual steroids from hERα. The present study also aims at an MD-based elucidation of atomic resolution history of structural changes of ER accompanying non-classical steroid actions.

Results and Discussion

Interaction networks in the steroid-free receptor. To study the effect of CA binding on structural dynamics of hERα (Fig. 2a,b and Supplementary Video S1), both the CA bound (CA+) and free (CA−) structures of the steroid-free LBD were investigated (Fig. 2c) for comparison. The p160-type CA[44,45] with crucial role in gene transcription was invloved in the present study. The C-terminus of the LBD was completed with a region called F domain extending the crystallographic structure using a modeling procedure described in Methods. In both cases, 1μs-long molecular dynamics (MD) calculations were performed to study the structural evolution of the LBD. Evaluations of the resulted trajectories showed (Fig. 2c) high root mean squared fluctuations (RMSF) of amino acid heavy atoms over the entire 1-μs domain at loops L1, L2, and in the F domain. Since loops are naturally flexible regions, and the F domain is a disordered region such fluctuations were expected. The flexibility of L1 can be explained mostly by its high exposition to the bulk. This loop is of high structural importance, as it has an indirect contact with the CBS through S329, and is also closely connected to helix H3, which is covering both the ABS and CBS (Supplementary Video S1).

Having MD results on both CA+ and CA− LBD structures, the influence of CA binding on the LBD was structurally analyzed paying special attention to the CA-connected helices H12, H3 and regions around the binding sites. Both termini of CA are connected to the LBD by salt bridges to E542 of H12, and by an H-bond to K362 of H3. In addition, CA forms hydrophobic contacts with I358 and M357 of H3 and L539 of H12 (Fig. 3a). The hydrophobic contacts with H3 are of particular interest, as I358 is in the vicinity of M357, which is part of the ABS. Therefore, comparison of their movement in CA+/CA− simulations may help to elucidate the mechanism of influence of CA on the process of ligand binding or dissociation to or from ABS. Accordingly, the movements of amino acids H356, M357, and I358 were quantified by calculating the distances between actual and initial positions of their side-chains (Fig. 3 bottom parts) along the MD simulations.

In order to maintain the hydrophobic interactions between the hERα, and CA (Fig. 3a top), I358 situated on H3 fluctuates between a distance of 2–3 Å measured from its initial position as a reference point (Fig. 3 bottom). The fluctuation (Fig. 3a bottom) is higher in the first part (0–400 ns) than in the second part of the simulation (400–1000 ns). The resulted 3 Å shift of I358 from its initial position causes the flipping of M357 into the ABS after 400 ns to further initiate the shift of H356 (Fig. 3a, bottom). In the CA− scenario (Fig. 3b top and bottom), it can be observed, that I358 highly fluctuates during the entire simulation. However, the above-mentioned shift of M357 into the ABS was not observed in the CA− scenario. Thus, the presence of CA can be perceived as a restricting factor, especially on M357. In contrast to M357, the orientation of H356 was not dependent on the presence or absence of CA at the end of the simulations. This can be explained by the contact between H356 (H3) and L327 (L1) through which L327 (L1) transfers its high mobility (Fig. 2c and Supplementary Video S1), to H356 (H3), then M357(H3). It was also found that in both CA+ and CA− simulations, H356 was oriented inside the ABS binding site by the end of 1μs simulation, but this switch occurs faster in the CA+ simulations (400 ns), than in CA− simulations (800 ns, Fig. 3 bottom parts) due to the movement of M357.

In the nucleus, sex steroids[45] bind to the CBS activating hERα which results in the occupancy of AF2 binding site[46] by CA (Fig. 1 top part). Such activation does not occur if hERα resides in the membrane and the AF2 site is left unoccupied. The membrane bound form of the estrogen receptor is involved[47,48] in non-classical effects such as antiapoptotis[16], cytoprotection, and neuroprotection[11] Kousteni and colleagues have also reported rapid, non-classical effect of E2[16], which require the extra-nuclear localization of the hERalpha, confirmed by confocal laser scanning microscopy studies[49]. The above-mentioned antiapoptotis is resulted by targeting[50] an ABS outside the CBS of hERα. Fluorescence experimental studies[51] also indicated the presence of ABS. Thus, ABS is linked to non-classical effects attributed to the membrane-bound form of hERα.

We found that ABS is available for ligand binding if AF2 is not occupied, otherwise it is dynamically blocked by both M357 and H356 side-chains. Thus, receptor dynamics at these two amino acids is responsible for the availability of ABS in membrane surrounding for certain ligands. In agreement with the herein presented results, experimental studies showed[16] that E2, EN and other sex steroids are capable to produce non-classical effects[22], occupying the ABS[27]. For this, sex steroids require extranuclear, membrane-bound localization of the estrogen receptor[22], where the AF2 binding site is not occupied by CA.

Binding sites of sexual steroids. Following structural dynamics investigations on the steroid-free receptor, a complete exploration of binding sites of sex steroids was performed on the entire surface of the apo LBD. Blind docking[52–54] was used for the search as this method does not require previous knowledge of the location of the binding sites. A representative structure of LBD was produced by MD simulation with subsequent clustering (Methods) and used as a target in the blind docking calculations (Fig. 4). The target structure was validated by

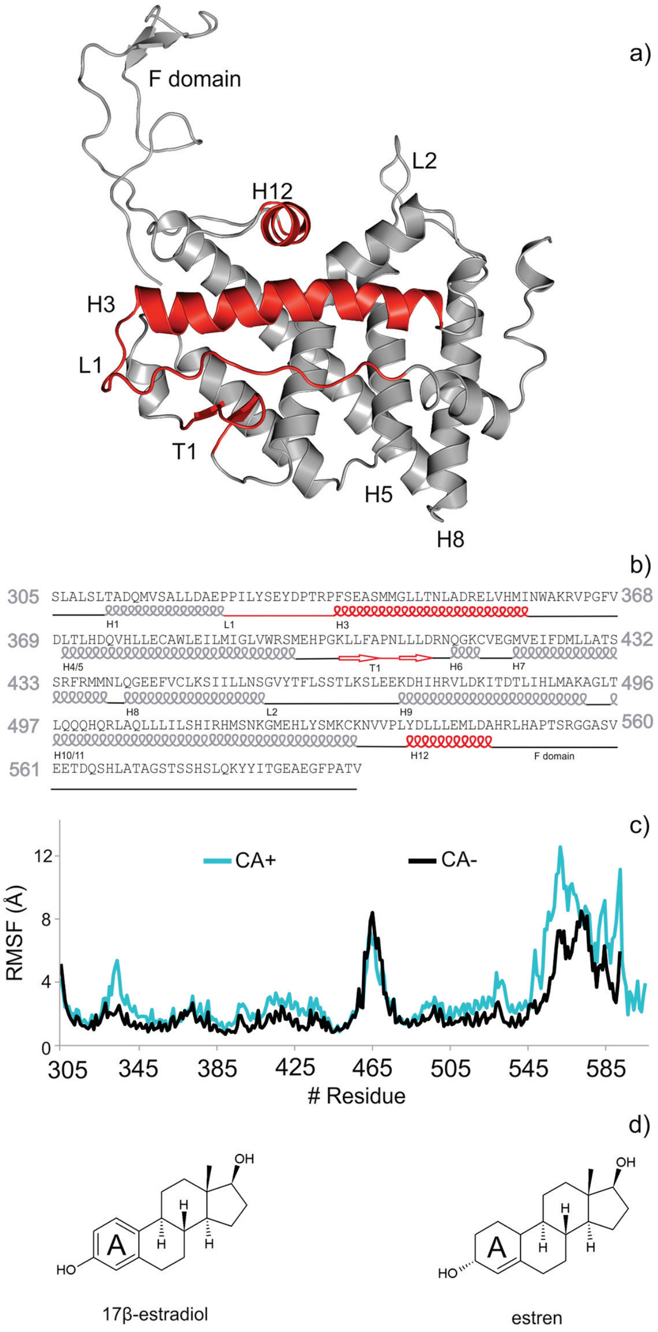

a)
F domain
L2
H12
H3
L1
T1
H5
H8
b)
305 SLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKRVPGFV 368
H1 L1 H3
369 DLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFAPNLLLDRNQGKCVEGMVEIFDMLLATS 432
H4/5 T1 H6 H7
433 SRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLMAKAGLT 496
H8 L2 H9
497 LQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDAHRLHAPTSRGGASV 560
H10/11 H12 F domain
561 EETDQSHLATAGSTSSHSLQKYYITGEAEGFPATV
c)
CA+
CA-
RMSF (Å)
0
4
8
12
305
345
385
425
465
505
545
585
# Residue
d)
OH
H
A
H
H
HO
17β-estradiol
OH
H
A
H
H
HO
estren

Figure 2. Molecules investigated in the present study. (**a**) The three-dimensional structure and the corresponding amino acid sequence (**b**) of ligand binding domain of human estrogen receptor alpha (hERα). Important helices H12, H3, loop L1, and β-turn T1 are highlighted in red. (**c**) Structural flexibility of hERα calculated as root mean squared fluctuations of all amino acid residues during the 1μs molecular dynamics simulations in co-activator bound (CA+) and unbound (CA−) forms of the steroid-free hERα. (**d**) Sex steroids 17β-estradiol, and an activator of non-classical estrogen-like signaling, estren.

blind docking of E2 (Fig. 4, magenta). The docking result was compared with the crystallographic ligand conformation in the CBS (Supplementary Fig. S1). One-hundred blind docking trials were performed with random initial positions of E2 around the target. The results were evaluated as described in previous works[52,53]. Briefly, the docked steroid copies were clustered and ranked by energy, resulting in a list of explored binding sites and ligand

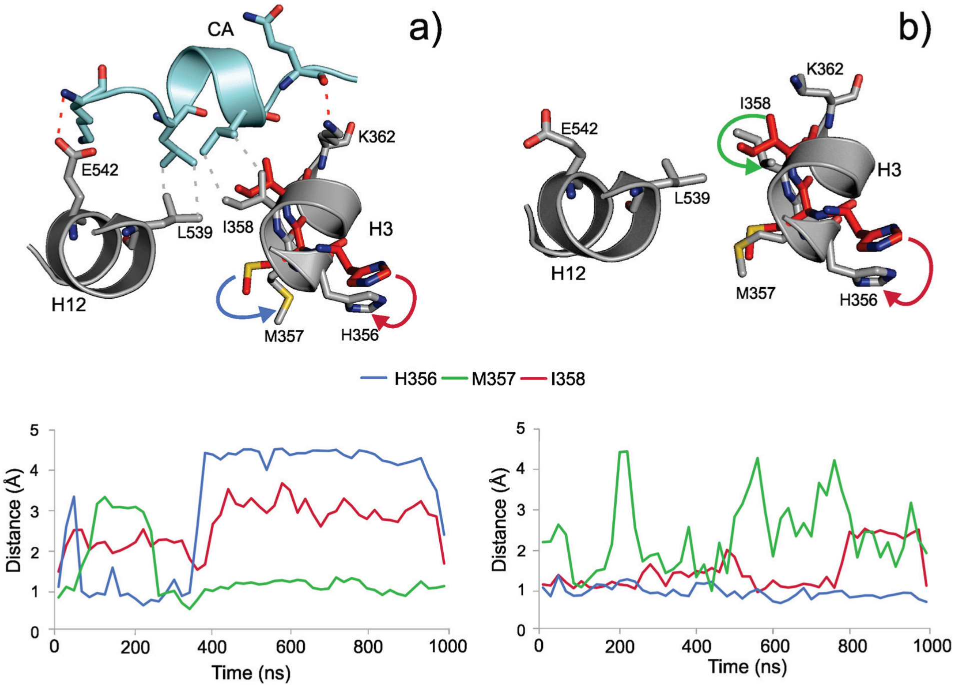


Figure 3. Detailed conformational changes at helices H3 and H12 of the human estrogen receptor alpha (hERα). 1 μs-long steroid-free molecular dynamics (MD) simulations were performed on co-activator bound (CA+, **a**) and unbound (CA−, **b**) hERα structures. The upper part shows conformation of important residues before (red) and after (grey) the MD simulations. CA is presented with cyan cartoon and sticks. A salt bridge, a H-bond (red dotted lines), and hydrophobic interactions (grey dotted lines) can be observed between CA and the helices (H12, H3) of hERα. In the bottom part, actual distances of M357, H356 and I358 from their initial positions are plotted during the MD simulation. Arrows on the upper parts have the same color codes as line charts on the bottom parts of (**a** and **b**).

poses with the strongest steroid-site interaction in the first rank. Besides E2, blind docking of EN (Fig. 4, teal) was also performed on the LBD. From the blind docking calculations 11 ranks were identified for E2 and 6 for EN (Fig. 4, Supplementary Table S1).

The CBS was found in the first rank of blind docking by both steroids. Reproduction of the binding mode of E2 in the CBS was successful as a root mean squared deviation (RMSD) of 2.1 Å (Supplementary Fig. S1) was measured between the heavy atoms of the blind docked and crystallographic (reference) steroid conformations. Such a good fit of the docked E2 to the experimental conformation shows that the target LBD structure is valid and blind docking predictions provide accurate results at atomic resolution. Analysis of docked molecules in CBS revealed that binding modes of E2 and EN are very similar to each-other (Supplementary Fig. S2). Both steroids occupy the same orientation with H-bonds formed between 3-hydroxyl of E2 and EN and hERα residues (F404, E353, and R394). Topologically, CBS is separated from the bulk by loop L1, ß-turn T1, H3 and H12. At the same time, structural differences between the steroids influence their hydrophobic interactions with the amino acids in the surroundings (Supplementary Fig. S2). For example, aromatic ring A of E2 (Fig. 2d) forms a perpendicular π-stacking with F404 situated on T1. The lack of aromatic ring in EN results its increased flexibility and weak hydrophobic interactions with F404, if compared to the π-stacking, observed at E2. This could also be part of the reason, why E2 is considered primarily as a CBS-binding ligand[26,44] and is selective for the classical pathway[16].

The ABS was found in the second rank during blind docking of both E2 and EN in a region proposed by previous studies[26,27]. ABS is located between H8 and H3 in the vicinity of the CBS (Supplementary Video S1). Exposition of ABS towards the bulk is higher than that of CBS as it is covered only by the highly flexible L1 (Fig. 2a). Similarly to the "ensemble model" of previous studies[26,27], the BD calculations showed that R394 and E353 separate the two sites (Supplementary Fig. S3). Furthermore, EN is bound to the ABS, with its 3-hydroxyl group oriented towards R394, which also agrees with previous studies[25,26]. Lipophilic residues (P324, L327, M357, W360, I386, P406) dominate this site, K449 is the only amino acid with polar side chain. Comparing the binding modes of the two analyzed steroids (E2 and EN) to the ABS, a head-tail swap can be observed between them (Fig. 4, bottom). Accordingly, a hydrogen bond is formed with the backbone amide of L327 with different groups of the steroids (17-hydroxyl of E2, and 3-hydroxyl of EN). In addition, 17-hydroxyl of EN forms another hydrogen bond with K449. This bond was not observed in the complex with E2. The H-bond with the backbone amide of L327 is common for the two ligands. As L327 is on loop L1 it is exposed to the bulk, mobile and susceptible to the thermal motion of the surrounding water molecules (see also Section Interaction networks in the

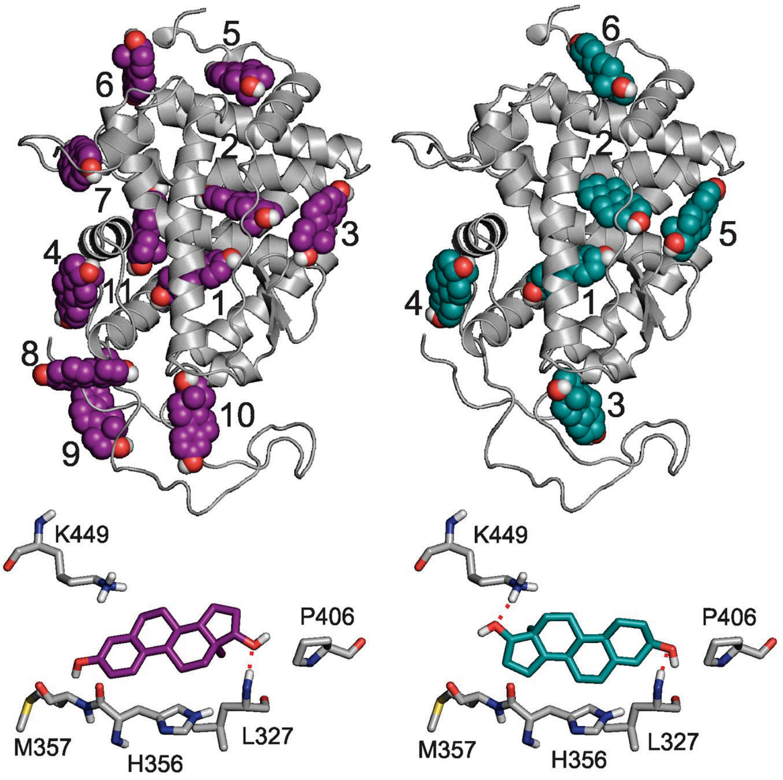


Figure 4. Results of blind docking calculations of steroids E2 (left) and EN (right) to hERα. At the top, cluster representant steroid conformations (spheres) and the corresponding rank numbers are shown. A small rank number corresponds to energetically favorable binding mode of the steroid. The receptor is shown as grey cartoon. At the bottom, a close-up of binding conformation of steroid E2 (magenta) and EN (green sticks) is shown in the ABS (Rank 2 in both cases). Neighbouring hERα residues are shown as sticks with grey carbon atoms.

Ligand	Present study		Mizwicki *et al.* 2004	
	ABS	CBS	ABS	CBS
EN	−37	−33	−66	−61
E2	−32	−35	−61	−66

Table 1. Interaction energies of sexual steroids with hERα (kcal/mol).

steroid-bound receptor and results on simulations with different velocity distributions). At the same time, the second H-bond specific for EN is formed with H8, buried in the pocket, inaccessible from the bulk stabilizing the interaction of EN with the LBD at the ABS. Concerning the location of ABS and CBS the results are in good agreement with previous studies[25–27]. A previous comparison of the binding interaction energies of E2 and EN produced by manual docking[25], showed that binding of EN is stronger to the ABS than that of E2 (Table 1). For the CBS, an opposite trend was observed (Table 1). Other docking[25,27] studies also confirmed E2 selectivity towards CBS. Experimental binding studies demonstrated[22,55] that E2 has a higher affinity towards ERα than EN. *In vitro* experiments[22,55] showed that E2 plays a role in classical effects associated with its CBS[50,51] binding. At the same time, despite the moderate binding affinity of EN[23] *in vivo* studies[22] also confirmed that it has a selectivity towards the non-classical pathway, lacking an effect on the reproductive organs which was confirmed by histological analysis of the uterus, and did not stimulate transcription of the C3 gene in the uterus[22]. In the present study, interaction energies were calculated using the docked and energy-minimized ligand structures. The differences in the energy values show good agreement with those obtained in previous docking (Table 1) and the affinity/selectivity preferences demonstrated by the above-mentioned *in vitro* and *in vivo* experimental studies.

Table 1 shows that EN binds 4 kcal/mol stronger to ABS than to CBS. At the same time, the binding of EN to ABS is 5 kcal/mol stronger than that of E2. This is in agreement with the above structural findings, and also with previous results[25], showing that EN has a larger affinity to ABS than CBS. These results suggest different binding modes at ABS and CBS which is consistent with the structural observations described above (Fig. 4).

All-in-all, for the top two ranks blind docking gave consensus results identifying the binding sites of both steroids as the CBS and the ABS, respectively. Both steroids bind to both sites with significant interaction energies, with E2 a classical effector on CBS, and EN preferring ABS as a non-classical effector[26]. In Rank 3 and beyond, steroids found different sites without a consensus result. Notably, binding of E2 to CBS had been precisely described[5,20] and the position of ABS was proposed in previous studies[25,27]. However, steroid binding to ABS has not been fully characterized. Here, atomic resolution structures of the complexed sites with both investigated ligands bound to ABS were provided (Fig. 4), highlighting crucial amino acids, for non-classical activity, and the binding difference between them. Moreover, binding mode of EN to CBS was also provided (Supplementary Fig. S2) and analyzed. Atomic resolution complex structures from the above blind docking calculations were piped in the investigations of the next Section dealing with the molecular dynamics of interaction networks of steroid binding.

Interaction networks in the steroid-bound receptor

Interaction dynamics. To effect the transcriptional activity in the classical, genomic pathway, a "long-lived"[16] steroid-CBS contact is needed in order to produce the specific conformational changes of hERα. At the same time, steroid ligands form "transient complexes" with the ABS, via a brief association to hERα in the non-classical pathway. However, investigation of such rapid effects of the non-classical pathway requires new approaches and techniques[16].

In the present study, we apply molecular dynamics calculations of the steroid-bound hERα surrounded by several thousand (explicit) water molecules. To investigate the interaction dynamics, docked steroid-bound receptor structures were adopted from Section Binding sites of sexual steroids as starting points. Besides singly occupied binding sites, additional complex structures were constructed (Methods) with both ABS and CBS simultaneously occupied for both EN and E2. All versions were produced both in the presence and absence of CA which yielded altogether twelve different complexes for the two sexual steroids (Fig. 1, bottom). For all complex structures, five parallel 100-ns-long MD calculations were performed to follow their trajectories. Thermal dissociation of the steroid ligands was expected by acquiring kinetic energy from its water and protein surrounding. The calculations were repeated five times using different initial velocity distributions resulting in a total of 6 μs MD calculation. Applying more than one starting initial velocity distribution for a starting structure is important to obtain statistically relevant, unbiased conclusions. In other words, five, independent dissociation trials were performed resulting in five, independent dissociation trajectories of the steroids in all twelve complexes.

From the dissociation trajectories (Fig. 5a, Supplementary Video S2), residence frequency (RF) values were calculated to quantify kinetic stability of the complexes in each trial of Fig. 1 (bottom). In drug discovery, assessment of kinetic stability described by the residence of a ligand in the binding site is crucial factor similarly to thermodynamic stability[56,57]. To calculate RF, the movement of the ligand was described by the distance between the centre of mass (d_{COM}) of its actual and starting positions at each time frame during the simulation time resulting in a COM-plot. The RF value of a binding site was directly obtained (Equation 3, Methods) from the COM-plots (Fig. 5bc) using a $d_{LIM} = 5$ Å for dissociation limit.

Results of the merged trajectories of a total of 500 ns simulation time per trial are listed in Tables 2 and 3. Per-trial and RMSD-based evaluations are presented in Supplementary Tables S2–S5. In the present study, the theoretical upper limit of RF was 10.0 ns^{-1}, which corresponds to the highest kinetic stability. The mean CBS RF values of E2 and EN (last column in Table 2, average of first four columns), are 10.0 ns^{-1} and 9.0 ns^{-1}, respectively. Experimental results are in agreement with our calculations (Table 2) affirming that E2 has a stronger affinity to CBS than EN[22,25]. Results in Table 1 are also in line with a key review by Norman and co-workers[26] presuming that steroids such as EN and E2 could have different "fractional occupancies" in the ABS and CBS pockets. Whereas both ligands show good binding stability at the CBS, a drop in RF values can be observed at ABS (Table 3) if compared with those at CBS (Table 2). In the case of ABS, the mean RF of EN is markedly higher than that of E2.

For structural interpretation of the results in Tables 2 and 3, representative individual trajectories were selected with RFs closest to that of the merged trajectory (bold in Supplementary Tables S2–5). As it was described in Section Binding sites of sexual steroids, EN has H-bonds with both hydroxyl groups, and is stabilized in the ABS at its both ends (Supplementary Video S2 and Fig. 5a). The two H-bonds are formed at the entrance with L327 of loop L1, and K449 of H8 helix, at the bottom of the pocket. Loop L1 is highly exposed to the bulk having a susceptibility to the thermal motion of the hydrating water molecules and it tends to pull out EN from the binding site. At the same time, forming an H-bond with K449, H8 acts as a counter balance and keeps EN in ABS. If the H-bond with K449 is broken, EN will be easily pulled out towards the bulk by the loop. After the breakage of the H-bond between EN and K449, a series of conformational changes are initiated by L327. Firstly, as L327 interacts with the side chain of H356 through hydrophobic interactions and H356 starts to move towards the ABS binding site, as fluctuation of L1 intensifies. Secondly, a conformational change is induced on M357 by H356. Here, the side chain of M357 flips into the ABS binding site, similarly to the apo simulations (see Secion 1). As M357 flips inside de binding site, sterically perturbs EN leading to its expulsion from the site. The above conformational changes were not observed in case of E2, and therefore, no role can be attributed to M357 in its dissociation. As the H-bond with K449 is missing in case of E2, the above described counter balancing effect does not take place. Hence, E2 is pulled out more easily than EN from ABS by the thermal motion of the loop.

The above analyses of the simulation trajectories highlighted that the conformational changes of hERα (Supplementary Video S2 and S3) have crucial role in the dissociation process of EN. In order to quantify the relationship between conformational changes of the receptor and dissociation of EN, d_{COM} was correlated with the movement of three residues (L327, H356, and M357) in the $d_{COM} < 5$ d_{LIM} interval. Correlation results are shown for the CA+/CBS+/ABS+ (Fig. 6) case with representative residue movements. Notably, similar correlations

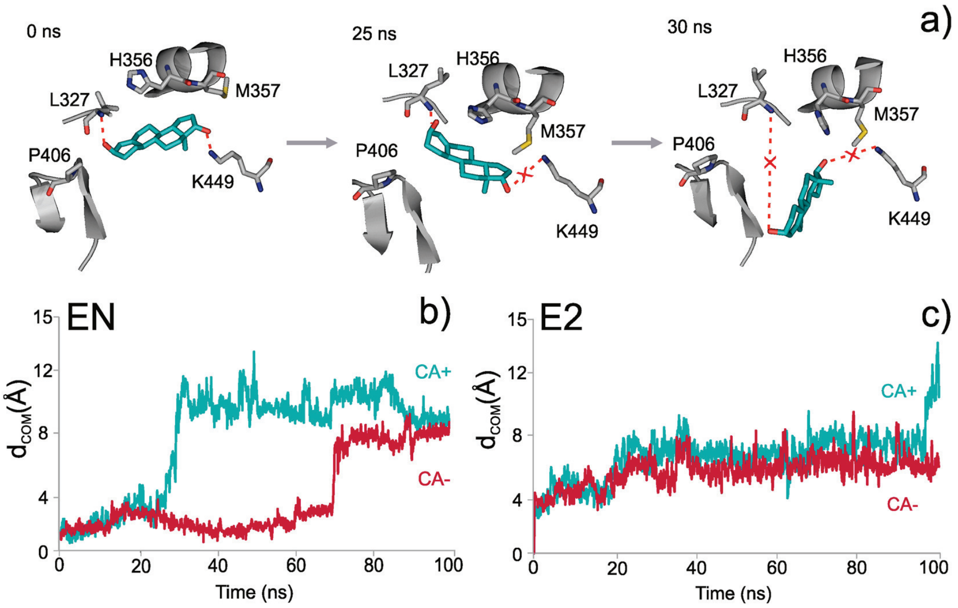


Figure 5. Dissociation of the steroids from the ABS. (**a**) Snapshots of the 100-ns-long molecular dynamics (MD) simulation of EN dissociation from the ABS. The simulation started from a CA+/CBS+/ABS+ starting complex. Disruption of H-bonds with K449, and L327 can be observed at 25 and 30 ns, respectively. Migrations of steroids EN (**b**) and E2 (**c**) out of ABS are represented as actual distances of their center of mass (d_{COM}) measured from their bound, starting position inside ABS. Evaluations of simulations both with (CA+) and without (CA−) the co-activator are shown. In the case of EN (**b**), an abrupt increase of d_{COM} can be observed, at 30 ns (CA+) and at 70 ns (CA−). Thus, the presence of CA promotes the dissociation of EN from ABS. Dissociation of E2 shows a different picture (**c**), as its d_{COM} increases in a stepwise manner, without fast jumps in the starting period of the simulation. This is due to the lack of strong, directed interactions between E2 and ABS.

CA	+		−		**Mean (SD)**
ABS	+	−	+	−	
EN	9.3	9.1	10.0	8.4	9.2 (0.7)
E2	10.0	10.0	10.0	10.0	10.0 (0)
Mean (SD)	9.7 (0.5)	9.6 (0.6)	10.0 (0)	9.2 (1.1)	

Table 2. Residence frequencies of the steroids in CBS (ns^{-1}).

CA	+		−		**Mean (SD)**
CBS	+	−	+	−	
EN	2.6	8.0	7.2	6.5	6.1 (2.4)
E2	1.8	2.9	5.1	2.3	3.0 (1.5)
Mean (SD)	2.2 (0.6)	5.5 (3.6)	6.2 (1.5)	4.4 (3.0)	

Table 3. Residence frequencies of the steroids in ABS (ns^{-1}).

were observed for the CA−/CBS+/ABS+ case (Supplementary Fig. S4), as well. The obtained correlations show that all three proposed residues are important in inducing EN dissociation. Due to their characteristic interaction networks there is a considerable difference in the dynamics of the three side-chains (Fig. 6). While M357 enters the ABS, which results in pushing EN out of its binding pose, L327 exerts a pulling effect on EN from the other side. H356 continuously fluctuates rotating inside the ABS.

In this Section, dissociation mechanisms of sexual steroids from ABS and CBS were uncovered by extensive molecular dynamics calculations. Differences in binding affinities[16,25] (Table 1) and kinetic stability (Tables 2 and 3) of steroid-hERα complexes was correlated with the differences in the dynamics of the corresponding interaction networks.

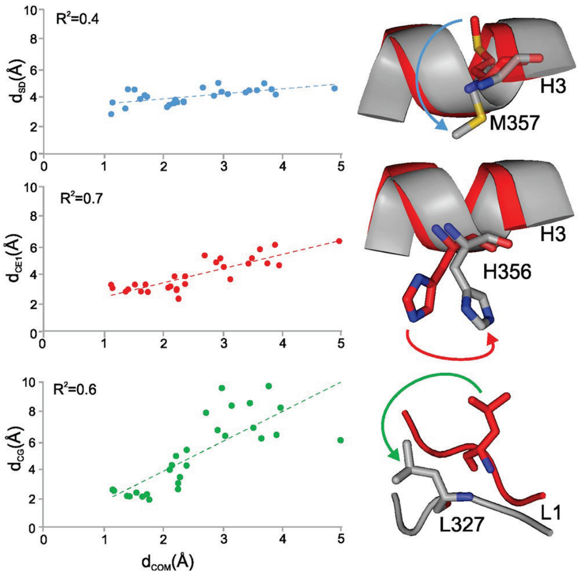


Figure 6. Correlation of movements of residues M357, H356, L327 with dissociation of EN. The movements of the residues are expressed as the distances of SD (M357), CE1 (H356) and CG (L327) atoms from their initial positions. Dissociation of EN is measured by d_{COM}, the distance of the center of mass of the ligand, measured from its initial position. The importance of M357, H356, L327 in the dissociation of EN is indicated by the above obtained correlations. Correlation plots contain points until dissociation ($d_{COM} < d_{LIM}$). The structural representations on the right correspond to $d_{COM} = d_{LIM}$. For simplicity, the points represent average distances calculated as described in Methods.

Effects of co-binders. The steroid-free MD simulations uncovered an interference between ABS and AF2. It was found that the ABS is available for ligand binding only if AF2 is not occupied by CA (Section Interaction networks in the steroid-free receptor). The results of Table 3 also show that occupancy of ABS is influenced by the presence of other ligands co-bound to hERα. Binding of CA to site AF2 or an additional steroid molecule to CBS has considerable effect on steroid binding to ABS (Table 3). In order to investigate the structural background of these effects, we examine how CA affects the binding dynamics of E2 and EN to hERα.

In the CA− scenario, remarkably high stability of E2 and EN binding to ABS was found especially if an additional E2 or EN copy was present in the CBS (CBS+, third column in Table 2). This situation is of particular importance as non-classical effects happen in the absence of CA (Section Interaction networks in the steroid-free receptor). Various experimental studies have suggested that fast, non-classical activity of streoids is exerted by their binding to the ABS[16,50]. It is also known that binding to ABS is not probable in the presence of CA[30,46,58]. Consequently, the presence of CA (CA+) would hinder steroid binding to ABS and facilitate dissociation. Our MD approach allowed the investigation of such a non-natural CA+ situation and the analysis of the reasons of the hindering effect of CA binding to AF2, as well. This finding is consistent with the effect of CA over the ABS binding site (Section Interaction networks in the steroid-free receptor) where the effect of CA bridge connecting helices H3 and H12 was demonstrated. As both helices are very close to the binding sites, they interfere with the CA bridge and ligand binding. CA binds to the LBD via ionic and hydrophobic interactions with H3 at M357 and I359, which are part of the ABS. It was also demonstrated (Section Interaction networks in the steroid-free receptor) that M357 tends to occupy ABS in presence of CA. The same mechanism was observed also in the ABS+ simulations, but only in case of EN, which indicates a dependency on the ligand type (see also Section Interaction dynamics, Fig. 5). If CA is present, M357 tends to move towards the center of the ABS, and I359 assists this process providing a steric restraint and keeping a hydrophobic contact with L693 of CA. On the other hand, if CA is missing from the above interaction networks, its influence on I359 and M357 is not there. Thus, I359 can move freely, and therefore, M357 can maintain its orientation towards the bulk, and it does not influence the stability of EN binding. All-in-all, CA changes the dynamic interaction network of the ABS leading to kinetic stability differences presented as RFs in Table 3. Although H356 has no direct contact with CA it also plays an important role in the dissociation mechanism as it is in the vicinity of M357, and also occupies the ABS promoting the dissociation of EN (Fig. 5a and Supplementary Video S2).

The above structural effects are also reflected by the velocity of EN during the dissociation process from ABS as calculated from the COM-plot (Fig. 5bc). The overall dissociation velocity of EN increased from 0.11 to 0.25 Å/ns (Supplementary Tables S6 and S7) in the CA+ case, due to the described destabilizing effect of CA binding to hERα. The dissociation process of EN can be divided into an initial ($d_{COM} \leq d_{LIM}$) and a terminal ($d_{LIM} < d_{COM} \leq 10$ Å) phase. Velocity of EN in the terminal phase is larger than it was in the initial phase, which is specific to EN. EN has higher v_2 values than E2 suggesting that final dissociation of E2 from ABS occurs slower

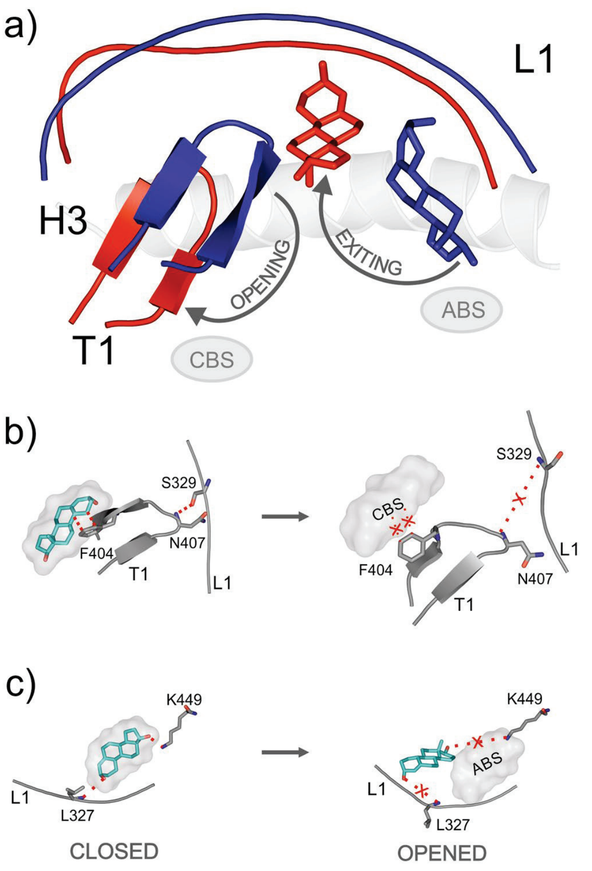


Figure 7. Dynamics of the flickering gate. (**a**) The three-dimensional structure of the flickering gate is composed by T1, L1, and H3 structural elements. Its closed (blue) and opened (red) states show considerable differences at T1 and L1 conformations. EN is represented with blue (ABS-bound) and red (unbound) sticks. (**b**) The absence of EN (sticks with green cartoon) from the CBS (grey cloud) results in the breakage of N407 and S329 in the opened state, releasing L1. (**c**) Consequently, L1 will not maintain its conformation necessary for the interactions with EN, which will readily leave ABS. (**b,c**) All atom coloring is used for sticks of EN and target residues (grey carbon). Red dotted lines highlight H-bonds, and hydrophobic interactions. Interactions present in the closed state, and absent in the opened state, are marked with red cross.

than in case of EN. The characteristic, abrupt movement of EN in the terminal phase can be explained by the sudden of breakage of the second stabilizing H-bond, the one with K449.

The effect of the presence of an additional steroid molecule in the CBS (CBS+) is coupled to that of the absence of CA and this CA−/CBS+ case shows the highest stability of the ABS-bound ligands (Table 3). To understand the effect of occupancy of the CBS simulations on EN were analysed for both CBS+ and CBS− cases (Supplementary Table S3, seed 1). Three structural elements T1, L1 and H3 were of particular interest, in analyzing the stabilizing effect of CBS over the ABS. These elements can be considered as parts of "flickering gate" (Fig. 7a and Supplementary Video S3) as proposed by a previous study[25]. T1 plays the role of the flickering wing, whereas L1 and H3 constitute the stable frame of the gate. Our calculations show that the gate is closed when CBS is occupied (Supplementary Video S3, blue), and opened when CBS is unoccupied (Supplementary Video S3, red). When EN binds to the CBS it is able to keep the "flickering gate" in a closed state, as it interacts with T1 (flickering wing) via a hydrophobic interaction with F404 (Fig. 7b). Therefore, in the closed state, stabilization of T1, by EN in the CBS, will further maintain an H-bond between T1 (N407) and L1 (S329). Stabilized by this H-bonding between T1 and L1 (Fig. 7b), L1 becomes less flexible, and its rigidity will further increase RF of EN in the ABS (Fig. 7c). This happens as L1 binds to EN in ABS via L327 (Fig. 7). See also Section Interaction dynamics showing that the movement of L327 correlates with ligand dissociation (Fig. 6). We found that the flickering gate adopts an opened state if CBS is not occupied (Fig. 7b). This is a consequence of the lack of hydrophobic interaction between T1 (F404) and the CBS-bound EN. The H-bonding between L1 (S329) and T1 (N407) becomes disrupted, and therefore, flexibility of L1 increases. As discussed above, a flexible L1 promotes dissociation of EN from ABS and lowers the corresponding RF (Table 3).

The above dynamic interaction network, especially between CBS, T1, L1, and ABS describe the working mechanism of the "flickering gate"[25]. Beside providing a detailed description of the opened and closed state of the gate, we were also able to detect two dissociation pathways of EN during exiting ABS. The first dissociation pathway towards F404 and P406 of T1 is shown in Supplementary Video S3, and the second pathway towards P323 of L1 can be followed in Supplementary Video S2. Both dissociation pathays require the opened state of the "flickering gate"[25]. In addition to the kinetic stability data of Tables 2 and 3, MD allowed the above in-depth analyses of changes of interaction networks at the ABS. The present approach provides a structural background of stability differences pointing to key residues of hERα affecting non-classical steroid action.

Conclusions

In the present study, elucidation of structural dynamics of non-classical effects of sex steroids was presented. Both classical and alternative binding modes were exhaustively mapped on the ligand-binding domain of human estrogen receptor alpha. Kinetic stability of the steroid –receptor complexes was investigated by molecular dynamics calculations. Real-time investigations of the complete interaction network at atomic resolution pointed to key residues of steroid binding mechanism. We showed how steroid binding to the alternative binding site of non-classical action is facilitated by the presence of a ligand in the classical binding site and the absence of the co-activator peptide. Uncovering such dynamic mechanisms behind steroid action will help the structure-based design of new drugs with rapid, non-classical responses.

Methods

Steroid-free systems.

Selection of target structure. There are 137 hERα LBD entries available in the Protein Databank (PDB, Supplementary Table S8) and among them structure 3q95 has the most amino acids solved with a good resolution of 2.05 Å. The 3q95 structure is co-crystallized with the native ligand (estriol) and CA. As 3q95 is the most complete structure, it was chosen to represent hERα LBD. The ligand-free hERα (2b23) is also available (Supplementary Table S8) and superimposing 2b23 and 3q95 on their backbone atoms with PyMol[59] has an excellent overall structural fit quantified by a root mean squared deviation (RMSD) of 0.5 Å. The RMSD was calculated between the two conformations according to Equation (1).

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{NH}}\sum_{\mathrm{i}=1}^{\mathrm{NH}}\left|\overrightarrow{\mathrm{C1}}_{\mathrm{i}} - \overrightarrow{\mathrm{C2}}_{\mathrm{i}}\right|^{2}} \tag{1}$$

where NH is the number of heavy atoms, C1 and C2 are space vectors of the i$^{\mathrm{th}}$ heavy atom of conformations 1 (C1) and 2 (C2), respectively.

Secondary structure prediction was performed on the amino acid sequence of the missing F-domain, the sequence was accessed from UniProt with accession ID of P03372, multiple sequence alignment was performed with Clustal Omega[60]. Prediction was performed on the PsiPred server[61], with the last two amino acids from X-ray structure added, to facilitate the fitting onto the protein after MD. Based on this prediction, the tertiary structure of the polypeptide chain was modelled with Tinker and equilibrated by a 10-ns-long molecular dynamics simulation. After equilibration, further 100 ns, unrestrained MD trajectory was generated for production (see next Section for details). After clustering, the representative structure of the C-terminal region, was merged with both X-ray structures of HERα (3q95 and 2b23) and these extended proteins were used throughout this study.

Both the ligand free and ligand bound PDB entries are appropriate representations of the LBD structure as E2 and estriol do not induce significant changes in the protein structure. In Section Interaction networks in the steroid-free receptor, the extended 2b23 was used, the holo simulations of Section Interaction networks in the steroid-bound receptor were performed with 3q95. The RMSF plot of 3q95 (1 μ simulation, without ligand, Supplementary Fig. S5) shows that overall dynamics of this protein structure is similar to that of 2b23.

Preparation of systems for energy minimization. Structures were solvated with the gmx solvate module of GROMACS 5.0.2[62] in a dodecahedral box with box edges 1 nm from the solute. Missing residues of 2b23 (except the C-terminal region) were not modelled. The box was filled with explicit TIP3P waters[63]. Parameters from the Amber99SB-ILDN[64] force field were used. Sodium or chloride counter ions were added to neutralize the system. The N-terminal region of the receptor proteins was capped; the co-activator peptide was modelled with charged termini.

Energy minimization. The optimization of the simulation boxes prior MD and docking calculations were done in two steps. This procedure was applied for all cases. In the first step a steepest descent minimization was performed on the solvated box, with convergence threshold set to 10^3 kJmol^{-1}nm^{-1}. It was followed by a conjugate gradient minimization, in this step, the convergence was set to 10 kJmol^{-1}nm^{-1}. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJmol^{-1}nm^{-2} in both steps.

Molecular dynamics (MD). After minimization, prior to the productive GROMACS MD calculations, a uniform equilibration procedure was performed. The optimized structure was equilibrated under NPT conditions for 10 ns (with 2 fs time step). The solvent and the solute was coupled separately to 300 K with the velocity-rescaling algorithm[65], with time constant of 0.1 ps. Pressure was kept at 1 bar with the Berendsen barostat[66] with time constant of 0.5 ps, and compressibility of 4.5×10^{-5} bar^{-1}. Long range interactions were cut off at 1.1 nm. Position restraints of 1000 kJmol^{-1}nm^{-1} were applied on all protein heavy atoms. After equilibration, productive NPT MD calculations were started using GROMACS, with position restraints removed. Pressure was coupled with the Parrinello-Rahman barostat[67] with time constant of 0.5 ps, and compressibility of 4.5×10^{-5} bar^{-1}. The temperature was coupled to 300 K with the velocity-rescaling algorithm[65], with time constant of 0.1 ps, with solvent and solute coupled separately. Coordinates were saved at regular time-intervals, at every 10 ps. Simulation on the ligand free structures were 1 μs-long, the terminal loop was simulated for 100 ns. Periodic boundary conditions were treated after the finish of the calculations.

Evaluation of MD results. A ligand free simulation of 1μs length contains 10^5 frames. RMSF calculation was performed with GROMACS gmx rmsf program. RMSF values of 462–471 and 297–300 residues in Fig. 2c were obtained from simulations with 3q95 (Supplementary Fig. S5). Distance calculations from the initial position of M357 (SD), H356 (CE1) and L327 (CG) sidechain atoms was followed throughout the 1μs steroid-free simulation. The distance was calculated using GROMACS rms program, having an alignment of heavy atoms on the initial structure, over H3 (341–361) and H12 (539–545) residues. For efficient presentation in Fig. 3 (bottom part) and Supplementary Fig. S6 average distances were plotted for every 200 frames.

Binding site search with blind docking. *Preparation of the target.* The most populous cluster from the 3q95 simulation after 100 ns with the modelled C-terminal was used as the target structure. Clustering was performed with Gromacs program cluster using the gromos method, and a 2 Å cut-off RMSD was set between clusters. Only polar hydrogens were treated explicitly, non-polar hydrogens were merged. Gasteiger-Marsili charges[68] were added to the protein.

Preparation of the ligand. The first step was a steepest descent optimization, with 10^4 steps. The next step was a conjugate gradient minimization, with a maximum of 10^4 steps, the with convergence threshold set to 10^{-7} kcal-mol^{-1}Å^{-1}. MMFF94 force field[69] was used in both steps. The third and last step was performed on semi-empirical quantum mechanical level with MOPAC2012[70] with PM7 parametrization[71]. Gradient norm was set to 0.01 kcal-mol^{-1}Å^{-1}. After optimization, force calculations were carried out, ensuring that in all cases, the force constant matrices were positive definite. This optimized structure was used in the dockings with Gasteiger-Marsili charges added.

Calculation of grid maps. The grid box around the protein was generated with AutoGrid 4.2[72]. The box was centred to cover the whole protein with 200 grid points along all axes, with a spacing of 0.375 Å.

Blind docking. Blind docking calculations[52–54] of the two steroids (E2 and EN) were performed. Docking calculations were performed with AutoDock 4.2.3[72], Lamarckian genetic algorithm with Solis-Wets local search was used in geometrical search. Dockings started with a population size of 250, the number evaluations were 10^7, and the number of generations was set to 10^7. 100 runs were performed in one docking. For RMSD calculation, between the crystallized and the docked estradiol, 1gwr hHERα was used, where estradiol is the co-crystallized ligand (Supplementary Fig. S1). The estradiol structure from 1gwr was taken after superimposing the Cα atoms of 1gwr on to the Cα atoms of hERα structure used for docking.

Calculation of interaction energy (E_{inter}). Calculation E_{inter} between docked steroids and hERα (Section Binding sites of sexual steroids) was performed after energy minimization with Gromacs (see previous Section, Energy minimization) of the docked complexes. A Lennard-Jones potential (Equation 2) was used with Amber parameters[73].

$$E_{inter} = \sum_{i,j}^{N_T N_L} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \tag{2}$$

$$A_{ij} = \varepsilon_{ij} R_{ij}^{12}$$

$$B_{ij} = 2\varepsilon_{ij} R_{ij}^{6}$$

$$R_{ij} = R_i + R_j$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$$

where N_T: number of target atoms, N_L: number of ligand atoms, r_{ij}: actual inter-nuclear distance, ε_{ij} = potential well depth at equilibrium between i and j atoms types combined from individual well depths, R_{ij} = inter-nuclear distance at equilibrium between i and j atom types combined from individual radii.

Interaction dynamics of steroid-bound systems. *Molecular dynamics.* The conditions of MD simulations were the same as described at the steroid-free calculations, except that the present steroid-bound trajectories were 100-ns-long each, and 1001 frames were sampled per trajectory. After each trajectory the periodic boundary effects were handled, the system was centred in the box and target molecules in subsequent frames were fit on the top of the first frame. In order to compare the "Open" and the "Closed" state between each other (Fig. 7a), after handing the periodic boundary effects, the first frame of "Open" state was superposed on the "Closed" state by their Cα atoms.

Kinetic stability. Residence frequency (RF, Equation 3) was calculated as a measure of kinetic stability. The movement of the ligand was described by the distance between the centre of mass (d_{COM}) of its actual and starting positions at each time frame during the total simulation time.

$$\mathrm{RF} = \frac{\text{Count of time frames with } d_{COM} \leq \ d_{LIM}}{\text{Simulation time (ns)}} \tag{3}$$

The value of dissociation limit d_{LIM} was set to 5 Å. The RF values were calculated for the five individual trajectories and also for a merged trajectory of 500 ns including all five trajectories. The theoretical upper limit of RF was 10.0 ns^{-1} (=1001/100 ns) in the present study which corresponds to the highest kinetic stability.

Correlation of movements of M357, H356, L327 residues, with dissociation of EN. The distance of the side chain atoms from the initial positions were calculated using Gromacs rms program, having an alignment of heavy atoms, on the initial structure, over H3 (341–361) and H12 (539–545) residues. Using the same technique as in the steroid-free evaluations (Methods), for efficient presentation, average distance values were calculated for every 10 frames, resulting in 100 distances for 100 ns of simulation. Correlation of movements of M357, H356, L327 residues, with dissociation of EN was followed when $d_{COM \leq}d_{LIM}$. The d_{LIM} corresponded to 27.2 ns, and correlation points (Fig. 6), were taken from 0 to 28 ns. Average distances of the investigated time interval (0–28 ns) resulted in 1.2 Å initial, and 4.9 Å final d_{COM} values, shown as abscissa in Fig. 6. Up until this frame, the ligand dissociation could be correlated with all three residues, and this is also the point when EN starts to abruptly dissociate from ABS (Fig. 5). The structural representations in Fig. 6, were taken from 18 ns. This was the frame when the movement of all three residues was the most representative.

Velocity calculations. In order to characterize the dissociation patterns of both E2 and EN, three types of velocities were calculated, and presented in Supplementary Table S6–7. The v_1 measures the ligand velocity in the initial dissociation phase, until d_{LIM} is reached. The second type of velocity (v_2) takes into account the necessary time for the ligand to reach total dissociation after reaching the d_{LIM}. The limit for final dissociation was set to 10 Å, and the time when this limit is reached, was collected in Supplementary Table S6. The v_2 characterizes the best, the differences between EN and E2 dissociation mode. In CA− simulations, d_{LIM} was not was not reached for E2, and therefore, v_2 was not calculated. The third type of velocity (v_3) describes the ligand velocity for the total dissociation, from the start of the simulation.

Data availability statement. The datasets generated during and/or analysed during the current study are included in this published article (and its Supplementary Information files) or available from the corresponding author on reasonable request.

References

1. Rossouw, J. E. *et al.* Risks and benefits of estrogen plus progestin in healthy postmenopausal women - Principal results from the Women's Health Initiative randomized controlled trial. *Jama-J Am Med Assoc* **288**, 321–333 (2002).
2. Moosmann, B. & Behl, C. The antioxidant neuroprotective effects of estrogens and phenolic compounds are independent from their estrogenic properties. *Proc. Natl. Acad. Sci. USA* **96**, 8867–8872 (1999).
3. Morale, M. C. *et al.* Estrogen, neuroinflammation and neuroprotection in Parkinson's disease: Glia dictates resistance versus vulnerability to neurodegeneration. *Neuroscience* **138**, 869–878 (2006).
4. Dubal, D. B. *et al.* Estrogen receptor alpha, not beta, is a critical link in estradiol-mediated protection against brain injury. *Proc. Natl. Acad. Sci. USA* **98**, 1952–1957 (2001).
5. Bourguet, W., Germain, P. & Gronemeyer, H. Nuclear receptor ligand-binding domains three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* **21**, 381–388 (2000).
6. Gronemeyer, H., Gustafsson, J. A. & Laudet, V. Principles for modulation of the nuclear receptor superfamily. *Nat. Rev. Drug. Discov.* **3**, 950–964 (2004).
7. Losel, R. & Wehling, M. Nongenomic actions of steroid hormones. *Nat. Rev. Mol. Cell Bio.* **4**, 46–56 (2003).
8. Singer, C. A., Figueroa-Masot, X. A., Batchelor, R. H. & Dorsa, D. M. The mitogen-activated protein kinase pathway mediates estrogen neuroprotection after glutamate toxicity in primary cortical neurons. *J Neurosci* **19**, 2455–2463 (1999).
9. Song, R. X. D. *et al.* Linkage of rapid estrogen action to MAPK activation by ER alpha-Shc association and Shc pathway activation. *Mol Endocrinol* **16**, 116–127 (2002).
10. Simoncini, T. & Genazzani, A. R. Non-genomic actions of sex steroid hormones. *Eur. J. Endocrinol.* **148**, 281–292 (2003).
11. Arevalo, M. A., Azcoitia, I. & Garcia-Segura, L. M. The neuroprotective actions of oestradiol and oestrogen receptors. *Nat Rev Neurosci* **16**, 17–29 (2015).
12. Mcewen, B. S. & Alves, S. E. Estrogen actions in the central nervous system. *Endocr. Rev.* **20**, 279–307 (1999).
13. Spence, R. D. *et al.* Neuroprotection mediated through estrogen receptor-alpha in astrocytes. *Proc. Natl. Acad. Sci. USA* **108**, 8867–8872 (2011).
14. Behl, C. Oestrogen as a neuroprotective hormone. *Nat Rev Neurosci* **3**, 433–442 (2002).
15. Cordey, M. & Pike, C. J. Neuroprotective properties of selective estrogen receptor agonists in cultured neurons. *Brain Res.* **1045**, 217–223 (2005).
16. Kousteni, S. *et al.* Nongenotropic, sex-nonspecific signaling through the estrogen or androgen receptors: Dissociation from transcriptional activity. *Cell* **104**, 719–730 (2001).
17. Abraham, I. M., Koszegi, Z., Tolod-Kemp, E. & Szego, E. M. Action of estrogen on survival of basal forebrain cholinergic neurons: promoting amelioration. *Psychoneuroendocrinology* **3**, S104–S112 (2009).
18. Kwakowsky, A. *et al.* Treatment of beta amyloid 1–42 (Aβ1–42)-induced basal forebrain cholinergic damage by a non-classical estrogen signaling activator *in vivo*. *Sci. Rep.* **6**, 21101, https://doi.org/10.1038/srep21101 (2016).
19. Ng, H. W., Perkins, R., Tong, W. & Hong, H. Versatility or promiscuity: The estrogen receptors, control of ligand selectivity and an update on subtype selective ligands. *Int. J. Environ. Res. Public Health* **11**, 8709–8742 (2014).
20. Celik, L., Lund, J. D. & Schiott, B. Conformational dynamics of the estrogen receptor alpha: molecular dynamics simulations of the influence of binding site structure on protein dynamics. *Biochemistry* **46**, 1743–1758 (2007).
21. Shiau, A. K. *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927–937 (1998).

22. Kousteni, S. *et al.* Reversal of bone loss in mice by nongenotropic signaling of sex steroids. *Science* **298**, 843–846 (2002).
23. Wessler, S., Otto, C., Wilck, N., Stangl, V. & Fritzemeier, K. H. Identification of estrogen receptor ligands leading to activation of non-genomic signaling pathways while exhibiting only weak transcriptional activity. *J. Steroid. Biochem.* **98**, 25–35 (2006).
24. Szego, C. M. & Davis, J. S. Adenosine 3′,5′-monophosphate in rat uterus: acute elevation by estrogen. *Proc. Natl. Acad. Sci. USA* **58**, 1711–1718 (1967).
25. Mizwicki, M. T. *et al.* Identification of an alternative ligand-binding pocket in the nuclear vitamin D receptor and its functional importance in 1 alpha,25(OH)(2)-vitamin D-3 signaling. *Proc. Natl. Acad. Sci. USA* **101**, 12876–12881 (2004).
26. Norman, A. W., Mizwicki, M. T. & Norman, D. P. Steroid-hormone rapid actions, membrane receptors and a conformational ensemble model. *Nat. Rev. Drug Discov.* **3**, 27–41 (2004).
27. van Hoorn, W. P. Identification of a second binding site in the estrogen receptor. *J. Med. Chem.* **45**, 584–589 (2002).
28. Heery, D. M., Kalkhoven, E., Hoare, S. & Parker, M. G. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* **387**, 733–736 (1997).
29. Warnmark, A. *et al.* Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *J Biol Chem* **277**, 21862–21868 (2002).
30. Fanning, S. W. *et al.* Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *Elife* **5**, e12792, https://doi.org/10.7554/eLife.12792 (2016).
31. Savkur, R. S. & Burris, T. P. The coactivator LXXLL nuclear receptor recognition motif. *J Pept Res* **63**, 207–212 (2004).
32. Balzarotti, F. *et al.* Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science* aak9913 (2016).
33. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature methods* **5**, 507–516 (2008).
34. Carlson, H. A. & McCammon, J. A. Accommodating protein flexibility in computational drug design. *Molecular pharmacology* **57**, 213–218 (2000).
35. Halle, B. Biomolecular cryocrystallography: structural changes during flash-cooling. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4793–4798 (2004).
36. Jensen, M. Ø. *et al.* Mechanism of voltage gating in potassium channels. *Science* **336**, 229–233 (2012).
37. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. USA* **112**, 386–391 (2015).
38. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* **139**, 4780–4788 (2017).
39. Kuzmanic, A. *et al.* Changes in the free-energy landscape of p38α MAP kinase through its canonical activation and binding events as studied by enhanced molecular dynamics simulations. *eLife* **6**, e22175 (2017).
40. Pabon, N. A. & Camacho, C. J. Probing protein flexibility reveals a mechanism for selective promiscuity. *eLife* **6**, e22889 (2017).
41. Niu, Y. *et al.* Revealing inhibition difference between PFI-2 enantiomers against SETD7 by molecular dynamics simulations, binding free energy calculations and unbinding pathway analysis. *Scientific Reports* **7**, 46547 (2017).
42. Shan, Y. *et al.* How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* **133**, 9181–9183 (2011).
43. Shan, Y. *et al. Molecular basis for pseudokinase-dependent* autoinhibition of JAK2 tyrosine kinase. *Nature structural & molecular biology* **21**, 579–584 (2014).
44. Nettles, K. W. *et al.* NFkappaB selectivity of estrogen receptor ligands revealed by comparative crystallographic analyses. *Nat. Chem. Biol.* **4**, 241–247 (2008).
45. McInerney, E. M., Weis, K. E., Sun, J., Mosselman, S. & Katzenellenbogen, B. S. Transcription Activation by the Human Estrogen Receptor Subtypebeta (ERbeta) Studied with ERbeta and ERalpha Receptor Chimeras. *Endocrinology* **139**, 4513–4522 (1998).
46. Nettles, K. W. & Greene, G. L. Ligand control of coregulator recruitment to nuclear receptors. *Annu. Rev. Physiol.* **67**, 309–333 (2005).
47. Madak-Erdogan, Z. *et al.* Design of pathway-preferential estrogens that provide beneficial metabolic and vascular effects without stimulating reproductive tissues. *Science signaling* **9**, ra53 (2016).
48. Watson, C. S. & Gametchu, B. Membrane-Initiated Steroid Actions and theProteins that Mediate Them. *Proceedings of the Society for Experimental Biology and Medicine* **220**, 9–19 (1999).
49. Yang, L.-c. *et al.* Extranuclear estrogen receptors mediate the neuroprotective effects of estrogen in the rat hippocampus. *PloS one* **5**, e9851 (2010).
50. Norris, J. D. *et al.* Peptide antagonists of the human estrogen receptor. *Science* **285**, 744–746 (1999).
51. Tyulmenkov, V. V. & Klinge, C. M. Interaction of tetrahydrocrysene ketone with estrogen receptors α and β indicates conformational differences in the receptor subtypes. *Archives of biochemistry and biophysics* **381**, 135–142 (2000).
52. Hetenyi, C. & van der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *Febs Lett.* **580**, 1447–1450 (2006).
53. Hetenyi, C. & van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **11**, 1729–1737 (2002).
54. Hetenyi, C. & van der Spoel, D. Toward prediction of functional protein pockets using blind docking and pocket search algorithms. *Protein Sci.* **20**, 880–893 (2011).
55. Centrella, M., McCarthy, T. L., Chang, W. Z., Labaree, D. C. & Hochberg, R. B. Estren (4-estren-3 alpha,17 beta-diol) is a prohormone that regulates both androgenic and estrogenic transcriptional effects through the androgen receptor. *Mol Endocrinol* **18**, 1120–1130 (2004).
56. Ganesan, A., Coote, M. L. & Barakat, K. Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug Discov. Today* **22**, 249–269 (2016).
57. Copeland, R. A. The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov* (2015).
58. Levin, E. R. Integration of the extranuclear and nuclear actions of estrogen. *Mol Endocrinol* **19**, 1951–1959 (2005).
59. The PyMOL molecular graphics system v. 1.7.4 (New York, NY, 2014).
60. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).
61. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, 349–357 (2013).
62. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
63. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
64. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **78**, 1950–1958 (2010).
65. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101, https://doi.org/10.1063/1.2408420 (2007).
66. Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A. & Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

67. Parrinello, M. & Rahman, A. Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
68. Gasteiger, J. & Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. *Tetrahedron* **36**, 3219–3228 (1980).
69. Halgren, T. A. Merck molecular force field.1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
70. Stewart, J. J. Mopac2012. Stewart Computational Chemistry, Colorado Springs, CO (2012).
71. Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19**, 1–32 (2013).
72. Morris, G. M. *et al*. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
73. Wang, J. *et al*. Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *J. Phys. Chem. B* **116**, 7088–7101 (2012).

Acknowledgements

We acknowledge a grant of computer time from CSCS Swiss National Supercomputing Centre, and NIIF Hungarian National Information Infrastructure Development Institute. We acknowledge that the results of this research have been achieved using the DECI resource Archer based in the UK at the National Supercomputing Service with support from the PRACE aisbl. The work was supported by the K123836, K112807, K120391 grants from the National Research, Development, and Innovation Office, Hungary and KTIA_NAP_13-2014-0001. The University of Pécs is acknowledged for a grant PTE ÁOK_KA/2017 and also support in the frame of "Viral Pathogenesis" Talent Centre program. We are thankful to the Gedeon Richter Pharmaceutical Plc. for a pre-doctoral scholarship to N.J.

Author Contributions

M.B., N.J., and I.H. performed research. M.B., C.H., N.J., and I.M.A. designed research. C.H. and I.M.A. organized research. M.B., C.H., I.M.A., and N.J. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-14840-9.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

D4

Bálint *et al. J Cheminform (2017) 9:65*
https://doi.org/10.1186/s13321-017-0255-6

Journal of Cheminformatics

METHODOLOGY

Open Access

Systematic exploration of multiple drug binding sites

Mónika Bálint[1,2], Norbert Jeszenői[3], István Horváth[4], David van der Spoel[5] and Csaba Hetényi[1*]

Abstract

Background: Targets with multiple (prerequisite or allosteric) binding sites have an increasing importance in drug design. Experimental determination of atomic resolution structures of ligands weakly bound to multiple binding sites is often challenging. Blind docking has been widely used for fast mapping of the entire target surface for multiple binding sites. Reliability of blind docking is limited by approximations of hydration models, simplified handling of molecular flexibility, and imperfect search algorithms.

Results: To overcome such limitations, the present study introduces Wrap 'n' Shake (WnS), an atomic resolution method that systematically "wraps" the entire target into a monolayer of ligand molecules. Functional binding sites are extracted by a rapid molecular dynamics shaker. WnS is tested on biologically important systems such as mitogen-activated protein, tyrosine-protein kinases, key players of cellular signaling, and farnesyl pyrophosphate synthase, a target of antitumor agents.

Keywords: Peptide, Search, Pocket, Pharmacodynamics, Water, Interaction, Structure, Complex, Dissociation, Flexibility

Background

Molecular docking complements experimental structure determination and it has become a standard tool of drug discovery for the determination of protein–ligand complex structures [1]. The technique in practice is a compromise between computational cost and accuracy. Its high speed necessitates the use of severe approximations such as (i) restriction of the search space to the surroundings of the binding site, (ii) no or inadequate explicit hydration of the ligand-target interface, (iii) partial or complete neglect of target flexibility [2–5] during ligand binding, (iv) and non-deterministic search algorithms [1, 6] based on random number generation. Approximations i–iv seriously limit the applicability of docking methods for the following reasons. Restriction of the search to a primary binding site requires knowledge of its location and also neglects multiple sites such as allosteric ones [7, 8]. Water molecules often play a role in ligand binding [9–11] and ignoring interfacial water positions during docking may drive the ligands into pockets which are or should be filled with water molecules, resulting in incorrectly docked ligand poses [12]. Potential water release is also important during ligand binding especially through its entropic contributions [13, 14]. Neglecting or limiting the flexibility of target molecules is obviously incorrect at binding situations with induced fit [15]. Eventuality of random number generation in search engines such as Monte-Carlo or genetic algorithms [1, 5, 6] is a natural barrier of the reproducibility and reliability of the results.

The blind docking (BD) approach was introduced [16, 17] to extend the docking search to the entire target surface. In BD, previous knowledge and restriction of the search to a primary binding site are not necessary, and therefore, it can be used in search of multiple binding sites, as well. Indeed, BD has gained popularity [18–20] and has been used for finding allosteric [21–23], or multiple [24–28] binding sites. Thus, BD addresses the above first challenge and performs a global search instead of a focused one at an increased computational cost. However, approximations ii–iv cannot be remediated

*Correspondence: csabahete@yahoo.com
[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, Pécs 7624, Hungary
Full list of author information is available at the end of the article

as simply as the first one. Promising approaches using explicit water molecules in the binding pocket [10] (approximation ii) and treating target flexibility (approximation iii) have been reported for focused docking [29]. However, such approaches have not been implemented in conjunction with solving the global search problem of BD on the entire target surface. Statistical evaluation of multiple docking trials has been shown to increase reproducibility of a BD search [17]; by using multiple randomized (approximation iv) initial ligand positions. Thus, it has become common to perform several docking trials with different initial positions in a BD search to ensure that the largest possible part of the target surface is scanned. However, even such a statistical evaluation cannot guarantee systematic and reproducible exploration of the entire target surface during BD.

Molecular dynamics (MD) simulations have an increasing impact on drug development [30–32]. A series of pioneering studies have reported the use of MD for tracking the ligand binding process [33–37], at atomic resolution. MD calculations also allow the use of explicit water molecules and flexible targets overcoming the above limitations from approximations ii and iii [38–40] potentially opening a new avenue for improvement of BD. MD simulations typically use random starting conformations for the ligands, likewise to BD. Generally, long MD calculation times are required for successful navigation of the ligand into the binding site such that the computational time necessary for accurate docking of a ligand may be prohibitive in practice. Pocket search methods were also developed, exploiting the above-mentioned advantages of MD [41]. A recent review [30] also concludes that “Improper preparation of the initial structure or insufficient equilibration of the initial structure(s) can impact the quality of the MD results”. The present study is aimed at overcoming the above uncertainties of present fast BD and molecular dynamics techniques, by combination of their advantages into a new strategy. Test applications are presented with successful identification of multiple binding sites on biologically important systems such as MAP and tyrosine-protein kinases, key players of cellular signaling as well as farnesyl pyrophosphate synthase, a target of antitumor agents.

Algorithm

Wrap ‘n’ Shake (WnS) is a new method composed of consecutive algorithms, the Wrapper and the Shaker (Fig. 1, Additional file 1: Supporting Movie 1) offering a systematic search for multiple binding sites and modes. WnS works in synergy with popular open source program packages AutoDock 4.2.3 [29] and GROMACS 5.0.2 [42].

Wrapper

Wrapper performs several fast BD cycles by AutoDock 4.2, and AutoGrid 4.2 [29] and systematically covers the entire surface of the target with a monolayer of ligand copies (Fig. 1). Each BD cycle is performed as described in Additional file 2: Table S1, and results in 100 docked ligand copies, which are ordered by their interaction energies with the target, and structurally clustered. To achieve a ligand monolayer, the ligand–ligand interactions are minimized through implementation of a weak repulsion between the docked ligand copies, and therefore blocking the formation of ligand aggregates (Additional file 2: Table S2). At the same time, target-ligand interactions are maximized (Additional file 2: Table S3) to ensure that the largest possible numbers of new ligand copies are placed on the surface in an actual BD cycle. The initial experiments (Additional file 2: Table S2) also showed that introduction of a weak repulsion is essential to avoid erroneous ligand geometries clashing with target atoms. Such unwanted clashes (Additional file 2: Table S2) were obtained if intermolecular electrostatic ($E_{Coulomb}$) and van der Waals (E_{LJ}, Eq. 1) interaction energy terms were simply switched off at the ligand atoms. Notably, calculation of total target-ligand intermolecular interaction energy (E_{inter}) in AutoDock 4.2 is based on the scaled $E_{Coulomb}$ and E_{LJ} terms of the Amber96 force field [43], and an estimate for de-solvation free energy changes (ΔG_{sol}, Eq. 1). E_{LJ} is the sum of Lennard-Jones potential energy values (V, Fig. 2) calculated for all target-ligand atom pairs.

$$E_{inter} = E_{Coulomb} + E_{LJ} + \Delta G_{sol}. \tag{1}$$

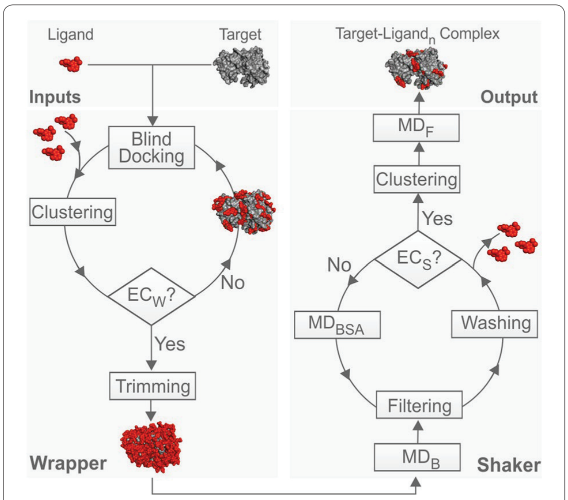


Fig. 1 Wrap ‘n’ Shake flowchart featuring the main steps of the method. A quick overview is also presented in Additional file 1: Supporting Movie 1

Finally, instead of the above-mentioned, oversimplified attempt of switching off all intermolecular terms of E_{inter} we elaborated a new protocol which produced the desired ligand monolayer by introduction of an excluded atom type (X). In this protocol, all ligand copies docked in a cycle and their surrounding target atoms are excluded from the next cycle (red in Fig. 2c), and only the unbound target surface (grey) is used for a next BD cycle. The neighboring target atoms are selected by an interface tolerance of 3.5 Å, the maximal distance between a target heavy atom and the closest docked ligand heavy atom. The above exclusion of certain atoms during docking is physically achieved by modification of the non-bonding terms of E_{inter}. For this, the new atom type X is assigned for excluded atoms (red in Fig. 2c) by a C program Wrp developed for this study. Wrp switches off $E_{Coulomb}$ by setting the partial charge of X to zero and also assigns new LJ parameters.

The new LJ parameters were fine-tuned for atom type X in order to produce the necessary weak repulsions described above. Briefly, the LJ parameters of X were calibrated considering the pairwise LJ potential between atom types X and Y (V_{XY}) at three common atom types (Y=O, C and H). A systematic search of both equilibrium potential well-depth (ε_X, Fig. 2a) and inter-nuclear distance (R_X) was conducted. Numerous docking runs were performed to evaluate the effect of the selected LJ parameters. A pre-defined value of r = 2 Å (ca. a covalent

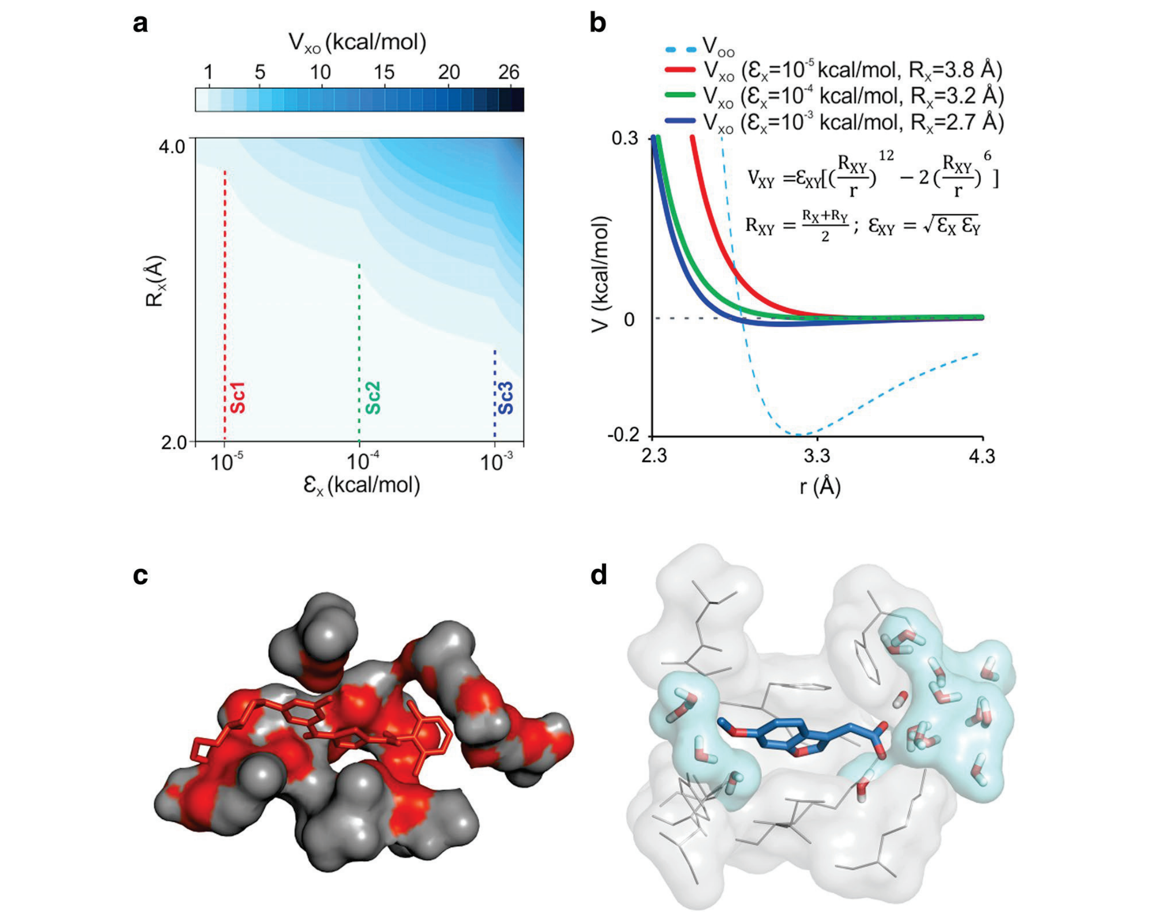


Fig. 2 Systematic calibration of ε_X and R_X. **a** A section of the $V_{XO}(r, \varepsilon_X, R_X)$ LJ potential function at r = 2 Å. Scenarios Sc1-Sc3 are shown with the magnitude of the R_X values corresponding to a short range repulsion of $V_{XO} \approx 1$ kcal/mol (dashed lines) **b** V_{XO} LJ potential functions of scenarios Sc1-Sc3. V_{OO} of an oxygen atom pair is also shown for comparison. **c** An example of excluded atoms X (red, Cycle 1, Rank 2, System 9). Docked ligand conformation, is presented with sticks and the binding pocket is shown as surface. **d** Ligand of System 2 (dark blue sticks) bound to its target farnesyl pyrophosphate synthase (grey lines and surface) after Shaker (MD_F step). Explicit water molecules surrounding the ligand within 7 Å are shown with sticks and light blue surface

bond + 0.5 Å) was used as a minimal distance where short-range repulsion should act at a desired maximal value not exceeding a $V_{XY} \approx 1$ kcal/mol. Three scenarios (Sc1-Sc3) were evaluated as shown in the r = 2 Å section of V_{XO} (r, ε_X, R_X) function (Fig. 2a) calculated for the XO atom type pair. Sc2 (green line, Fig. 2a, b) was identified as an optimal scenario with an $\varepsilon_X = 10^{-4}$ kcal/mol and an R_X of 3.2 Å (approximate distance between heavy atoms in an H-bond). In this case, available target surface is optimally used without large ligand-free zones in the monolayer. A short-range repulsion was achieved (green line in Fig. 2b) with a zero value beyond the repulsion zone. If R_X was too large (Sc1, red in Fig. 2a, b) then the repulsion zone around the docked ligand copies would also increase with a V_{XO} curve shifted to the right if compared to the green curve of Sc2 (Fig. 2b) resulting in large ligand-free zones, i.e. a non-optimal arrangement of the ligand copies in the monolayer. Importantly, the repulsion zone in the optimal V_{XO} curve of Sc2 starts at lower distances (r) than in the V_{OO} curve. V_{OO} is shifted to the right of the red curve (Sc1), which would result in even larger ligand-free regions than Sc1. Thus, using only a repulsion term of V_{OO} would have not been adequate for exclusions of atoms in wrapping. On the other hand, if R_X was too small (Sc3, blue in Fig. 2a, b), then unwanted attractive effects such as aggregation between docked ligand copies would still happen similar to Trial 1, in Additional file 2: Table S2. Accordingly, in Sc3 the corresponding blue curve is shifted to the left from the green Sc2 curve (Fig. 2b). The same procedure was repeated for atom types Y = C and H and an average R_X value of 3.6 Å was concluded (Additional file 2: Table S3) and used in Wrapper along with the above $\varepsilon_X = 10^{-4}$ kcal/mol.

These calibrated LJ parameters of X allowed elimination of the above-mentioned unwanted interactions between the newly docked ligand copies and the already filled binding pockets (Fig. 2c). As the introduced repulsive potential acts on a short range, the ligands can still dock to other, unbound parts of the target surface. The new atom type and parameters also maximize target-ligand interactions adding the maximal number of ligand copies to the mono-layer during a BD cycle.

Wrapper cycles are terminated by either the drop of uncovered surface area of the target below one percent of its total (ligand-free, initial) surface area, or positive target-ligand interaction energy in every cluster representative (EC_W in Fig. 1). As a last step, a trimming is performed to remove all ligand copies situated more than 3.5 Å from the target. Wrapper results in a target wrapped in N ligand copies (target-ligand$_N$ complex) provided as a single Protein Databank (PDB) file. Wrapper is implemented in a new open source package WnS as shell scripts and a C program Wrp available for download together with a User's Manual at www.wnsdock.xyz.

Shaker

Shaker selects functional binding sites by removing nonspecific, loosely bound ligand copies from the target surface. The target-ligand$_N$ complex is placed in a box filled with water and subjected to MD simulations in consecutive cycles. The cycles are performed until a 75% of the ligand copies are eliminated (Exit Criterion of Shaker, EC_S Fig. 1). In each Shaker cycle, distance and energy metrics are calculated describing target-ligand interactions at each time step (frame) of a trajectory. The metrics include the closest distances between the target and the ligand as well as E_{LJ}, calculated using Amber parameters. Based on these metrics, filtering (Additional file 2: Table S4) and subsequent removal of the corresponding ligand co-ordinates (Washing, Fig. 1) are applied to exclude ligand positions dissociated from their starting binding positions. The filtering involves two distance-based steps and two final steps based on E_{LJ}.

Before the first cycle a 5-ns target backbone-restrained MD (MD_B) is used to grossly shake off the weakly bound ligands. In cases where this initial MD is not enough to reach the required EC_S (Additional file 2: Table S1 and Additional file 2: Table S7), multiple cycles with 20-ns simulated annealing (MD_{BSA}) simulations are performed, using position restraints on the target backbone atoms. Depending on the molecular weight (MW, Table 1) of the ligands, SA was done, using two temperature protocols, up to 50 °C (MW ≤ 300) or 80 °C (MW ≥ 300). High temperature in SA accelerated the dissociation process as expected. After MD_{BSA} cycles, a clustering and ranking step is performed, using the last frames of the remaining ligands. A refinement of 20-ns MD with full protein flexibility (MD_F) is also performed on every target-ligand complex resulted after clustering (Additional file 2: Table S7 and Additional file 2: Table S8). The Shaker protocol (Additional file 2: Table S9) was formulated during multiple trials (Additional file 2: Tables S5 and S6) and results in a final solution structure of a target-ligand$_n$ complex, where n is the total number of final cluster representatives.

Systems and test metrics

A diverse set of ten target-ligand systems were selected (Table 1) and prepared (Additional file 2: Table S1) as test cases of WnS. Challenging systems with multiple (prerequisite or allosteric) binding sites were included (Table 1). Our selection contains both small ligands and bulky, flexible ones. Apo protein structures were used as targets except System 8. In the case of System 5 another

Table 1 Test systems

#	PDB ID[a]	Target	Ligand	MW[b]
1	3ptb	bovine β-trypsin	benzamidine	120
2	3n3 l	farnesyl pyrophosphate synthase	(6-methoxy-1-benzofuran-3-yl) acetic acid (MS0)	206
3a	3hvc	mitogen-activated protein kinase	4-[3-(4-fluorophenyl)-1 h-pyrazol-4-yl]pyridine (GG5)	239
3b	4f9w	mitogen-activated protein kinase	4-[3-(4-fluorophenyl)-1 h-pyrazol-4-yl]pyridine (GG5)	239
4	3cpa	carboxy-peptidase	GY	256
5	1qcf	haematopoetic cell kinase (HCK)	1-ter-butyl-3-p-tolyl-1 h-pyrazolo[3,4-d]pyrimidin- 4-ylamine (PP1)	281
6	1h61	pentaerythritol tetranitrate reductase	Prednisone®	358
7	2bal	mitogen-activated protein kinase	[5-amino-1-(4- Fluorophenyl)-1H-Pyrazol-4- yl] [3-(piperidin-4-yloxy) phenyl]methanone	380
8	1hvy	thymidylate synthase	Ralitrexed®	459
9	3g5d	tyrosine-protein kinase Src	Dasatinib®	488
10	1be9	PDZ-domain	KQTSV	544

[a] PDB ID of the holo X-ray structure

[b] Molecular weight of the ligand

protein tyrosine-protein kinase was used as apo structure similar to a previous study [33].

Three standard metrics were used to quantify the results of tests. (1) root mean squared deviation (RMSD) measures structural precision of WnS results by comparison of atomic positions of ligand conformations produced by WnS and those of the crystallographic reference. Prior to calculation of RMSD, a structural alignment (Additional file 2: Table S10) was performed on the holo and apo target residues surrounding the ligand within 5 Å similarly to a previous work [33]. (2) Shaker Rate (SR = N/n) is a ratio of counts of the N ligand copies residing on the target surface (N) after Wrapper and the n final cluster representatives (n) produced by Shaker. The larger the SR, the more efficiently Shaker eliminated ligand copies from the target surface. (3) Rank serial number (#Rank) is calculated using relative ligand-target interaction energies corresponding to the docked ligand positions. WnS ranks docked ligand copies by their interaction energies with the target. The smaller the #Rank, the stronger the target-ligand interaction is at a ligand position. The #Rank of the docked ligand copy of the lowest RMSD is listed for all systems in Table 2. In the final rank lists, docked ligand copies with small RMSD, i.e. close to the crystallographic conformations should be preferably placed at the top of the rank lists, with small #Rank values.

Results and discussion

Association or dissociation?

Encouraged by results of pioneering MD studies [31, 33, 34], association of ligand benzamidine to bovine trypsin was followed in three MD simulations. Benzamidine is an easy case for docking and it has also been used in tests of recent approaches [44]. The present MD simulations were 1-µs-long and benzamidine was placed at three different starting positions (Fig. 3, Additional file 2: Table S11), at various distances (Fig. 3a) from the crystallographic binding site.

Analysis of the trajectories shows that the crystallographic binding position was found in two out of the three simulations after 81 and 690 ns simulation time (drop of red and green lines in Fig. 3b), respectively. In the 3rd case with the largest starting distance, 1 µs was not enough to dock to the native site by association (blue line). Thus, the usefulness of association MD runs for docking strongly depends on the starting ligand position even in the easy case of benzamidine. MD needs a simulation time comparable to the real association time of the ligand (Fig. 3b). This can be considerable, as migration of the ligand is hindered by friction in the surrounding water. Previous studies [33, 36, 45], have also reported simulations of several hundreds of nanoseconds for navigation of the ligand to the desired binding pocket.

All-in-all, the necessary time for successful docking by association MD depends on the actual starting position of the ligand, the size and shape of the target, ligand etc. To overcome such uncertainties on simulation length and still use the benefits of MD we elaborated a new strategy, the Wrap 'n' Shake (WnS, Fig. 1). Instead of simulating the association process, WnS is based on the dissociation of the ligand. Dissociation is fast and reproducible at binding sites of low stability.

A systematic approach

Naturally, a dissociation approach requires a set of ligand copies bound to the target. Systematic mapping of all possible ligand positions (sites) cannot be guaranteed in a single BD cycle (Introduction) even if it contains hundreds of fast BD trials [17]. A truly systematic algorithm

Table 2 Results for the test systems

#	N[a]	CLS[b]	#Rank[c]		n[d]	SR[e]
			MD_{BSA}	MD_F		
1a	68	6	1	1	6	11
1b[g]	74	5	1	–	4	19
1c[g]	71	6	1	–	5	14
2	300	18	2	4	13	23
3a	222	46	3	4	21	11
3b	222	46	9	12	21	11
4 [h]	155	12	1	1	8	19
5	143	25	2	1	12	12
6[i]	116	26	1	2	12	10
7	123	26	4	4	12	10
8	106	25	1	1	10	11
9[j]	92	23	2	1	10	9
10	49	11	2	1	4	12

[a] Total count of ligand copies after Wrapper

[b] Count of ligands surviving the Shaker, after MD_{BSA}

[c] Rank serial number of the structure with the best RMSD value, after MD_{BSA} and after MD_F

[d] Count of cluster representatives (final solutions) Shaker

[e] Shaker Rate

[f] Total computational time required for MD_B, MD_{BSA} and MD_F, as explained in Additional file 2: Table S12

[g] For System 1, WnS was performed with different seeds for data reproduction purposes

[h] Final clustering was done using van der Waals and Coulomb interactions due to interactions of zinc ion with the ligand

[i] Wrapper process was done, using the LJ interaction as a scoring function, instead of AD4 (Additional file 2: Table S13)

[j] Final clustering was done with 6 Å distance limit between clusters

should completely wrap the entire surface of the target in a monolayer of copies of the ligand molecule. Our initial guess of such a Wrapper algorithm was based on a previous finding [17] that the coverage of the target can be increased with several, successive fast BD cycles where accumulated docked ligand copies from the previous cycle are considered as part of the target in the next cycle. However, additional experiments with such successive BD cycles showed that previously and newly docked ligand copies can easily form multi-layer aggregates with each-other instead of the target (Additional file 2: Table S2). The formation of such aggregates hinders wrapping of the target surface into the desired monolayer of ligand copies.

During the wrapping process, parts of the target surface already covered with ligand copies has to be excluded from interactions with ligand copies docked in a next BD cycle. This task is not trivial as potential functions of the docking force fields normally cannot distinguish between target sites unbound and covered with ligands. After extensive experimentation including an optimization of the force field ("Wrapper" section, Additional file 2: Table S3, Appendix 1) we arrived at a new algorithm called Wrapper (Figs. 2, 4). Wrapper performs a systematic coverage of the target surface in several, consecutive fast blind docking cycles (Fig. 4). The algorithm continuously monitors the status of coverage of target surface (Fig. 4a) and results in the desired monolayer of N ligand copies not interacting with each-other. Figure 4b shows an example of such a monolayer. Ligands constituting the monolayer have physically realistic arrangement (Fig. 4c), maximized interactions with the target and no contacts with each-other. Thus, the target is systematically and rapidly wrapped in a monolayer of N (Table 2) ligands.

Having a realistic input geometry, the resulting target-$ligand_N$ complex is transferred to the Shaker including MD simulation(s) with explicit water ("Shaker" section), filtering, and clustering steps. These steps eliminate ligands dissociated during MD and result in a strong binder at each pocket (Additional file 2: Table S7). Final results are shown in Table 2 using test metrics described in "Systems and test metrics". Parameter SR characterizes efficiency of removal of loose binders. SR values of Table 2 indicate that a considerably large part of the weak binders were efficiently removed at all test systems beyond the default EC_S of 75% (SR = 4). Other important metrics are RMSD and #Rank. In most of the systems analyzed, ligand conformations with the lowest RMSD

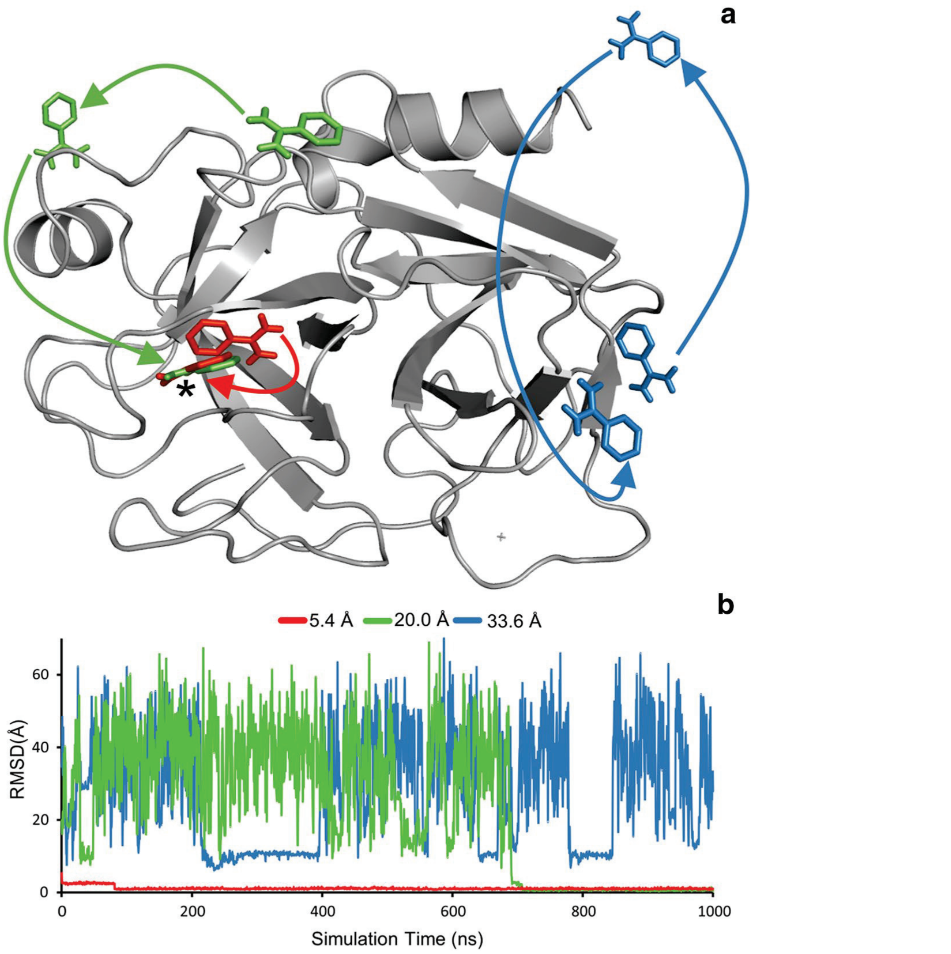


Fig. 3 Pilot molecular dynamics simulations. Benzamidine ligand (sticks) started the MD simulations from three positions at different distances (as indicated in the legend) from the native binding site on the trypsin target (grey cartoon). Arrows in **a** point from starting (t = 0 ns) to final (t = 1000 ns) ligand positions. Only two of the three 1000 ns-long simulations with the closest starting position succeeded in finding the reference binding pose (*) known from the crystallographic structure (3ptb). **b** Time-dependence of root mean squared deviation (RMSD) of the ligand measured from its reference pose

were placed into the first two ranks (Table 2, Fig. 5, and Additional file 2: Table S8). For stable ligand copies, good structural matches to the corresponding reference conformations (Fig. 5 and Additional file 2: Table S8), as well as low #Rank values (Table 2) were found. Fair results were obtained for challenging cases too (Systems 2 and 3). The somewhat lower rank in these cases may be explained by the relatively high B-factor of the ligands of these systems (Additional file 2: Table S1) suggesting an increased mobility and a less stable target-ligand interaction.

For example, B-factors of measured atomic positions of ligand MSo (System 2) vary in a range between 54 and 95 $Å^2$ (Additional file 2: Table S1). During MD_F simulations we found that the RMSD varied between 2.5 and 5.1 Å (Additional file 2: Table S8), and a final #Rank of 4 and an RMSD of 3.1 Å were obtained. Considering the above high B-factor values, it is realistic to assume that ligand MSo adopts various conformations when bound to farnesyl phosphate synthase (System 2) including the one close to the assigned position found with an RMSD of 2.5 Å. This conformational variability of the bound

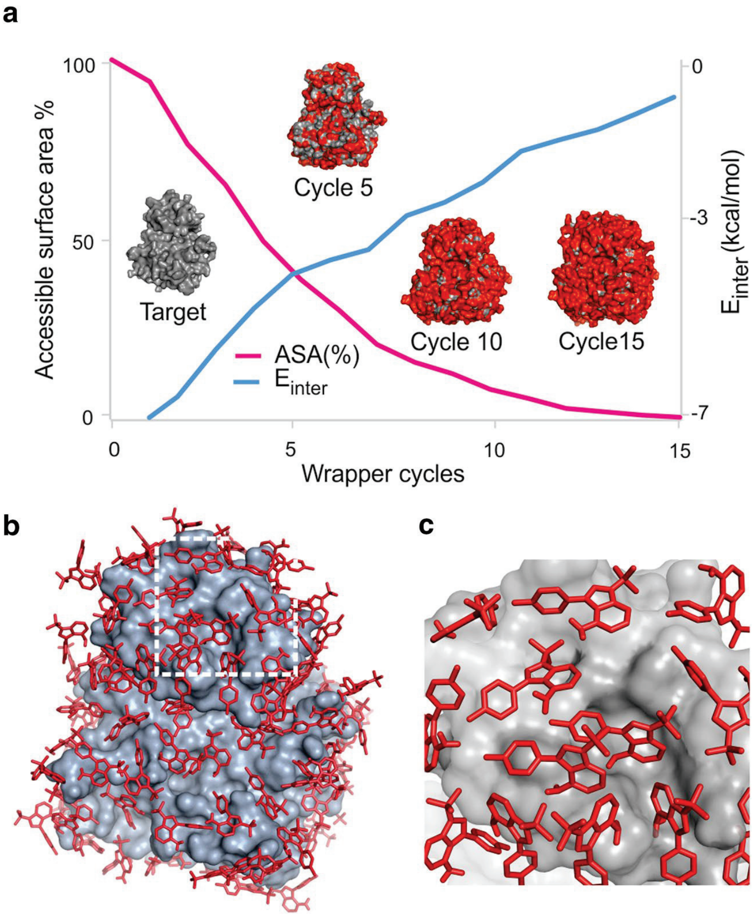


Fig. 4 Wrapping tyrosine-protein kinase Src target into a mono-layer of ligand copies (System 5). **a** Unbound (ligand free) accessible surface area (ASA) of the target and the lowest E_{inter} of the cluster representatives in consecutive wrapping cycles. Target-ligand interaction energy (E_{inter}) increases with increasing number of cycles finding strong binding sites in the first few cycles, before the final, saturation region. ASA finally decreases below 1%. Structural images show the wrapping of the target (grey surface) with ligands (red). **b** The monolayer arrangement of the ligands (red sticks) wrapping the entire target surface (grey) after the final cycles. **c** A close-up of a section of the monolayer showing that the ligand copies are evenly arranged without overlap

MSo is probably due to its carboxylate group with the highest B-factor of 95 Å^2. This group is hydrated by bulk water molecules, helping the dissociation of MSo from the target (Fig. 2d). At the same time, MD simulations with explicit water molecules also account for a hydrophobic, anchoring interaction between the benzofuran part of MSo (no waters present, Fig. 2d) and the target. This example shows the necessity of use of explicit water model during the shaking process in order to account for all, even antagonistic interactions.

In our pilot study ("Association or dissociation?" section) it was demonstrated that MD methods following the association pathways often need large amount of computational time and/or a fortunate starting conformation in order to find the primary site correctly for System 1. WnS yielded the correct solution for this system (Additional file 2: Table S8) in a 5-ns-long MD_B simulation which is at least one order of magnitude shorter than the lengthy association times discussed in "Association or dissociation?" section. Elimination of ligand excess (dissociation of ligand copies) (Tables S14 and S15) at an SR of 11 was facilitated by hydrogen bonding with explicit water molecules [46, 47]. Thermal motion of water molecules also contributed to fast "shake off" of the ligand copies especially in the cases of Systems with small ligands with the application of the simulated annealing

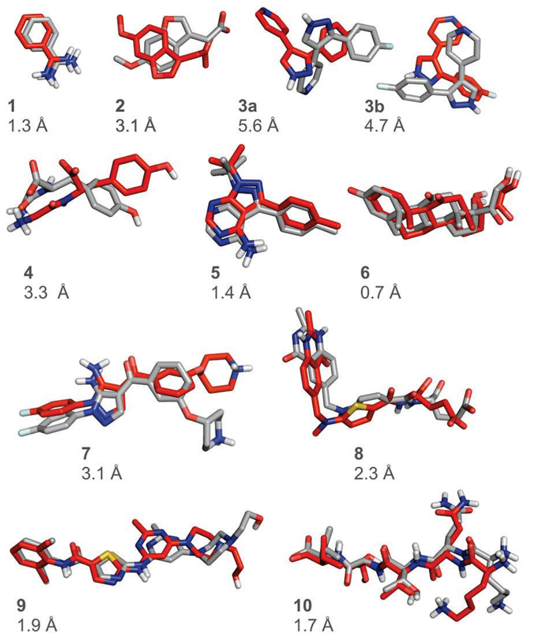


Fig. 5 Structural fits quantified as root mean squared deviation (RMSD) with values given in Å. Ligand conformations after Shaker (grey) compared to the crystallographic references (red sticks). System# is bold

protocol (MD_{BSA}, see an SR of 23 in case of System 2 in Table 2).

A case with a small ligand

WnS was tested on tyrosine protein kinase target with a pyrazolopyrimidine 1 ligand (PP1, System 5). Regulation of kinase activity is important in numerous human diseases [48, 49]. At the same time, this kinase is a challenging test target for WnS as it has multiple sites including an allosteric one identified in previous studies [50, 51]. The native, PP1 site was found (Fig. 5) at an excellent RMSD agreement (1.4 Å, Fig. 5) with the crystallographic position. Besides obtaining very good RMSD (Fig. 5), the #Rank was improved from second to first place (Table 2) during the final MD_F simulation (Additional file 2: Table S16). Apart from the primary site, our goal was to find other, prerequisite binding sites, as well. As described in a previous MD study [33], such sites correspond to poses on the binding pathway leading to the primary site. WnS found both low- and high energy prerequisite sites described previously [33] (Fig. 6). Besides structural matches, #Rank and the corresponding energy values are also comparable to the previous results. Notably, WnS can predict multiple binding sites beyond experimentally observable ones. These binding sites can be considered as prerequisite or allosteric binding sites. Previous MD results [33, 52] concluded, that finding prerequisite binding sites is a substantial advantage of the MD simulations.

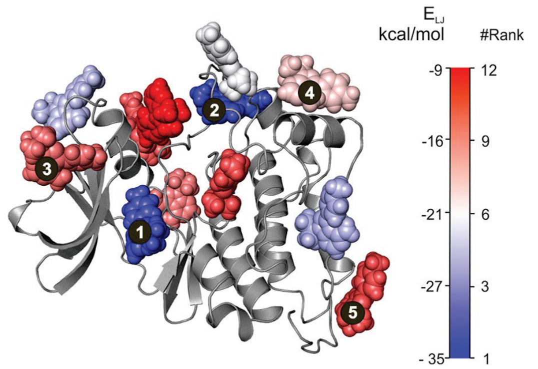


Fig. 6 Haematopoetic cell kinase (HCK, System 5) with ligand copies remaining after Shaker. Ligand copies are colored by the calculated target-ligand interaction energy E, and the #Rank assigned. The previously reported pockets 1(ATP), 2(A-loop), 3(PIF site), 4(G-loop) and 5(MYR) are numbered by their increasing E_{LJ}

Cases with large ligands

Tyrosine kinase also binds dasatinib (System 9), a bulky ligand, for which an SR of 9 was obtained (Table 2), after six simulated annealing cycles (Additional file 2: Table S12). The same four binding pockets were found for dasatinib as for the above PP1 (Additional file 2: Table S17). After the final MD_F step, local conformational refinement of dasatinib was observed, improving the RMSD from 2.3 to 1.9 Å. Similar to PP1, this could be partially explained by the role of the water molecules and the enhanced target motion during MD_{BSA}. WnS was further tested on the challenging System 10 with a pentapeptide ligand with twenty-three flexible torsions. The correct binding position of the ligand was obtained after the MD_F stage of Shaker with an improvement of RMSD from 6.8 to 1.7 Å (Fig. 7, Additional file 3: Supporting Movie 2).

A re-ranking (Table 2) from Rank 2 to Rank 1 was also observed after MD_F. For comparison, the wrapped target-ligand$_N$ complex of System 10 was subjected directly to an MD_F simulation skipping the MD_B and MD_{BSA} steps of Shaker. In this case, an RMSD of 11.3 Å (Line 10b in Additional file 2: Table S8) was obtained which was worse than the RMSD obtained with the complete Shaker protocol (1.7 Å, Fig. 5). This demonstrates that both MD_B and MD_{BSA} steps of Shaker are necessary to find the correct position. After Wrapper, the pentapeptide was in a closed, cyclic conformation (Fig. 7, Snapshot 1). This unrealistic arrangement was opened up (Snapshots 2 and

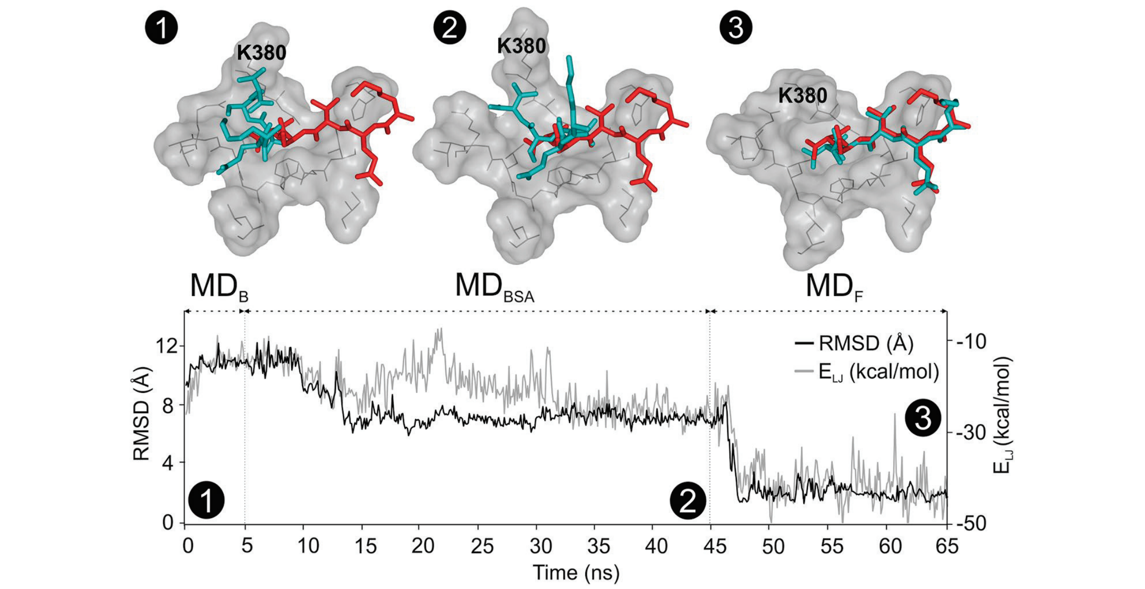


Fig. 7 During Shaker, conformational changes of the pentapeptide KQTSV are observed, while remains bound to its pocket on the PDZ domain (System 10). Red sticks represent the native ligand conformation from PDB (1be9). Teal sticks represent ligand conformations at different Shaker stages starting with the conformation right after Wrapper (**1**), and continuing with conformation after MD$_{BSA}$ (**2**), and after MD$_F$ (**3**). The changes of target-ligand interaction energy (E_{LJ}) and the RMSD during the MD stages in the Shaker protocol are plotted below the structural snapshots. See also Additional file 3: Supporting Movie 2 for further details of conformational changes

3) by interacting water molecules. It can be also observed that limited protein flexibility during MD$_B$ and MD$_{BSA}$ allowed only moderate reduction of the ligand RMSD by improvement of the target-ligand interactions. Most of the RMSD and interaction energy improvement was achieved after MD$_F$, and rearrangement of K380 inside the pocket was necessary, to improve the conformation of the simulated ligand (Fig. 7). All-in-all, MD steps including target flexibility have a significant influence on the results of WnS for large ligands. Introduction of MD$_F$ considerably improved structural precision, in the above case studies of large ligands (Systems 9 and 10).

Conclusions

In the present study, a systematic strategy, the Wrap 'n' Shake was introduced for exploration of multiple binding sites and modes of drugs on their macromolecular targets. Wrap 'n' Shake systematically wraps the target into a monolayer of ligand copies using a modified blind docking approach and selects stable positions by shaking off loose binders. The method offers a computationally feasible solution for the present problems of the field (Introduction). Wrapper requires only fast blind docking cycles with a program package such as AutoDock 4.2.3. The Shaker process is fairly short and can be performed by available MD packages. Shaker is further accelerated by simulated annealing and uses all benefits of explicit water model and target flexibility. Wrap 'n' Shake is suitable to study interactions of protein targets with even large peptide ligands. We have started the extension of the method towards protein ligands using a fragment-based approach with post hoc reconstruction of the ligand. In future applications, Wrap 'n' Shake could be also used for general pocket search, besides docking of individual ligands. We envision that Wrap 'n' Shake can become the tool of choice for systematic exploration of multiple binding sites and modes of ligands in drug design and structural biology.

Additional files

Additional file 1. Supporting Movie 1 featuring the processes of Wrapper and Shaker in the case of System 5. The first part presents the results of 15 wrapping cycles. The second part contains MD$_B$ and two MD$_{BSA}$ cycles of Shaker. Final cluster representatives are the outputs of WnS. Additional refinement steps are shown in Supporting Movie 2 (Additional file 3).

Additional file 2. Supporting Tables S1–S17 and Appendix 1–4 with detailed methods and results.

Additional file 3. Supporting Movie 2 featuring conformational changes of pentapeptide KQTSV, bound to PDZ-domain (System 10) during 65 ns simulations performed Shaker. The binding pocket of KQTSV on the PDZ domain is presented with grey surface. The simulated and crystallographic reference structures of KQTSV are presented as teal and red sticks.

Abbreviations

BD: blind docking; EC_W: exit criterion of Wrapper; EC_S: exit criterion of Shaker; E_{inter}: intermolecular interaction energy; E_{LJ}: Lennard-Jones potential; MD_B: molecular dynamics with backbone position restraints; MD_{BSA}: molecular dynamics with backbone position restraints and simulated annealing; MD_F: molecular dynamics without restraints (flexible simulation; PDB: protein data bank; RMSD: root mean squared deviation; SR: Shaker Rate; WnS: Wrap 'n' Shake.

Authors' contributions

MB, CH, NJ, and IH performed research. MB, CH, and DvdS designed research, and wrote the manuscript. All authors formulated the research. All authors read and approved the final manuscript.

Author details

[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, Pécs 7624, Hungary. [2] Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest 1117, Hungary. [3] MTA NAP-B Molecular Neuroendocrinology Group, Institute of Physiology, Szentágothai Research Center, Center for Neuroscience, University of Pécs, Szigeti út 12, Pecs 7624, Hungary. [4] Chemistry Doctoral School, University of Szeged, Dugonics tér 13, Szeged 6720, Hungary. [5] Uppsala Center for Computational Chemistry, Science for Life Laboratory, Department of Cell and Molecular Biology, University of Uppsala, Box 596, 75124 Uppsala, Sweden.

Acknowledgements

We acknowledge a grant of computer time from CSCS Swiss National Supercomputing Centre, and NIIF Hungarian National Information Infrastructure Development Institute. We acknowledge that the results of this research have been achieved using the DECI resource Archer based in the UK at the National Supercomputing Service with support from the PRACE aisbl.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

A software package is released under the GNU GPL, freely accessible with examples and a manual at http://www.wnsdock.xyz.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

The work was supported by the K123836, K112807, K120391 grants from the National Research, Development, and Innovation Office, Hungary. The University of Pécs is acknowledged for a grant PTE ÁOK_KA/2017 and also support in the frame of "Viral Pathogenesis" Talent Centre program. We are thankful to the Gedeon Richter Pharmaceutical Plc. for a pre-doctoral scholarship to N.J.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 2 August 2017 Accepted: 16 December 2017
Published online: 28 December 2017

References

1. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935–949
2. Fischer M, Coleman RG, Fraser JS, Shoichet BK (2014) Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. Nat Chem 6:575–583
3. Hou XB, Li KS, Yu X, Sun JP, Fang H (2015) Protein flexibility in docking-based virtual screening: discovery of novel lymphoid-specific tyrosine phosphatase inhibitors using multiple crystal structures. J Chem Inf Modeling 55:1973–1983
4. Pan AC, Borhani DW, Dror RO, Shaw DE (2013) Molecular determinants of drug–receptor binding kinetics. Drug Discov Today 18:667–673
5. Halperin I, Ma BY, Wolfson H, Nussinov R (2007) Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins Struct Funct Genet 47:409–443
6. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373
7. Iorga B, Herlem D, Barre E, Guillou C (2006) Acetylcholine nicotinic receptors: finding the putative binding site of allosteric modulators using the "blind docking" approach. J Mol Modeling 12:366–372
8. Othman R, Kiat TS, Khalid N, Yusof R, Newhouse EI, Newhouse JS et al (2008) Docking of noncompetitive inhibitors into dengue virus type 2 protease: understanding the interactions with allosteric binding sites. J Chem Inf Modeling 48:1582–1591
9. Mancera RL (2007) Molecular modeling of hydration in drug design. Curr Opin Drug Discov Dev 10:275–280
10. Jeszenoi N, Bálint M, Horváth I, Van Der Spoel D, Hetényi C (2016) Exploration of interfacial hydration networks of target–ligand complexes. J Chem Inf Modeling 56:148–158
11. Jeszenoi N, Horvath I, Balint M, van der Spoel D, Hetenyi C (2015) Mobility-based prediction of hydration structures of protein surfaces. Bioinformatics 31:1959–1965
12. Hetenyi C, van der Spoel D (2011) Toward prediction of functional protein pockets using blind docking and pocket search algorithms. Protein Sci 20:880–893
13. Ahmad M, Helms V, Lengauer T, Kalinina OV (2015) Enthalpy–entropy compensation upon molecular conformational changes. J Chem Theory Comput 11:1410–1418
14. Ahmad M, Kalinina O, Lengauer T (2014) Entropy gain due to water release upon ligand binding. J Cheminform 6(1):P35
15. Rastelli G, Pacchioni S, Sirawaraporn W, Sirawaraporn R, Parenti MD, Ferrari AM (2003) Docking and database screening reveal new classes of plasmodium falciparum dihydrofolate reductase inhibitors. J Med Chem 46:2834–2845
16. Hetenyi C, van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11:1729–1737
17. Hetenyi C, van der Spoel D (2006) Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett 580:1447–1450
18. Grinter SZ, Zou X (2014) Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. Molecules 19:10150–10176
19. Yuriev E, Holien J, Ramsland PA (2015) Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. J Mol Recognit 28:581–604
20. Yuriev E, Ramsland PA (2013) Latest developments in molecular docking: 2010–2011 in review. J Mol Recognit 26:215–239
21. Hocker HJ, Rambahal N, Gorfe AA (2014) LIBSA—a method for the determination of ligand-binding preference to allosteric sites on receptor ensembles. J Chem Inf Model 54:530–538
22. Schindler CE, de Vries SJ, Zacharias M (2015) Fully blind peptide-protein docking with pepATTRACT. Structure 23:1507–1515
23. Whalen KL, Tussey KB, Blanke SR, Spies MA (2011) Nature of allosteric inhibition in glutamate racemase: discovery and characterization of a cryptic inhibitory pocket using atomistic MD simulations and pK(a) calculations. J Phys Chem B. 115:3416–3424
24. Garcia-Sosa AT, Sild S, Maran U (2008) Design of multi-binding-site inhibitors, ligand efficiency, and consensus screening of avian influenza H5N1 wild-type neuraminidase and of the oseltamivir-resistant H274Y variant. J Chem Inf Modeling 48:2074–2080
25. Roumenina L, Bureeva S, Kantardjiev A, Karlinsky D, Andia-Pravdivy JE, Sim R et al (2007) Complement C1q-target proteins recognition is inhibited by electric moment effectors. J Mol Recognit 20:405–415
26. Bugatti A, Giagulli C, Urbinati C, Caccuri F, Chiodelli P, Oreste P et al (2013) Molecular interaction studies of HIV-1 matrix protein p17 and heparin:

identification of the heparin-binding motif of p17 as a target for the development of multitarget antagonists. J Biol Chem 288:1150–1161

27. Kovacs M, Toth J, Hetenyi C, Malnasi-Csizmadia A, Sellers JR (2004) Mechanism of blebbistatin inhibition of myosin II. J Biol Chem 279:35557–35563
28. Agarwal T, Annamalai N, Khursheed A, Maiti TK, Bin Arsad H, Siddiqui MH (2015) Molecular docking and dynamic simulation evaluation of Rohinitib—Cantharidin based novel HSF1 inhibitors for cancer therapy. J Mol Graph Modelling 61:141–149
29. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
30. Ganesan A, Coote ML, Barakat K (2017) Molecular dynamics-driven drug discovery: leaping forward with confidence. Drug Discov Today 22:249–269
31. Dror RO, Dirks RM, Grossman J, Xu H, Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41:429–452
32. Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. BMC Biol 9:71
33. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE (2011) How does a drug molecule find its target binding site? J Am Chem Soc 133:9181–9183
34. Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci USA 108:10184–10189
35. Limongelli V, Bonomi M, Parrinello M (2013) Funnel metadynamics as accurate binding free-energy method. Proc Natl Acad Sci USA 110:6358–6363
36. Shan Y, Gnanasambandan K, Ungureanu D, Kim ET, Hammaren H, Yamashita K et al (2014) Molecular basis for pseudokinase-dependent autoinhibition of JAK2 tyrosine kinase. Nat Struct Mol Biol 21:579–584
37. Jensen MØ, Jogini V, Borhani DW, Leffler AE, Dror RO, Shaw DE (2012) Mechanism of voltage gating in potassium channels. Science 336(6078):229–233
38. Borhani DW, Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26:15–26
39. Casasnovas R, Limongelli V, Tiwary P, Carloni P, Parrinello M (2017) Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations. J Am Chem Soc 139:1480–4788
40. Kuzmanic A, Sutto L, Saladino G, Nebreda AR, Gervasio FL, Orozco M (2017) Changes in the free-energy landscape of p38α MAP kinase through its canonical activation and binding events as studied by enhanced molecular dynamics simulations. eLife 6:e22175
41. Prakash P, Hancock JF, Gorfe AA (2015) Binding hotspots on K-ras: consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. Proteins Struct Funct Bioinform 83:898–909
42. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B et al (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1:19–25
43. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM et al (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 117:5179–5197
44. Soderhjelm P, Tribello GA, Parrinello M (2012) Locating binding poses in protein-ligand systems using reconnaissance metadynamics. Proc Natl Acad Sci USA 109:5170–5175
45. Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan YB et al (2012) Pathway and mechanism of drug binding to G-protein-coupled receptors. Biophys J 102:410
46. van der Spoel D, van Maaren PJ, Larsson P, Timneanu N (2006) Thermodynamics of hydrogen bonding in hydrophilic and hydrophobic media. J Phys Chem B 09(110):4393–4398
47. Schmidtke P, Luque FJ, Murray JB, Barril X (2011) Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design. J Am Chem Soc 133:18903–18910
48. Cohen P (2002) Protein kinases—the major drug targets of the twenty-first century? Nat Rev Drug Discov 1:309–315
49. Shukla D, Meng Y, Roux B, Pande VS (2014) Activation pathway of Src kinase reveals intermediate states as targets for drug design. Nat Commun. 5:3397
50. Foda ZH, Seeliger MA (2014) Kinase inhibitors: an allosteric add-on. Nat Chem Biol 10:796–797
51. Sadowsky JD, Burlingame MA, Wolan DW, McClendon CL, Jacobson MP, Wells JA (2011) Turning a protein kinase on or off from a single allosteric site via disulfide trapping. Proc Natl Acad Sci USA 108:6056–6061
52. Tiwary P, Limongelli V, Salvalaglio M, Parrinello M (2015) Kinetics of protein-ligand unbinding: predicting pathways, rates, and rate-limiting steps. Proc Natl Acad Sci USA 112:E386–E391

hetenyi.csaba_83_23

Chapter 8

Systematic Exploration of Binding Modes of Ligands on Drug Targets

Csaba Hetényi and Mónika Bálint

Abstract

Exploration of binding sites of ligands (drug candidates) on macromolecular targets is a central question of molecular design. Although there are experimental and theoretical methods available for the determination of atomic resolution structure of drug-target complexes, they are often limited to identify only the primary binding mode (site and conformation). Systematic exploration of multiple (allosteric or prerequisite) binding modes is a challenge for present methods. The Wrapper module of our new method, Wrap 'n' Shake, answers this challenge by a fast, computational blind docking approach. Beyond the primary (orthosteric) binding mode, Wrapper systematically produces all possible binding modes of a drug scanning the entire surface of the target. In several fast blind docking cycles, the entire surface of the target molecule is systematically wrapped in a monolayer of N ligand copies. The resulted target–ligand$_N$ complex structure can be used as it is, or important ligand binding modes can be further distinguished in molecular dynamics shakers. Wrapper has been tested on important protein targets of drug design projects on cellular signaling and cancer. Here, we provide a practical description of the application of Wrapper.

Key words Pocket, Peptide, Enzyme, Interaction, Inhibitor, Receptor, Mechanism, Action, Agonist, Antagonist

1 Introduction

There is a continuous increase in the number of atomic resolution structures of biomolecules available in public repositories such as the Protein Databank (PDB [1]). This promising trend is further facilitated by emerging cutting-edge techniques such as cryo-electron microscopy [2] allowing determination of structures of large biological entities such as viruses [3]. Despite the increase in the number of solved biomolecules, and high-throughput automation of X-ray crystallography [4], the measurement of structures of biomolecular targets in complex with their ligands remains a challenge and requires considerable time and money in many cases.

Molecular docking has been introduced as a computational counterpart of experimental techniques for the determination of

Zoltán Gáspári (ed.), *Structural Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, vol. 2112, https://doi.org/10.1007/978-1-0716-0270-6_8, © Springer Science+Business Media, LLC, part of Springer Nature 2020

target–ligand complex structures [5]. Thanks to its high speed, docking has been extensively applied in high-throughput screening campaigns of drug design projects [6] focusing on a known binding pocket of the target. Besides focused projects, docking has produced useful results if the search space was extended to the entire surface of the target molecule and the corresponding approach was named as blind docking [7, 8]. Blind docking has been extensively applied for finding allosteric [9–11] or multiple [12–16] binding sites. Like all other methods, docking also has numerous limitations coming from its approximations. First of all, it has been designed for focused search for drugs, and a systematic coverage of the entire target surface has not been implemented. Furthermore, starting ligand positions and steps of the search algorithm are mostly randomized which decreases reproducibility of the results. Modeling of flexibility (induced fit) and hydration of the target molecule is also oversimplified in docking programs to ensure fast results [17, 18]. Application of molecular dynamics simulations [19, 20] for blind docking is a reasonable approach to overcome the above hydration and flexibility problems of the fast methods. Nowadays, it is quite common to use realistic explicit water models with molecular dynamics, and flexibility can be obviously taken into account on both target and ligand sides. While these features of molecular dynamics considerably improve the precision of the calculated complex structure, they still cannot guarantee a systematic coverage of the entire surface of the target and correct location of the real binding pocket(s) during a single docking simulation [21].

To answer all these challenges of the blind docking problem, a new method Wrap 'n' Shake [21] was developed. The Wrapper module of Wrap 'n' Shake systematically finds all possible binding modes (sites and conformations) of a drug in several fast blind docking cycles. Wrap 'n' Shake has been tested on important protein targets of drug design projects on cellular signaling and cancer [21]. In the present paper, a detailed description of the protocol of the Wrapper module is provided to help future applications.

2 Materials

2.1 Preparation of Target and Ligand Molecules

Wrapper requires complete target and ligand molecules for proper results. Unfortunately, PDB structures of targets often have missing atoms or residues, which need to be inserted (*see* **Note 1**). In cases of missing terminal amino acids, acetyl and amide (*N*-methyl) capping groups need to be added to the N- and C-terminus, respectively. Such molecular editing and addition of hydrogen atoms can be performed by freely available modeling software such as Swiss-PdbViewer [22] or Schrödinger Maestro program package v. 9.6 [23]. Preparation of target structures is completed by energy minimization using free program packages such as GROMACS [24, 25]. For most of the protein targets, a uniform procedure with an AMBER99SB-ILDN force field [26], TIP3P explicit water model [27], and no restraints on the heavy atoms is appropriate. Ligand molecules can be built and edited by the above Maestro or other software. Protonation of the ligands (where applicable) is often helped by the pK_a plug-in in Marvin Sketch [28]. Fast energy minimization of the hydrogenated ligand structures is usually sufficient. In the first stage, molecular mechanics minimization with Maestro software is performed, using OPLS force field [29], followed by a quantum chemistry program package such as MOPAC [30] with a semiempirical parametrization such as PM6 or above.

2.2 Wrapper

The Wrapper module is available as part of a stand-alone, open source software package Wrap 'n' Shake freely downloadable from the web page of the program [31] along with full documentation. It is distributed under the terms of GNU General Public License. At present, Wrap 'n' Shake 1.1 contains software for the Wrapper module. Wrapper contains two bash scripts (pre-wrapper.sh and wrapper.sh) and a C program (wrp). After downloading the package (wns.tgz), it can be extracted using the following command:

```
$ tar -xvf wns.tgz
```

Pre-wrapper.sh and wrapper.sh can be found in wns/scripts and are readily usable under the Linux operating system. The source code of wrp can be compiled and installed into a $HOME/bin using the following commands:

```
$ cd wns/wrp/src
$ make
$ make install
```

The present version of Wrapper requires installation of external programs AutoGrid 4.2 and AutoDock 4.2 (Release 4.2.3) of the AutoDock 4.2 [32, 33] package, Python scripts of AutoDockTools [34], editconf and sasa programs of the GROMACS program package. All external programs are freely available. Organization of the components of Wrapper is shown in Fig. 1 and the programs are described as follows:

1. Script pre_wrapper.sh requires standard PDB files as input and prepares the files required by wrapper.sh. The necessary inputs for wrapper.sh are the PDBQT files of the ligand and target molecules and also grid (GPF) and docking (DPF) parameter files. The PDBQT file has the similar format to the regular PDB file, with additional columns containing the partial charges and the atom type. In wrapper, Gasteiger partial charges and the atom types of the modified AD4_parameters.dat (*see* also

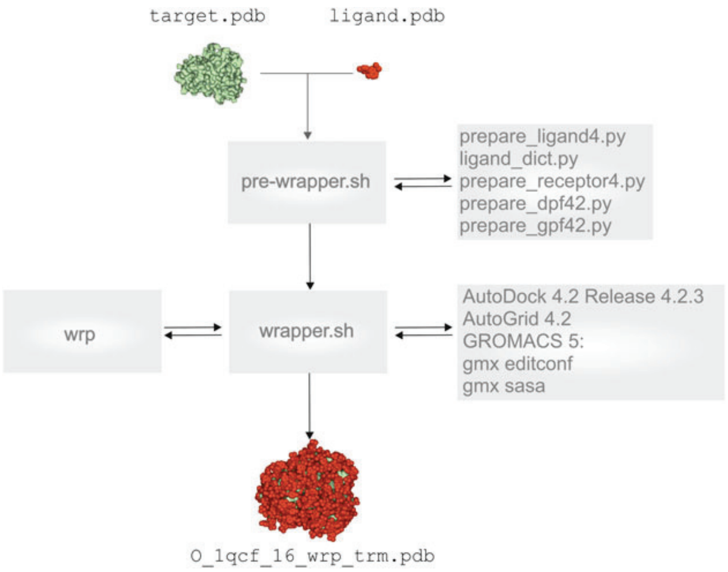


Fig. 1 Components of Wrapper and their connection with external shell scripts and programs. The figure was reproduced from the website of Wrap 'n' Shake with permission

Subheading 3.3) file are used. Notably, the original version of AD4_parameters.dat can be found in the source code folder of the AutoDock4.2 package. Both the ligand and target PDBQT apply united atom representation, which means that only the polar hydrogens are explicitly kept in the docking input file. The GPF file is the input of AutoGrid 4.2 and contains the docking (grid) box parameters. The grid box defines the search space where the docking calculations are performed. The GPF file also lists the names of target and ligand files and their atom types. The DPF file is the input file of AutoDock 4.2 and contains the parameters of the search algorithm and docking runs. The DPF also contains the names of map files generated by AutoGrid 4.2 for each atom type.

2. Wrapper.sh is the director of the Wrapper module. In several blind docking cycles, it covers the entire surface of the target with a monolayer of numerous ligand copies. Wrapper.sh works in symbiosis with program wrp of the present package detailed in the next point. The blind docking cycles are performed by external programs of the AutoDock 4.2 package and performed in separate working directories. After each cycle, free surface area of the target is calculated by external programs of the GROMACS package. Wrapper.sh reads PDBQT files of the ligand and target molecules and supplies the results as a single PDB file. For the ligand, a template file (ligand_templ.pdbqt) is also required for post-processing the wrapped target and used

in the trimming mode of Wrapper. During Wrapper, all ligand copies are renamed as "LIG" by the wrp program, and after ligand minimization, all atom names are renamed by MOPAC. Thus, the ligand template file is used for renumbering and renaming the ligand atoms and residue name after Wrapper. This ensures an exact match of the ligand atom names and ligand residue name with the molecular dynamics topology, which is required if the user merges the target ligand complex to use in a Shaker step. The atoms of the template file must have exactly the same order and number of heavy atoms as the input ligand.pdbqt file. The template file can be prepared by following the same input preparatory steps as for the ligand.pdb, except MOPAC minimization. Note that all hydrogen atoms must be added (Subheading 2.1) and the MOPAC energy minimization step is not required. After adding all hydrogen atoms, the PDB template file can be converted to a PDBQT file, using the command line of the python script below or the graphical interface of ADT program:

```
$pythonsh $PATH_TO/prepare_ligand4.py -l ligand_templ.pdb -o ligand_templ.pdbqt -v
-d $PATH_TO/ligand_dict.py -F
```

In this way, the same number and order of atoms is obtained in the template file as in the input PDBQT of the ligand. Wrapper.sh also produces log files containing reports on finished cycles with interaction energy and accessible surface values.

3. Wrp is an open source C program and serves as the background engine of the Wrapper module. It is called by wrapper.sh and performs clustering and ranking of the docked ligand conformations and subsequent assignation of excluded atoms. In wrapping mode, wrp results in a PDBQT file including the target, and all ligand copies accumulated up to the actual cycle and also a statistical file with ranking and intermolecular energy results (E_{inter}), calculated by the AutoDock 4.2 scoring function [35]. Wrp can also work in trimming mode where excess ligand copies not interacting with the target are removed after the final cycle and the results are written into a single PDB file identical with that one mentioned at wrapper.sh. This step is also initiated by script wrapper.sh. Repeated use of wrp in wrapping mode provides the target structure systematically covered in a monolayer of ligand copies. The work of wrp is adjusted by distance tolerance values as described in Subheading 3.4.

4. External python scrips (Table 1) of AutoDock Tools (ADT) are required by pre-wrapper.sh. The scripts are freely available [32, 34]. After ADT installation, these scripts can be found in

Table 1
Python scripts of ADT

Python script name	Input	Output
prepare_ligand4.py	PDB	PDBQT
ligand_dict.py	PDB	PDBQT
prepare_receptor4.py	PDB	PDBQT
prepare_dpf42.py	PDBQT	DPF
prepare_gpf4.py	PDBQT	GPF

a separate directory of the user: `$USER_HOME/MGLTools-1.5.6/MGLToolsPckgs/AutoDockTools/Utilities24`

The pythonsh binary is also installed, and insertion of an alias line in the .bashrc system file is advised, for easy access: `alias pythonsh=$USER_HOME/MGLTools-1.5.6/bin/pythonsh`

The python scripts generate PDBQT, DPF, and GPF files required by AutoDock 4.2 using the parameters described in Table 2. Based on the generated PDBQT files, ADT scripts also prepare grid and docking parameter files as required by AutoDock 4.2 [32].

We recommend the use of flexible ligand structures with torsional restriction on the aromatic and amide bonds only. Accordingly, branching of the torsion tree in the DPF files is generated with all default torsions of the ligand molecules as automatically assigned by ADT.

5. Blind docking runs of wrapper cycles are performed by external program package AutoDock 4.2. including program AutoGrid 4.2 for calculation of grid maps of the target molecule with pre-calculated energy values and the docking engine AutoDock 4.2 with a Lamarckian genetic algorithm. Docking parameters were used as described in a previous study [8]. The source code of the package was modified in order to be able to produce all the necessary map files in case of multiple target files. Original source code limits the number grid map generation to 14 atom types. Therefore, to produce grid map for all atom types, in autocomm.h file, line number 93 needs to be changed as follows.

   Original source code:

```
#define MAX_ATOM_TYPES (14 - NUM_NON_VDW_MAPS)
```

   Replaced by:

```
#define MAX_ATOM_TYPES (34 - NUM_NON_VDW_MAPS)
```

Table 2
Blind docking parameters

Parameter	Value
Grid parameters	
Grid spacing	0.375 Å
Number of grid points (*x*,*y*,*z*)	200,200,200
Docking parameters	
Search method	Lamarckian genetic algorithm
Population size	250
Maximum number of energy evaluations	20 million
Maximum number of generations	2000 million
Number of top individuals to survive to next generation	1
Rate of gene mutation	0.02
Rate of crossover	0.8
Alpha parameter of Cauchy distribution	0.0
Beta parameter of Cauchy distribution	1.0
Number of iterations of Solis and Wets local search	300
Consecutive successes before changing rho	4
Consecutive failures before changing rho	4
Size of local search space to sample	1
Lower bound on rho	0.01
Probability of performing local search on individual	0.06
Number of hybrid GA-LS runs	100

6. External GROMACS programs editconf and sasa [25] are called for calculation of accessible surface area of the target–ligand complex using a PDB file as input. The editconf command transforms the input pdb file intro gromacs .gro file, and the sasa program performs the calculations. GROMACS sasa calculates the ASA for the entire target–ligand complex, but wrapper.sh will eliminate the surface calculated for the ligand, by deleting rows, with residue name “LIG” from the total_atomarea_lig.xvg file obtained from GROMACS. Wrapper.sh also produces a log file containing the free target surface not covered by ligand copies.

3 Methods

3.1 Overview

Wrapper builds a monolayer of ligand copies covering the entire target molecule. Wrapper performs a series of automated, fast blind docking cycles. The algorithm ensures a complete and systematic coverage of the surface of the target with ligand copies. Wrapper uses a modified docking force field and clustering allowing maximal ligand–target and minimal ligand–ligand interactions. The popular docking program package AutoDock 4.2 is piped into Wrapper and performs consecutive fast blind docking cycles without the need of initial ligand positions or any other interventions of the user. The outcome of Wrapper is a single PDB file including the structure of the target wrapped in a monolayer of ligand copies, i.e., the structure of a target–ligand$_N$ complex. The application of Wrapper is described using an example (Fig. 2) of the complex of hematopoietic cell kinase (HCK, target, in green) and 1-ter-butyl-3-p-tolyl-1H-pyrazolo[3,4-D]pyrimidin-4-ylamine (PP1, ligand, in red). The complex structure was published under PDB code 1qcf, and this code will be used in the names of input and output files of the example also provided for download on the web page of the program [31].

3.2 Input Files

Wrapper requires complete, energy-minimized structures of the ligand (1qcf_ligand.pdb, red) and target (1qcf_target.pdb, green) molecules in Protein Databank (∗.pdb) format. Preparation of target and ligand molecules is described in Subheading 2.

3.3 Pre-wrapper.sh

From both target and ligand structures, pre-wrapper.sh produces PDBQT input files (1qcf_target.pdbqt, 1qcf_ligand.pdbqt) and parameter files (1qcf_target.gpf, 1qcf_target.dpf) as required by AutoDock 4.2 called by wrapper.sh. The docking box is set to cover the entire surface of the target molecule. For this, the center of the box is set to that of the target molecule (default option), and grid maps of 200 grid points in all three spatial directions are generated. Notably, if the size of the target exceeds ca. 450 amino acids corresponding to the largest proteins of our test set (Fig. 3), the number of grid points of 200 should be increased in the following command of the pre-wrapper.sh script calling prepare_gpf4.py in order to cover the whole target in one BD cycle.

```
$SCRIPTPATH/pythonsh $SCRIPTPATH/prepare_gpf4.py
-l $ligand_name.pdbqt -r $target_name.pdbqt -p spacing=0.375
-p npts='200,200,200' -p ligand_types='A,..,YY,LL' -v
```

With this, the numbers of grid points are specified in GPF for all three directions of space. The user must also consider the shape of the target and change the box dimensions in one or all directions

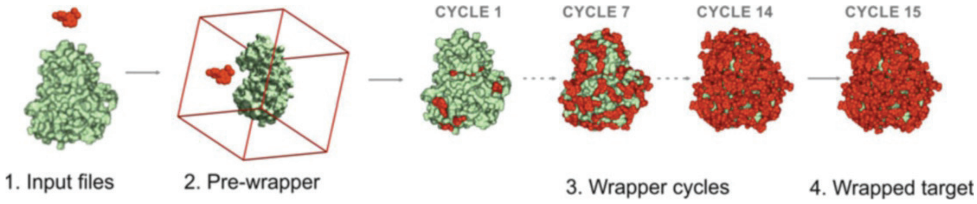
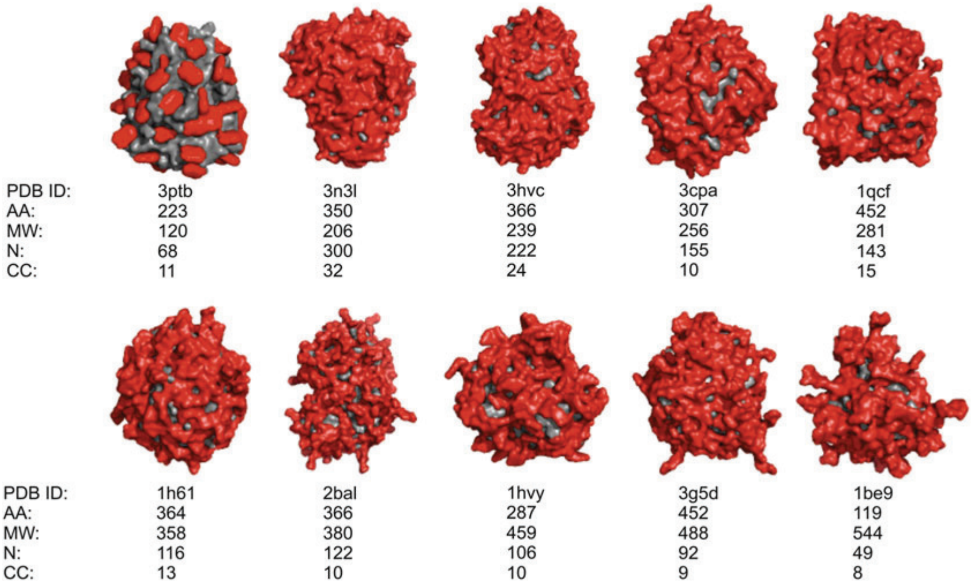


Fig. 2 Main stages of Wrapper. The target (hematopoietic cell kinase, green) is wrapped in numerous copies of the ligand (1-ter-butyl-3-p-tolyl-1H-pyrazolo[3,4-D]pyrimidin-4-ylamine, red) molecule in several blind docking cycles. The docking box (red lines) covers the entire surface of the target molecule. The figure was reproduced from the website of Wrap 'n' Shake with permission



PDB ID:	3ptb	3n3l	3hvc	3cpa	1qcf
AA:	223	350	366	307	452
MW:	120	206	239	256	281
N:	68	300	222	155	143
CC:	11	32	24	10	15

PDB ID:	1h61	2bal	1hvy	3g5d	1be9
AA:	364	366	287	452	119
MW:	358	380	459	488	544
N:	116	122	106	92	49
CC:	13	10	10	9	8

Fig. 3 Targets (grey) wrapped in a monolayer of ligand (red) copies. *AA* count of amino acids of the target, *MW* molecular weight of the ligand, *N* number of ligand copies, *CC* count of cycles. (The figure was reproduced from the website of Wrap 'n' Shake with permission)

accordingly (*see* **Note 2**). An edge of the box can be calculated in Ångström as the product number of grid points and grid spacing (a value of 0.375 Å was used; Table 2). Pre-wrapper.sh also adds new entries of excluded atom types LL and YY (commonly marked as X in our original publication [21]) to the DPF and GPF files. This step is performed only once, as the same parameter files can be used in all wrapping cycles later. This step is necessary for generation of the map files of the new atom types. Gasteiger partial charges are added to both the ligand and target. Addition of hydrogen atoms to the ligand or target is skipped as the minimized PDB files already have all atoms. The nonpolar hydrogens are merged (Subheading 2.2). All default active torsions are kept for the ligand, but the target is treated rigidly, without active torsions. Parameter files have the settings as described in Subheading 2.2 (Table 2).

Pre-wrapper also performs three important administrative modifications on text files.

1. The first row of the parameter files (both the DPF and GPF) is updated to the actual path of the modified AD4_parameters.dat.

   Default:

```
autodock_parameter_version 4.2
```

   Modified:

```
parameter_file $USER_DEFINED_PATH/AD4_parameters.dat
```

2. New lines of atom types LL and YY are inserted after the last line of standard atom type maps.

```
map 1qcf_target.YY.map
map 1qcf_target.LL.map
```

3. Two lines of atom types LL and YY are inserted to the end of AD4_parameters.dat file (the modified file can be also downloaded from our web page [31]).

```
atom par YY 3.60 1E-04 00.0000 0.00000 0.0 0.0 0 0 0 0
atom par LL 3.60 1E-04 00.0000 0.00000 0.0 0.0 0 0 0 0
```

   The user may decide to prepare the input PDBQT, DPF, and GPF using the graphical interface of ADT instead of pre-wrapper.sh. In this case, after generating the DPF and GPF, the above detailed three changes should be also done by manual editing of the files. Whereas the use of pre-wrapper.sh is not mandatory as file preparations can be arranged as described above; however, the use of pre-wrapper.sh is recommended to avoid human mistakes especially if multiple target files or a library of ligand structures are handled.

3.4 Wrapper.sh and wrp

Wrapper.sh performs the coverage of target surface with a monolayer of N ligand copies ending up in a target–ligand$_N$ complex. Several fast BD cycles are performed all of them resulting in 100 docked ligand copies. The count of necessary BD cycles (CC) depends on the size and shape of the target molecule as indicated in Fig. 3. Ligand copies and interacting target surface elements are excluded from successive BD cycles via assignation of a new "excluded" atom type to the atoms involved. In this way, unbound target sites can be distinguished from those covered with ligand copies, ligand-ligand interactions are minimized, and target–ligand interactions are maximized for the largest possible coverage of the target surface. Further details on structural and physical chemistry of the Wrapper algorithm can be found in the original publication of Wrap 'n' Shake [21].

The BD cycles follow a uniform protocol. Grid map files (1qcf_target_∗.map) of chemical and excluded (YY, LL) atom types are calculated by Autogrid 4.2 along with a log file. The corresponding ∗.YY.map and ∗.LL.map files are generated before the docking runs. One hundred BD runs are performed in each cycle, and the docked ligand structures are collected in a log file (1qcf_1.dlg for the first cycle) by AutoDock 4.2. The log file is evaluated by the wrp program, which first ranks and clusters the docked ligand conformations.

Docked ligand conformations of the DLG file are clustered and ranked based on their interaction energy (E_{inter}, the first energy component of estimated free energy of binding in the DLG file) values with the target and the closest distance between each heavy atom of the ligand copies (dmin). In the initial clustering phase, wrp (wrapper mode) sorts the 100 docked ligand conformations according to $E_{\mathrm{inter.}}$ Ligand conformation of the lowest E_{inter} from among the 100 docked ligand copies is selected as the representative of Cluster 1. Ligand conformation of the second lowest E_{inter} is selected as a representative of a new Cluster 2 if dmin>drnk, where drnk is a ranking tolerance, a measure of separation of clusters from each other. If dmin≤drnk, then ligand conformation of the second lowest E_{inter} is placed into Cluster 1. In this way, all 100 ligand conformations are clustered, and the representatives are evenly spread over the target surface without clashing each other. In our protocol, drnk was set to 2 Å, which is approximately a covalent bond distance (1.5 Å) plus a 0.5 Å added. The results of clustering are summarized in .sta file type (O_1qcf_1_wrp.sta) after each wrapper cycle.

Wrp in wrapper mode assigns the new atom type (YY, LL) of the abovementioned excluded atoms in the target file (YY) and the docked ligand copies (cluster representatives LL). Excluded atoms are assigned using a target–ligand interface tolerance and an assignation tolerance. Both of these tolerance values were set to 3.5 Å in our default settings. Merging of the modified target and ligand copies results in a target–ligand complex O_1qcf_1_wrp.pdbqt file. This file is moved from the working directory of the current cycle into the directory of the next cycle and used as target input for programs AutoGrid 4.2 and AutoDock 4.2 if none of the exit criteria described below are achieved. After each cycle, the free (unliganded) accessible surface area (ASA) is calculated by external GROMACS program sasa, as described in Subheading 2, Point 6 (Msroll in the 1.0 version). Wrapping ends if ASA ≤ 1% or the interaction energy E_{inter} value of any cluster representative in the cycle is ≥0 kcal/mol. Otherwise, the resulted PDBQT file is forwarded to the next cycle as described above. ASA and E_{inter}

evaluations are calculated for each wrapper cycle and stored in two separate files (O_1qcf_1_surface_percentage.log and O_1qcf_1_lowest_energy.log). These files are generated in the working directory of each cycle and moved to "stats" folder where statistical evaluation of Wrapper takes place.

For our test system 1qcf, wrapping finished in 16 cycles and 1qcf_16_wrp.pdbqt is the result after the last cycle. All files of the complete Wrapping process of 16 wrapping cycles can be downloaded as a single compressed package (O_1qcf_wrp.tgz).

After the last (16th) wrapping cycle, a trimming mode of wrp is involved to remove ligand copies positioned far from the target surface. This is necessary, as some ligand copies may dock to distant regions of the docking box depending on the actual target. The trimming step also performs formal post-processing of the 1qcf_16_wrp.pdbqt file using a template file (1qcf_ligand_templ.pdbqt) described in Subheading 2.2. The resulted O_1qcf_16_wrp_trm.pdb file has all atoms renamed according to the standards of PDB file format allowing the use of this file of the molecular dynamics steps of a Shaker process (*see* **Note 4**).

3.5 Output, Benchmark

In our example, the target structure was wrapped in a monolayer of $N = 143$ ligand copies in 15 cycles (Fig. 3). The CPU time of a cycle of 100 docking runs took 11 h for this system on an Intel Xeon E5520. In general, CPU times of a cycle varied between some hours and 1–2 days for the test systems listed in Fig. 3 depending on the size of the target molecule and the size and number of rotatable torsion of the ligand (*see* **Note 3**). The count of cycles (CC in Fig. 3) necessary for complete wrapping depends both on the size and geometry of the partners. The largest ligand (system 1be9) fully covered its relatively small target in less than ten cycles. The largest CC of 32 was found for system 3n3l, where the ligand is relatively small and the target is large. The special geometry of ligand benzamidine is probably a reason for the unique wrapping pattern corresponding to unexpectedly low N and CC values obtained in the case of system 3ptb.

4 Notes

1. During pre-wrapper.sh, it is useful to check the net charge (sum of partial charges of all atoms in the PDBQT file) of the target and ligand molecules. The value of the net charge of a PDBQT file should be close to an integer. For example, a net charge of 3.5 indicates that the structure of the molecule is erroneous (missing/extra atoms), or partial charges could not be assigned correctly by ADT. In this way, checking of net charge helps the detection of error occurring during the preparation of target or ligand structures. Special attention must also be given to the

charge assigned on systems with coordinating ions (e.g., Fe^{3+}, Ni^{2+}, etc.) as the partial charges assigned for such atoms by ADT are not always correct [36].

2. The user should check if the grid box covers the whole target; otherwise, parts of the target surface excluded from the box will not be analyzed for possible binding sites. The grid box can be visualized by a python script called gbox.py downloadable from the website of Wrap 'n' Shake [31].
3. We suggest running pre-wrapper.sh on a simple workstation (personal computer, PC) as it requires only some seconds to finish. Wrapper.sh can also be run on a simple PC under Linux. However, as complete wrapping of a target usually takes several hours or days of CPU time, its frequent application may require a dedicated PC or a server node.
4. The Shaker protocol of Wrap 'n' Shake [21] can be used for distinction of important binding modes and structural refinements on hydration and induced fit effects in successive molecular dynamics steps. The wrapped target is placed in a simulation box and hydrated with explicit water molecules. The hydrated complex is subjected to a series of simulations and filtering steps between the MD runs, where loosely bound ligand copies are removed. Refinement of bound ligand structure can be performed with all target atoms released.

Acknowledgments

We acknowledge a grant of computer time from CSCS Swiss National Supercomputing Centre and the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research have been achieved using the DECI resource Archer based in the UK at the National Supercomputing Service with support from the PRACE aisbl. The work was supported by the K123836 grant from the National Research, Development, and Innovation Office, Hungary. The University of Pécs is acknowledged for a support by the 17886-4/23018/FEKUTSTRAT excellence grant. The work was supported by EFOP-3.6.2-16-2017-00008 and the ÚNKP-19-4 New National Excellence Program of the Ministry for Innovation and Technology. The work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

1. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS et al (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39(Suppl 1):392–401
2. Nogales E (2018) Profile of Joachim Frank, Richard Henderson, and Jacques Dubochet,

2017 Nobel laureates in chemistry. Proc Natl Acad Sci U S A 115(3):441–444

3. Cheng Y, Glaeser RM, Nogales E (2017) How Cryo-EM became so hot. Cell 171 (6):1229–1231
4. Hui R, Edwards A (2003) High-throughput protein crystallization. J Struct Biol 142 (1):154–161
5. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935–949
6. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC et al (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45(11):2213–2221
7. Hetényi C, van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11 (7):1729–1737
8. Hetényi C, Van Der Spoel D (2006) Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett 580:1447–1450
9. Hocker H, Rambahal N, Gorfe AA (2014) LIBSA – a method for the determination of ligand-binding preference to allosteric sites on receptor ensembles. J Chem Inform Model 54 (2):530–538
10. Schindler CEM, De Vries SJ, Zacharias M (2015) Fully blind peptide-protein docking with pepATTRACT. Structure 23 (8):1507–1515
11. Whalen KL, Tussey KB, Blanke SR, Spies MA (2011) Nature of allosteric inhibition in glutamate racemase: discovery and characterization of a cryptic inhibitory pocket using atomistic MD simulations and pKa calculations. J Phys Chem B 115(13):3416–3424
12. García-Sosa AT, Sild S, Maran U (2008) Design of multi-binding-site inhibitors, ligand efficiency, and consensus screening of avian influenza H5N1 wild-type neuraminidase and of the oseltamivir-resistant H274Y variant. J Chem Inf Model 48(10):2074–2080
13. Roumenina L, Bureeva S, Kantardjiev A, Karlinsky D, Andia-Pravdivy JE, Sim R et al (2007) Complement C1q-target proteins recognition is inhibited by electric moment effectors. J Mol Recognit 20(5):405–415
14. Bugatti A, Giagulli C, Urbinati C, Caccuris F, Chiodelli P, Oreste P et al (2011) Molecular interaction studies of HIV-1 matrix protein p17 and heparin: identification of the heparin-binding motif of p17 as a target for the development of multitarget antagonists. J Biol Chem 288(2):1150–1161
15. Kovács M, Tóth J, Hetényi C, Málnási-Csizmadia A, Seller JR (2004) Mechanism of blebbistatin inhibition of myosin II. J Biol Chem 279(34):35557–35563
16. Agarwal T, Annamalai N, Khursheed A, Kumar T, Bin H, Haris M (2015) Molecular docking and dynamic simulation evaluation of Rohinitib—Cantharidin based novel HSF1 inhibitors for cancer therapy. J Mol Graph Model 61:141–149
17. Rastelli G, Ferrari AM, Costantino L, Gamberini MC (2002) Discovery of new inhibitors of aldose reductase from molecular docking and database screening. Bioorg Med Chem 10 (5):1437–1450
18. García-Sosa AT, Mancera RL (2010) Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex. Mol Inform 29(8–9):589–600
19. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE (2011) How does a drug molecule find its target binding site? J Am Chem Soc 133(24):9181–9183
20. Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci U S A 108 (25):10184–10189
21. Bálint M, Jeszenői N, Horváth I, van der Spoel D, Hetényi C (2017) Systematic exploration of multiple drug binding sites. J Cheminform 9(1):65
22. Guex N, Peitsch MC, Schwede T (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis 30 (Suppl 1):162–173
23. Maestro, Schrödinger, LLC (2017) New York, NY
24. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B et al (2015) Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25
25. GROMACS (2018). Available from: http://manual.gromacs.org/current. Accessed 01 Oct 2018
26. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO et al (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78(8):1950–1958
27. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of

simple potential functions for simulating liquid water. J Chem Phys 79(2):926–935

28. Chemaxon (2014) Marvin Sketch, v. 6.3.0. Chemaxon, Budapest
29. Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci U S A 102 (19):6665–6670
30. MOPAC (2012) MOPAC. Stewart JJP, computational chemistry
31. Wrap'n'Shake (2017). Available from: http://www.wnsdock.xyz. Accessed 01 Oct 2018
32. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 28(1):73–86
33. Autodock 4.2 (2009). Available from: http://www.autodock.scripps.edu. Accessed 28 Sept 2018
34. AutoDock Tools 1.5.6 (2009). Available from: http://mgltools.scripps.edu/downloads. Accessed 28 Sept 2018
35. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19 (14):1639–1662
36. Bikadi Z, Hazai E (2009) Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. J Cheminform 1(1):1–16

hetenyi.csaba_83_23

D6

pharmaceuticals

MDPI

Article

Prerequisite Binding Modes Determine the Dynamics of Action of Covalent Agonists of Ion Channel TRPA1

Balázs Zoltán Zsidó [1], Rita Börzsei [2], Erika Pintér [1] and Csaba Hetényi [1,*]

[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary; zsido.balazs@pte.hu (B.Z.Z.); erika.pinter@aok.pte.hu (E.P.)

[2] Department of Pharmacology, Faculty of Pharmacy, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary; rita.borzsei@gmail.com

[*] Correspondence: hetenyi.csaba@pte.hu

Abstract: Transient receptor potential ankyrin 1 (TRPA1) is a transmembrane protein channeling the influx of calcium ions. As a polymodal nocisensor, TRPA1 can be activated by thermal, mechanical stimuli and a wide range of chemically damaging molecules including small volatile environmental toxicants and endogenous algogenic lipids. After activation by such compounds, the ion channel opens up, its central pore widens allowing calcium influx into the cytosol inducing signal transduction pathways. Afterwards, the calcium influx desensitizes irritant evoked responses and results in an inactive state of the ion channel. Recent experimental determination of structures of apo and holo forms of TRPA1 opened the way towards the design of new agonists, which can activate the ion channel. The present study is aimed at the elucidation of binding dynamics of agonists using experimental structures of TRPA1-agonist complexes at the atomic level applying molecular docking and dynamics methods accounting for covalent and non-covalent interactions. Following a test of docking methods focused on the final, holo structures, prerequisite binding modes were detected involving the apo forms. It was shown how reversible interactions with prerequisite binding sites contribute to structural changes of TRPA1 leading to covalent bonding of agonists. The proposed dynamics of action allowed a mechanism-based forecast of new, druggable binding sites of potent agonists.

Keywords: TRPA1 receptor; prerequisite binding; covalent binding

check for updates

Citation: Zsidó, B.Z.; Börzsei, R.; Pintér, E.; Hetényi, C. Prerequisite Binding Modes Determine the Dynamics of Action of Covalent Agonists of Ion Channel TRPA1. *Pharmaceuticals* **2021**, *14*, 988. https://doi.org/10.3390/ph14100988

Academic Editor: Arpad Szallasi

Received: 23 August 2021
Accepted: 24 September 2021
Published: 28 September 2021

1. Introduction

Mammalian neurons of the pain pathway detect potentially dangerous environmental signals. In the peripheral nervous system, there are specialized nociceptive neurons, that recognize either noxious chemical signals, thermal or mechanical stimuli. Pathological processes, such as tissue damage and inflammation elicit the formation and subsequent release of a wide variety of mediators (arachidonic acid derivatives, free radicals, H_2O_2, H_2S, etc.). These molecules depolarize the nerve terminals of nociceptors, which transmit the signals to the central nervous system. The transient receptor potential ankyrin 1 (TRPA1) is a Ca^{2+}-permeable cation channel that was identified as the chemical nocisensor, expressed by primary afferent nerve fibers [1–9]. Activated TRPA1 promotes pain itching and induces local neurogenic inflammatory response via the release of neuropeptides, such as substance P, calcitonin gene-related peptide and neurokinins.

TRPA1 receptor is activated by the binding of electrophile ligands (Figure 1) to its N-terminus cytoplasmic binding site (Figure 2A), which is characterized by three nucleophilic cysteine residues (C621, C641 and C665) [6,7]. This binding event induces a local conformational change, that is translated to the whole of the receptor, and a 15° rotation of the transmembrane domain is observed [7], resulting in a pore widening, that facilitates Ca^{2+} influx, which first potentiates, then desensitizes agonist-induced responses [7], resulting in an inactive state of the TRPA1 ion channel [10]. On the local scale, a cytoplasmic

A-loop near the transmembrane region of the receptor overlays the binding site cavity [7] and initially sterically hinders agonist binding. Upon the binding of the electrophile agent, a flip of the A-loop (residues 666–680, Figure 2B) was observed towards the cell membrane, leaving more binding space for the agonists. Thus, flipping of A-loop contributes to the widening of the pore of the ion channel, and the above-mentioned activation process [7] at the same time, and therefore, it is important in agonist design.

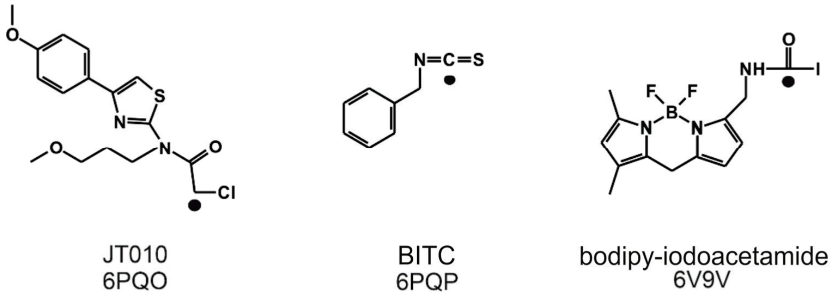


Figure 1. The Lewis structures of three TRPA1 agonists, JT010, benzyl-isothiocyanate (BITC) and bodipy-iodoacetamide, and the PDB IDs of their complexes with the TRPA1 receptor. The original molecular structures were restored prior to the covalent bond formation. A chlorine was added to JT010, an iodine to bodipy-iodoacetamide and the geometry of the N=C=S bond of BITC was restored. The atoms that participate in the formation of the covalent bond are marked by black dots.

Known electrophile agonists of the TRPA1 receptor include dimethyl trisulfide (DMTS) and allyl isothiocyanate (AITC) [4,5,8]. However, given the size of these molecules, they would not bind selectively to only one cysteine amino acid residue in the body. Thus, the need for a selective site-specific electrophile agonist (JT010) was first met in 2015 by Takaya et al. [4], and a potent thiazol derivate agonist (EC_{50} = 0.65 nM, Figure 1) with a covalent alkyl-halide warhead was designed. Since then, the binding position of JT010 was experimentally found by cryo-electron microscopy [7]. Recent structural studies [2,3,6,7] provided additional details of the agonist binding mechanism and consequential receptor activation. Besides JT010, the binding of the other two site-specific covalent agonists BITC [6] and bodipy-iodoacetamide [7] (Figure 1) was investigated, which preferentially bind to the active site cysteine C621 of the TRPA1 receptor.

Covalently binding agonists are subjects of intense research [11], and they often form covalent bonds with nucleophilic cysteine residues. As cysteine is abundant in the human proteome, a careful design has to be performed to achieve binding site specificity [11,12] to avoid unwanted promiscuity of the agonists via non-selective covalent bonds with non-targeted cysteines. Thus, a common strategy of covalent agonist design adopts known agonists having selective non-covalent interactions [11] with the receptor. Both covalent and non-covalent interactions have been fully described [6,7] at the final, irreversible binding mode of the agonists in Figure 1. However, the binding routes leading to the final binding pocket have not been mapped at atomic resolution. The above-mentioned, agonist-induced structural changes of TRPA1 at the A-loop (Figure 2B) suggest that the agonist may form dynamic interactions with prerequisite sites on their route to the final, covalent positions.

The present study is focused on the elucidation of binding dynamics of covalent agonists using experimental structures of their complexes with TRPA1 (holo form) and compared to prerequisite interactions with the apo form. Covalent and non-covalent molecular docking techniques are tested at forecasting prerequisite and final, covalent binding modes of the agonists. The docked agonist-TRPA1 complexes are subjected to molecular dynamics calculations to complete the binding mechanism at the atomic level. We also aim at the mechanism-based forecasting of prerequisite binding sites which can become druggable targets of agonists in drug design projects.

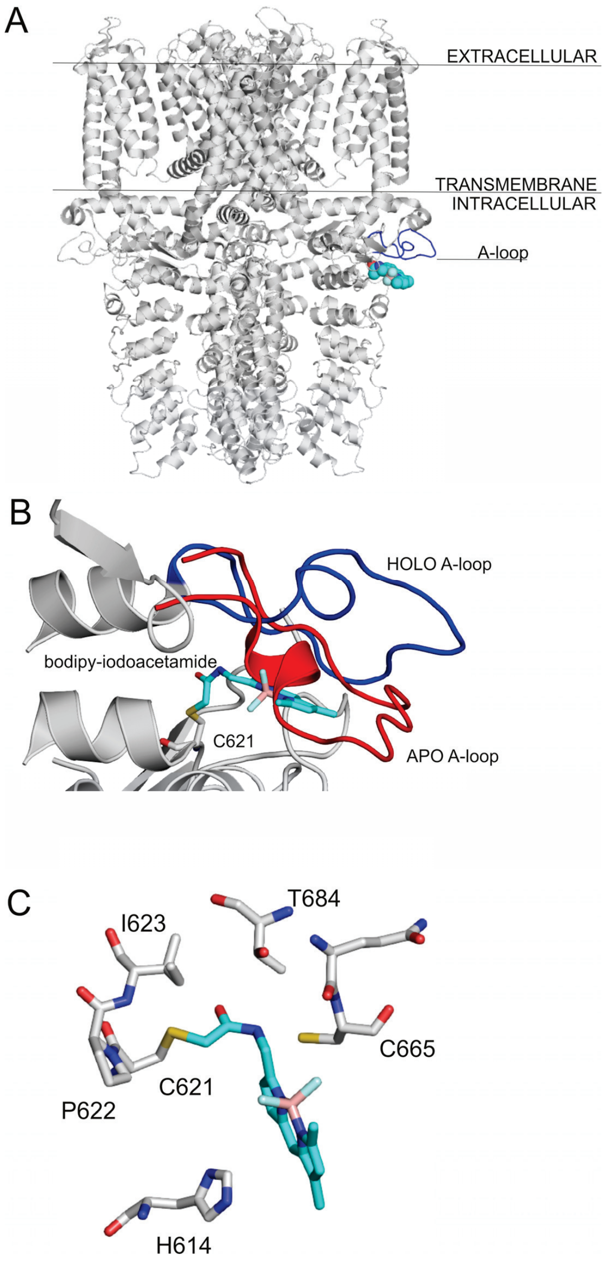


Figure 2. (**A**) The TRPA1 ion channel shown as grey cartoon representation, the agonist binding site is shown by the binding of bodipy-iodoacetamide (PDB: 6v9v) as teal spheres. The A-loop is highlighted with blue; (**B**) the movement of the A-loop during ligand binding. The figure was prepared with the superposition of 6v9w on 6pqo. The blue loop is A-loop in the holo, open form, and the red is A-loop in the apo, closed form. The rest of the binding site is shown with grey cartoon. bodipy-iodoacetamide is shown in teal sticks, and C621 in all atom representation grey sticks. (**C**) The close-up of the binding of bodipy-iodoacetamide to the agonist binding site of TRPA1, interacting amino acids are shown as grey thick lines in all atom representation, bodipy-iodoacetamide is shown as teal all atom representation sticks.

2. Results and Discussion

2.1. Final Covalent Binding Modes

The atomic resolution structures of three covalent agonists JT010 [4], BITC [6] and bodipy-iodoacetamide [7] (Figure 1) bound to the TRPA1 ion channel are available in the Protein Databank (PDB). JT010 and bodipy-iodoacetamide have an alkyl-halide covalent warhead, the bonds made with chlorine and iodine atoms break up during covalent binding to C621 (Figure 3). BITC participates in an isothiocyanate bond with C621 (Figure 3), the N=C double bond diminishes, and the C atom forms a covalent bond with the S atom of the target C621, and the N atom of BITC gains a H atom. As the nucleophilic C621 binds all three electrophilic agonists covalently [6,7], our docking studies focused on the surrounding binding pocket.

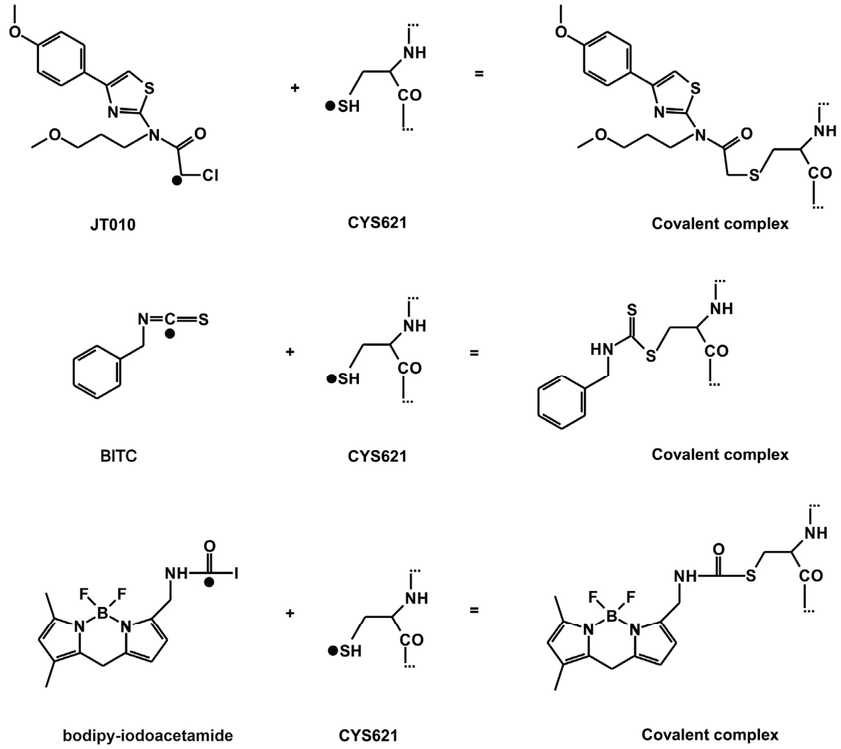


Figure 3. The reaction schemes of JT010, BITC, bodipy-iodoacetamide. Atoms, that participate in the formation of the covalent bond are highlighted by a black dot (●). The distance of these two atoms is referred to as d.

As a first step of our investigation, the popular program package FITTED [13–15] was tested using the experimental PDB structures as references for comparison with the docking results. FITTED was first tested to reproduce the final binding modes of the covalently bound agonists. A standard evaluation protocol was applied to all covalent docking results. Firstly, the structural match of the calculated binding mode (bind position, orientation, and conformation of the ligand) to the crystallographic reference was calculated and the best match was expressed as a root mean squared deviation ($RMSD_{\text{best}}$, see Section 4 for the definition of all metrics used in the Tables) [16]. Secondly, it was tested if the docking method identified $RMSD_{\text{best}}$ as an energetically favorable binding mode and ranked it at the top of the list of all binding modes. The second criterion reports on the applicability of the binding free energy (scoring) function of the docking method.

The covalent docking of the agonists to the holo form of TRPA1 was structurally successful as the $RMSD_{\text{best}}$ values were comparable/below 2.5 Å (Table 1), a threshold accepted in the literature [17–21]. Only bodipy-iodoacetamide showed a slightly elevated *RMSD* (Table 1), which is due to the mobility of the -SH group around the C_{β} of C621 during docking. The rotation of the S atom around the C_{β} of C621 also turns bodipy-iodoacetamide from its experimental position around its longitudinal axis by ca. 180°. The C_{β}-S-$C_{\text{bodipy-iodoacetamide}}$ angle is also smaller than that observed in the experimental structure. The binding modes with $RMSD_{\text{best}}$ were positioned to the first place on the ranking lists for all three agonists (Table 1). The calculated free energy of binding of JT010 is the most favorable, closely followed by that of BITC (Table 1), apparently, the alkyl-halide covalent bond formed by the alkyl-iodine pharmacophore of bodipy-iodoacetamide is less favorable compared to that formed by the alkyl-chlorine pharmacophore of JT010. The efficiency index (EI_{NHA}) of BITC is the best out of the three reference ligands.

Table 1. Covalent docking calculations performed by FITTED [13–15].

Ligand Name	JT010	BITC	Bodipy-Iodoacetamide
	HOLO target		
AA_{match} (%)	100%	100%	100%
$RMSD_{\text{best}}$ (Å)	2.28	2.05	3.87
$\text{Rank}_{\text{best}}$	1/3	1/3	1/3
ΔG_{FD} (kcal/mol)	−84.1	−77.7	−44.3
NHA [c]	23	10	22
EI_{NHA} [d] (kcal/mol)	3.66	7.77	2.01
d_{covalent} (Å)	1.8 (1.8) [a]	1.8 (1.8) [a]	1.8 (1.8) [a]
	APO target [b]		
AA_{match} (%)	100%	60%	66.6%
$RMSD_{\text{best}}$ (Å)	6.82	4.75	6.55
$\text{Rank}_{\text{best}}$	1/5	1/5	1/5
ΔG_{FD} (kcal/mol)	−77.4	−73.8	−43.1
NHA [c]	23	10	22
EI_{NHA} [d] (kcal/mol)	3.36	7.38	1.96
d_{covalent} (Å)	1.8	1.8	1.8

[a] The experimental covalent bond lengths are shown in brackets; [b] Without A-loop; [c] Number of heavy atoms; [d] Efficiency index.

The large conformational flexibility of target molecules is a challenge for fast docking programs and can be handled by the involvement of molecular dynamics approaches [22] which require longer calculation times. Conformational flexibility is important when induced fit occurs during agonist binding. Here, the A-loop is a flexible element that covers the binding site (Introduction, Figure 2B) in the apo TRPA1 conformation, and sterically prevents the agonist from reaching its destination (holo position) at the bottom of the pocket resulting in unacceptably large *RMSD* values and positive energies (Tables S1–S3). Therefore, the A-loop was removed from TRPA1 and covalent docking calculations were repeated using FITTED. Although the A-loop consists of the amino acids from 660–680, in the apo docking calculations, only the amino acids 665–677 were removed, as these are the ones that elicit the greatest movement during apo to holo transition. In the absence (Tables 1–3) of the A-loop, all covalent bonds were formed with the apo TRPA1 (Table 1), with *RMSD* values larger than those observed in the case of the holo TRPA1 (Table 1). The corresponding ΔG_{best} values were lower than those of the apo TRPA1 by 6% on average. These findings emphasize the role of A-loop-agonist interactions in the final binding position. However, the removal of the A-loop did not influence the EI and ΔG_{best} order of the ligands.

Table 2. Non-covalent docking calculations performed by FITTED.

Ligand Name	JT010	BITC	Bodipy-Iodoacetamide
		HOLO target	
ΔG_{FD} (kcal/mol)	−46.1	−32.4	−13.7
$\text{Rank}_{\text{best}}$	10/10	1/10	8/10
AA_{match} (%)	100%	100%	100%
d_{best} (Å)	3.6	4.0	8.7
		APO target [a]	
ΔG_{FD} (kcal/mol)	−33.4	−26.7	0.5
$\text{Rank}_{\text{best}}$	3/5	1/5	4/5
AA_{match} (%)	100%	60%	33.3%
d_{best} (Å)	3.5	3.9	3.3

[a] Without A-loop.

The formation of the covalent bond between the agonist and TRPA1 (Figure 3) is a quantum mechanical phenomenon, which is hard to treat adequately by docking programs based on molecular mechanics scoring functions [23,24]. Despite the above challenges, the covalent docking methodology of FITTED performed well for the above test cases and supplied relevant structural and scoring (ranking) results. Encouraged by the above test results, we expect that FITTED will also help in mapping the prerequisite binding modes on route to the final binding pocket.

Table 3. Non-covalent docking calculations performed by AutoDock 4.2 [25].

Ligand Name	JT010	BITC	Bodipy-Iodoacetamide
		HOLO target	
ΔG_{AD} (kcal/mol)	−6.8	−3.8	−5.9
$\text{Rank}_{\text{best}}$	1/5	1/1	4/4
AA_{match} (%)	100%	80%	66%
d_{best} (Å)	3.6	6.5	4.0
		APO target [a]	
ΔG_{AD} (kcal/mol)	−5.16	−3.74	−5.26
$\text{Rank}_{\text{best}}$	1/3	1/2	3/5
AA_{match} (%)	50%	40%	0%
d_{best} (Å)	7.5	7.2	7.3

[a] Without A-loop.

2.2. Prerequisite Binding Modes

As it was discussed in Section 1, the entrance to the binding cavity (outer prerequisite binding mode) and the formation of the final ligand-target covalent bond (inner prerequisite binding mode) is hindered by the position of the A-loop (Figure 2B) in the apo form of the TRPA1 target. During a successful binding process, the ligand initiates the flipping of the A-loop via intermolecular interactions with the loop. To develop such interactions, the ligand needs to occupy a prerequisite binding mode outside the final binding pocket. Two different programs, FITTED and AutoDock 4.2.6 (The Scripps Research Institute, La Jolla, CA, USA) [25], were involved in the mapping of possible prerequisite binding modes by non-covalent docking and the results are shown in Tables 2 and 3, respectively. Both programs are based on physico-chemical principles. FITTED is a genetic algorithm-based docking method, that includes an ESFF [26] force field-based search engine, called CDiscoVer [27] to perform conjugate gradient minimizations [13]. AutoDock also uses a (Lamarckian) genetic algorithm and AMBER-based intermolecular force field terms for scoring [25].

For prerequisite binding modes *RMSD* was not calculated, the distance (d) between TRPA1 C621 S atom and the atom of the agonist that participates in the covalent binding was used as a measure of ligand position instead. The d_{best} value indicates the closest distance between the aforementioned atoms (black dots in Figure 3) achieved by subsequent docking calculations. The match of the docked binding mode was expressed as the percentage of matching amino acids (AA_{match}) compared with the experimental binding pocket. Both programs ranked the binding mode of BITC with the d_{best} as the top 1st in all prerequisite docking calculations (Tables 2 and 3). However, BITC is considerably smaller, than JT010 and bodipy-iodoacetamide. The head-to-tail docking orientation of larger agonists, like JT010 and bodipy-iodoacetamide might cause elevated d_{best} values.

In all cases and both scorings, the holo prerequisite docking calculations yielded better AA_{match} and ΔG_{FD} and ΔG_{AD}, than the apo prerequisite docking calculations (Tables 2 and 3). This was expected, as the holo conformation of the binding site is already prepared to accept the agonists. The FITTED prerequisite holo docking calculations yielded an AA_{match} of 100% in the cases of all three agonists. The d_{best} values of the prerequisite docking calculations were under 4.0 Å in all cases, with the only exception of bodipy-iodoacetamide holo docking (Table 3). In the case of AutoDock, the d_{best} values of the holo prerequisite calculations were below 7.0 Å, and slightly above it in the apo prerequisite docking runs. The $d_{best} \geq 7.0$ Å values are due to head-to-tail binding mode of the agonists (Tables 2 and 3).

The comparison of the prerequisite binding modes produced by FITTED on both the holo and apo docking calculations of the three compounds resulted in C621 as the only common binding site amino acid for all three agonists. All holo prerequisite docking calculations found C621 with only the exception of bodipy-iodoacetamide prerequisite holo docking with AutoDock. These findings suggest that different prerequisite binding modes (Tables S5–S7) might result in good final covalently binding positions (Table 1). By investigating more agonists with two docking programs, one might expect to discover common amino acids that indicate a larger prerequisite binding area. If an agonist at least partially interacts with the prerequisite binding area it has a chance to find its way to the final binding pocket.

Amino acids C665, P666 and F669 of the A-loop are part of the binding pocket of bodipy-iodoacetamide, and also of most prerequisite binding modes of bodipy-iodoacetamide and JT010 (Tables S5–S7) found by both programs. C665 is also highlighted in the literature [7] as an important amino acid both in agonist binding and receptor activation. The absence of interaction of BITC with the above-mentioned amino acids might be due to the smaller size of BITC compared to the other two agonists (Figure 1). These results suggest a previously unexplored structural role of the amino acids P666 and F669 co-operating with C665 in agonist binding and consequent flipping of the A-loop, leading to conformational changes and receptor activation. These findings were also strengthened by virtual mutation and docking (Figure S1). However, BITC interacts with another part of the A-loop as prerequisite binding site found by apo docking with AutoDock. This BITC site was also sufficient to open the binding pocket in molecular dynamics (MD) simulations (Section 2.3).

Although not an accepted medicine, JT010 has a remarkable EC_{50} of 0.65 nM [4]. Thus, in novel drug design, JT010 can be regarded as a reference point, and therefore, it was further investigated from the mechanism viewpoint of this Section. During the transition from a non-covalent, prerequisite binding mode (d = 3.6 Å) to the final, covalent binding mode of JT010, its interactions with F669 and Y680 diminish (Table S4), and new interactions with K620, I623 and E625 are formed (cut-off distance of interaction of 3.5 Å for heavy atom–heavy atom distance). Interactions with the two binding site cysteines, C621 and C665 [7] are observed for both non-covalent and covalent binding modes of JT010. As it was highlighted in the previous section, F669 is part of the A-loop and has a possible role in the flipping of the loop during agonist binding to the TRPA1 receptor, based on the example of JT010. It can be hypothesized, that interaction with F669 is only important in the initial prerequisite binding of the agonist, later during covalent bond formation this interaction diminishes, and the agonist penetrates deeper into the binding site, interacting with amino acid residues that are in close proximity of C621. This observation is strengthened by MD simulations also (see Section 2.3). The covalent ΔG_{FD} of JT010 almost doubles (and consequently its EI also), compared to that of the prerequisite binding mode.

Regardless of their potency, all three agonists can activate the TRPA1 ion channel, however, to prevent xenobiotic overload of the body it is advisable to administer the lowest possible dose of a drug, which is only effective if the agent is highly potent. If using JT010 as a reference for future studies, the following limit values can be concluded for the selection of potent agonists. A prerequisite EI value of at least 2 kcal/mol, and a prerequisite $d_{best} \leq 4.0$ Å, and a prerequisite ΔG_{FD} of at least -35 kcal/mol forecast a strong agonist. Notably, the ΔG_{FD} value of JT010 is approximated by that of BITC, and the EI of BITC even surpasses that of JT010 (Table 1). bodipy-iodoacetamide somewhat lags behind JT010 and BITC. These findings are in good agreement with the literature, as the EC_{50} value of iodoacetamide (without the bodipy label) is 357 µM [28], which is substantially larger, than that of JT010. The EC_{50} of allyl-isothiocyanate (a similar compound to BITC) is 37 nM [7], which is also in the nanomolar range, as the EC_{50} of JT010.

The docking performance of FITTED slightly outperformed AutoDock as seen in Tables 2 and 3. However, FITTED requires a probe previously placed within the binding site to select the binding site amino acids, which obviously helps the search. At the same time, AutoDock did not require such information, and an unrestricted search could be performed for the prerequisite binding mode within the docking box. Thus, we decided to use the prerequisite binding mode of BITC found by AutoDock apo calculation with A-loop (Table S3) for further MD simulations in the next, Section 3.

2.3. Ligand Migration Dynamics Connecting Prerequisite and Final Binding Modes

MD simulations (100 ns, unrestrained, with explicit waters and simulated annealing protocol as described in Section 4) were performed on both apo and holo forms of TRPA1 (Table 4) to further explore the binding dynamics of the agonist BITC of the best EI value (Table 1). As the results of the previous Section indicated that the prerequisite binding modes affect A-loop, we were particularly interested in the structural changes of the loop, and the communication between the distinct prerequisite and final binding modes. An MD_{apo} simulation was used as a reference, to observe if there are any changes in the conformation of the A-loop in the absence of the agonist. Then, two MD simulations were started from two prerequisite binding modes of BITC on the apo TRPA1 found in the previous Section, one of them interacting with the loop (MD_{rank1}). Finally, an MD simulation was started from the experimental binding position of BITC (MD_{holo}), with the covalent bond cut and the geometry of the N=C=S bonds restored. In the MD_{holo} simulation, an unbinding-binding event occurred and the A-loop remained stable throughout the simulation (Figure 4). The interaction with the original five amino acids of the TRPA1 pocket gradually diminished and appeared once again in a very short time interval of the first 0.7 ns (Figure 4). During the entire MD_{apo} (Table 4) simulation, no significant changes were observed in the conformation of the A-loop, while, in the case of MD_{rank1}, the loop moved upward (the red and blue arrows show the movement of A-loop and the teal arrows the movement of BITC on Figure 5). In the starting position of MD_{rank1}, BITC interacted with the loop and was positioned beneath it (marked with 0 ns at Figure 5). After a very short time (0.3 ns) the ligand dissociates from the TRPA1 surface, dragging down the loop with itself. After 17.6 ns the loop moved upwards, approximating the open position of the binding site which is present in the holo structure (Figures 2 and 5). Finally, (at 38 ns) BITC finds its way back into the binding pocket and resides there for 2 ns until dissociation. At the same time, in the case of MD_{rank3} (Table 4) in which the docked position of BITC did not interact with the A-loop, BITC dissociated after 1.6 ns and afterwards, no changes were observed in the structure of the loop. Thus, the above results showed how the binding of an agonist to the A-loop induces its motion towards opening the binding pocket and allowing the entrance of the agonists.

Table 4. The details of the MD simulations performed to unravel the binding mechanism of BITC.

Simulation Name	TRPA1	Ligand	Change in A-Loop	Movement of the Agonist
MD_{apo}	Apo protein	-	No change in A-loop conformation	-
$MD_{holo,PSA}$	Holo protein	Experimental	No change in A-loop conformation	Unbinding–binding
MD_{rank1}	Apo protein	Rank 1 docked ligand binding mode	A-loop flipping to the active conformation	Dissociation–association
MD_{rank3}	Apo protein	Rank 3 docked ligand binding mode	No change in A-loop conformation	Dissociation

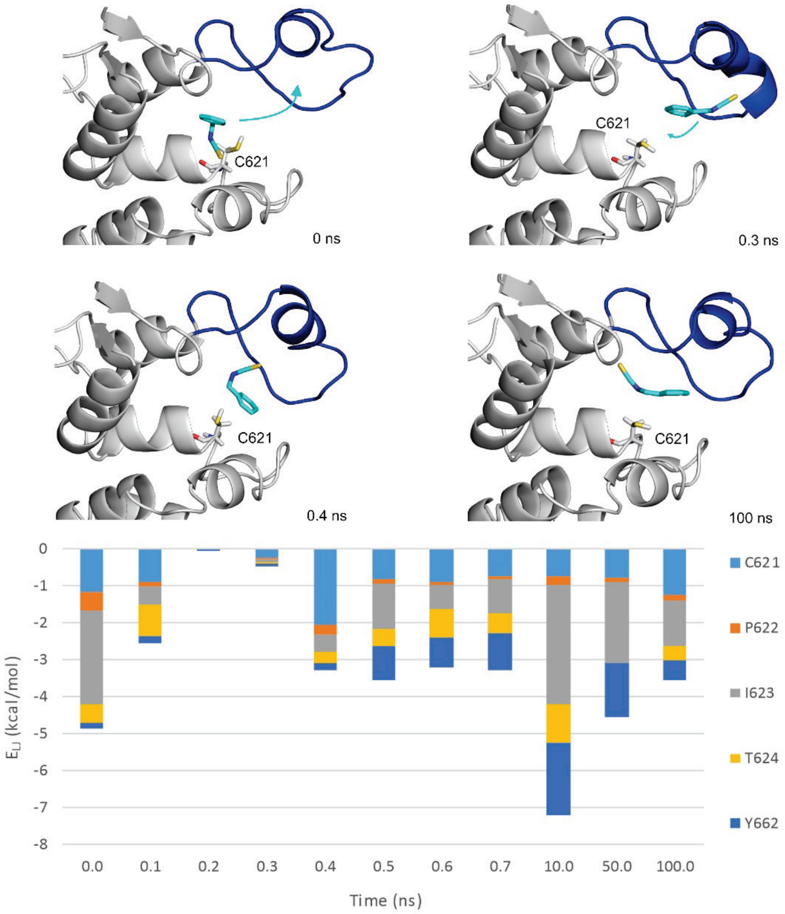


Figure 4. The MD_{holo} simulation starting from the experimental binding position, with the covalent bond cut. Interaction energy distribution of the interacting amino acids of the pocket (inner prerequisite site) is shown during the MD simulation. The structural figures are snapshots of the binding position of BITC at the stated time frame of the MD simulation. The protein is shown in grey cartoon and BITC with teal sticks. The A-loop is marked with blue in the open conformation. C621 amino acid is also shown as all atom sticks representation. The teal arrows indicate the movement of BITC. Lennard–Jones interaction energies calculated between BITC and the TRPA1 target amino acids are shown per residue on the diagram.

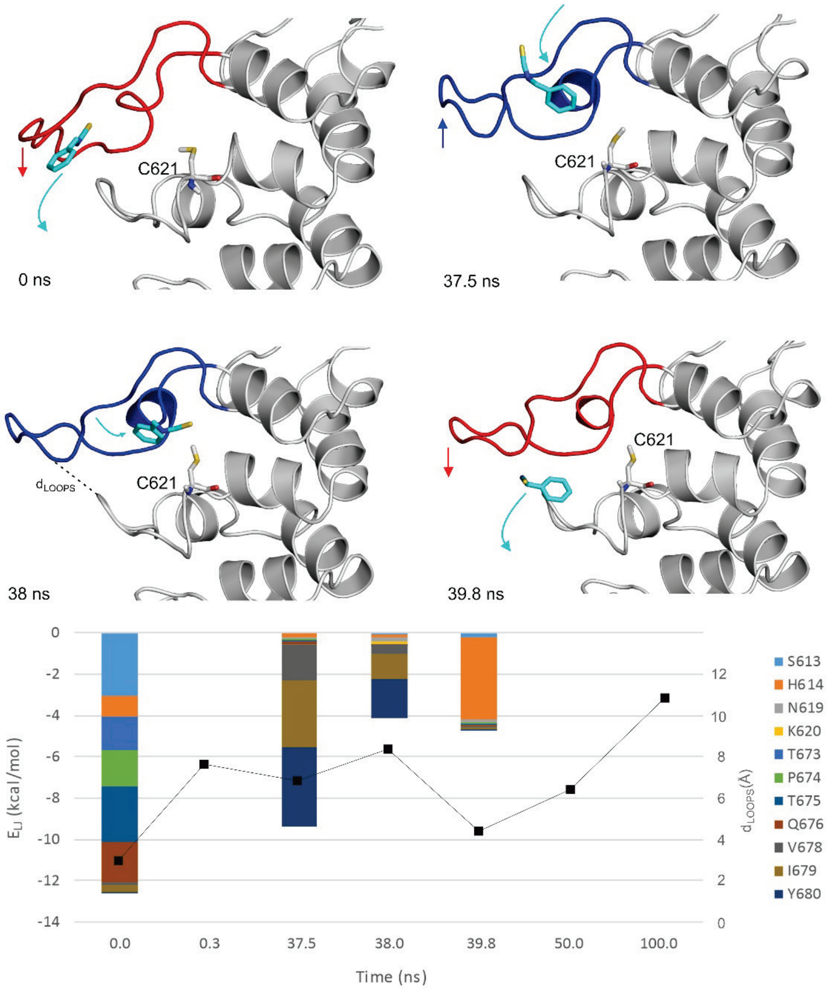


Figure 5. The MD$_{rank1}$ simulation starting from the top 1st docked prerequisite binding mode of BITC on the apo TRPA1. Interaction energy distribution of the interacting amino acids of the outer prerequisite site are shown during the MD simulation. The loop motion is quantified by the distances between the N atom of N615 of the A-loop and O atom of Q676 of the opposite loop (black dotted line on the t = 38 ns structural plot, d$_{LOOPS}$) shown as joint black boxes. The structural figures are snapshots of the binding position of BITC at the stated time frame of the MD simulation. The protein is shown in grey cartoon and BITC with teal sticks. The A-loop is marked with blue and red in open and closed conformations, respectively. C621 amino acid is also shown as all atom sticks representation. The red and blue arrows show the movement of the A-loop and the teal arrows the movement of BITC. Lennard–Jones interaction energies calculated between BITC and the TRPA1 target amino acids are shown per residue on the diagram.

The interaction of the prerequisite binding mode with the P669 amino acid side chain (first mentioned in the previous Section) was observed in the starting frame of MD$_{rank1}$ and was diminished both upon dissociation and the penetration of BITC towards its final binding pocket. BITC in the prerequisite binding mode forms mainly polar interactions with amino acids of the A-loop and the other loop (Figure 5), such as S613, P674, T675 and Q676. Towards the final binding mode, however, BITC interacts with hydrophobic amino acids such as V678, I679 and Y680. This latter observation is strengthened by the MD$_{holo}$

run (Figure 4), where interactions with I623 and Y662 dominate. The distances between the N atom of N615 of the A-loop and the O atom of Q676 of the opposite loop (Figure 5) appear to be good indicators for A-loop opening and closing.

3. Materials and Methods

3.1. Preparation of the Ligand Structures

Ligand conformations were obtained from their respective atomic coordinate structure files of TRPA1 receptor-ligand complexes. The ligands were modified to regain their original structure, before the covalent interaction. To JT010 [4] a chlorine was added. In the case of benzylisothiocyanate (BITC, [6]) the proper bond orders and hybridization states were restored, as R-N=C=S, the hybridization states of N and S were set to sp2 and that of C to sp. Finally, in the case of bodipy-iodoacetamide [7] an iodine was added. These modifications were carried out using the builder function of PyMol (Schrödinger, New York, NY, USA) [29]. All the ligands were energy minimized by a quantum chemistry program package, MOPAC [30,31] with PM7 parametrization [31]. Hydrogens and Gasteiger–Marsili [32] partial charges were added by OpenBabel [33]. In the case of FITTED program package [13–15,34], the built in preparation steps were used with default settings. For BITC the molecular mechanics force field parameters were obtained from the general AMBER force field (GAFF) [35]. The ligand was built in Maestro [36], then semi-empirical quantum mechanics optimization was performed with MOPAC [30,31] using PM7 parametrization [31], with the gradient norm set to 0.001. After energy minimization, a further force calculation step was included, the force constant matrices were positive definite. The RED-vIII.52 [37] software was used for restrained electrostatic potential (RESP) charge calculations, using RESP-A1B fitting (compatible with the AMBER99SB-ILDN force field) after ab inito geometry optimization by GAMESS [38]. Acpype [39] was used to assign atomtypes, bond and angle parameters for topology of ligand. The missing bond stretching, angle bending and torsional parameters were calculated by the antechamber [39,40] and parmchk utilities of AmberTools program package similarly as described in [41]. Torsional parameters for R-N=C=S moiety were manually added.

3.2. Target Preparation

The atomic coordinate file of the ligand free TRPA1 receptor was obtained from the Protein Data Bank (PDB, [42]), under the accession code 6V9W [7]. As the four chains of the target are symmetrical (homotetramer), only one chain was used to reduce computational costs. The amino acids of a chain do not interfere with the binding of the ligand to another chain. The missing atoms and residues [43] were rebuilt using SWISS MODEL [44], and energy minimized with GROMACS [45]. The convergence threshold of the steepest descent optimization was set to 10^3 kJ mol^{-1} nm^{-1}, and that of the conjugate gradient optimization to 10 kJ mol^{-1} nm^{-1}. AMBER99SB-ILDN force field [35] was used for the calculation, and a position restraint at a force constant of 10^3 kJ mol^{-1} nm^{-2} was applied on heavy atoms. The targets were further optimized by ProCESS tool of FITTED, with the original settings [13]. In the case of AutoDock (The Scripps Research Institute, La Jolla, CA, USA), the added H atoms and partial charges were kept from energy minimization.

3.3. Covalent Docking with FITTED

Covalent docking calculations were carried out using FITTED [13–15,34]. The covalent residue (C621) and adjacent basic residue (P622) were adjusted in the graphical user interface of the program. Root mean squared deviation (*RMSD*) values were calculated between the crystallographic and representative ligand conformations, if available. All other settings were used as the default of the program. In the PREPARE step of the program, the binding site interacting amino acids were identified by leaving the crystallographic ligand in the structure, which was then removed after this step. The non-covalent docking was performed similarly, with the exception, that the covalent mode of the program was switched off.

3.4. Prerequisite Docking with AutoDock 4.2

Prerequisite binding calculations were performed by AutoDock [25,46–49]. The number of grid points was set to 60 × 60 × 60 with a 0.375 Å grid spacing. Lamarckian genetic algorithm was used, flexibility on all active torsions was allowed on the ligands. Ten docking runs were performed for all ligands, and the resulting ligand conformations were ranked based on their calculated free energy of binding values. The binding mode with the most favorable calculated energy of binding was ranked in the lowest rank.

3.5. Molecular Dynamics Simulations

The apo TRPA1 and dry docked complexes of BITC were subject to a two step energy minimization, including steepest descent and conjugate gradient algorithms as described in "Target preparation". After energy minimization, the apo and dry docked complexes were subject to 100-ns-long MD simulations. The simulation box was filled up with explicit TIP3P [50] water molecules, and to neutralize the systems, counter-ions (sodium or chloride) were added. The maximum step size of the steepest descent algorithm was 0.5 nm, and that of the conjugate gradient algorithm was 0.05 nm. The exit tolerance level of the steepest descent algorithm was set to and 10^3 that of the conjugate gradient algorithm to 10 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-1}}$. Movement of the solute Cα atoms were restrained at a force constant of 10^3 $\mathrm{kJmol^{-1}nm^{-2}}$, except for that of the A-loop. Calculations were performed with programs of the GROMACS [45] software package, using the AMBER99SB-ILDN [35] force field. After energy minimization, 100-ns-long NPSA MD simulation was carried out with a time step of 2 fs. Simulated annealing temperature scheme was applied as described in [22]. Simulated annealing temperature was rescaled and controlled for both solvent and solute. The temperature was gradually increased up to 323.15 K, then lowered back to 300 K in the first 20 ns, then the simulation was continued to 100 ns with constant temperature of 300 K. Pressure was coupled by the Parrinello–Rahman algorithm and a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} $\mathrm{bar^{-1}}$ and reference pressure of 1 bar. Particle Mesh-Ewald summation was used for long range electrostatics. Van der Waals and Coulomb interactions had a cut-off at 11 Å. Coordinates were saved at regular time-intervals of 1 ps yielding 1×10^3 frames. Periodic boundary conditions were treated before analysis to center whole and recovered hydrated solute structures in the box. The original protein structure served as the basis of Cα fitting.

3.6. Scoring

AutoDock [25] estimates the binding free energy of the ligand (ΔG_{AD}) with Equation (1) as a scoring function.

$$\Delta G_{AD} = W_{vdW} \sum_{ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbound} \sum_{ij} E_{(t)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{ij} \frac{q_i q_j}{\varepsilon_{(rij)_{rij}}} + W_{sol} \sum_{ij} (S_i V_j + S_j V_i) e^{\left(-\frac{r_{ij}^2}{2\partial^2}\right)} \tag{1}$$

W terms are weighting constants calibrating to an experimentally determined set of free energies. Ligand atoms are represented by *i*, and protein atoms by *j*. A Lennard–Jones 12-6 dispersion/repulsion term, a directional 12–10 h-bonding term and a screened Coulombic electrostatic potential are included. *A* and *B* parameters are taken from the Amber force field. *E*(*t*) is a directional weight based on the angle, *t*, between the probe and the target atom. *C* and *D* parameters are assigned for well-depth calculations. The final term is a desolvation potential, *V* is the volume of the atoms surrounding a given atom and *S* is a solvation parameter for weighting [51]. δ is a distance weighting factor. The actual distance between the ith (ligand) and jth (target) atoms is marked with *r*.

The FITTED [13–15,34] scoring function estimates (ΔG_{FD}) with the sum of various terms including the number of rotatable bonds, van der Waals and electrostatic interactions and directional H-bonding contributions as described in Equation (2).

$$\Delta G_{FD} = \Delta G_0 + 0.14 N_{rot} + \sum (scale\ factor) \left[\left(0.26\ U_{vdW}^{inh-prot} + 0.035\ U_{elec}^{inh-prot} + 0.80 f_{hb}(\Delta r, \Delta\alpha) \right) \right] \tag{2}$$

N_{rot} is the number of rotatable bonds, U_{vdw} and U_{elec} are the van der Waals and electrostatic interactions based on the AMBER94 force field. The last term is the solvation contribution to the free energy of binding. Where f_{hb} is the electrical field strength of hydrogen bonds, r is the length and α is the angle of hydrogen bonds.

$$E_{ij} = \sum_{ij}^{N_I N_L} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] \\ A_{ij} = \varepsilon_{ij} R_{ij}^{12}; B_{ij} = 2\varepsilon_{ij} R_{ij}^{6}; R_{ij} = R_i + R_j; \varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \tag{3}$$

where *εi* and *εj* are the potential well depths in the equilibrium distance of atom pairs of identical types; *εij* is the potential well depth in equilibrium between the *i*th (ligand) and *j*th (target) atoms; *Rij* is the internuclear distance at equilibrium between *i*th (ligand) and *j*th (target) atoms; *Ri* and *Rj* are half equilibrium distances between *ii* and *jj* atom pairs of identical types, respectively; *rij* is the actual distance between the *i*th (ligand) and *j*th (target) atoms; N_T is the number of target atoms; N_L is the number of ligand atoms.

3.7. Ranking

The basis of the structural clustering and ranking of the docked ligand conformations was their AutoDock 4.2 binding free energy values. In the respective Tables, the serial number of ranks are represented. To create one rank [41], the ligand structure with the lowest calculated free energy of binding, and its neighboring docked ligand structures within 2 Å [52] were selected. Then new ranks were opened for the remaining structures, and clustering was repeated with the same protocol. The low serial number of a rank indicates an energetically favorable binding conformation. The actual rank (N) selected from all the ranks (M) is given in the format N/M.

RMSD. In all cases, the structural match of the docked (***D*** in Equation (4)) binding mode to the crystallographic reference (***C***) was expressed as a root mean squared deviation (*RMSD*) value according to Equation (1).

$$RMSD = \sqrt{\frac{1}{N} \sum_{n=1}^{N} |D_n - C_n|^2} \tag{4}$$

In Equation (4), N is the number of ligand heavy atoms, ***C*** is the space vector of the nth heavy atom of the crystallographic reference ligand molecule, ***D*** is the space vector of the nth heavy atom of the calculated ligand conformation. $RMSD_{\mathrm{best}}$ is the *RMSD* value of the ligand binding mode with the lowest *RMSD*.

The distance (d) between the S atom of C621 amino acid and the ligand atom that participates in the covalent binding was also measured to check the presence of covalent bond and to estimate the degree of translation necessary to move the prerequisite binding mode into the covalent binding mode (Figure 3). The $\mathrm{d_{best}}$ value is the smallest distance observed.

The $\mathrm{AA_{match}}$ (%) is the rate of identical AAs present in two different binding pockets interacting with the ligand in a 3.5 Å cut-off distance. It is calculated by the results of Tables S5–S7.

NHA Number of heavy atoms of the agonist counts all the atoms except for hydrogens.

$\mathrm{EI_{NHA}}$ Efficiency index, the calculated free energy of binding is divided by the NHA of the respective agonist. The dimension is kcal/mol.

$\mathrm{d_{covalent}}$ The length of the covalent bond in Å.

$\mathrm{Rank_{best}}$ The scoring function of the program collects the results into ranks based on their calculated free energy of binding. The lowest rank contains the best energy. The rank that contains the model with the best *RMSD* value is the $\mathrm{Rank_{best}}$.

4. Conclusions

Agonist binding to TRPA1 is a dynamic process involving structural changes of the target, first at the smaller scale of the A-loop, necessary for binding site activation, then at the whole of the target, required for channel activation. The present study identified the prerequisite binding modes of three agonists and showed how the binding of a ligand to the prerequisite site can forecast its successful docking to the final binding pocket. The prerequisite binding sites proved to be milestones on the association/dissociation pathway of the agonists, important in mechanism-based design. The present study also showed how the prerequisite binding modes affect the opening of the A-loop region, a central scene of the agonist binding mechanism. The time step measured in nanoseconds necessary for binding site activation is currently hidden from experimental methods, and only the co-operation with in silico approaches can shed light on them. Thus, amino acids identified along the dynamic binding pathway will serve as new target sites for the design of reversible binding of future agonists, beyond the well-known target of the covalent binding pocket of TRPA1.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ph14100988/s1. Figure S1: The close-up of binding of BITC to TRPA1 mutant structure, Table S1: Covalent docking calculations performed by FITTED on the apo target with Aloop, Table S2: Non-covalent docking calculations performed by FITTED on the apo target with A-loop, Table S3: Non-covalent docking calculations performed by AutoDock on the apo target with A-loop, Table S4: Interacting (≤3.5 Å) amino acid residues of 6PQO with the non-covalent top ranked binding mode of JT010 (FITTED), Table S5: The interacting amino acids (within 3.5 Å) of the experimental binding position, the covalently docked and prerequisite binding modes of bodipy-iodoacetamide, Table S6: The interacting amino acids (within 3.5 Å) of the experimental binding position, the covalently docked and prerequisite binding modes of BITC, Table S7: The interacting amino acids (within 3.5 Å) of the experimental binding position, the covalently docked and prerequisite binding modes of JT010.

Author Contributions: Conceptualization, B.Z.Z., C.H. and E.P.; methodology, B.Z.Z., C.H. and R.B.; formal analysis, C.H.; investigation, B.Z.Z., R.B. and C.H.; resources, C.H. and E.P.; writing—original draft preparation, B.Z.Z. and C.H.; writing—review and editing, C.H. and E.P.; visualization, B.Z.Z., C.H.; supervision, C.H.; project administration, C.H.; funding acquisition, C.H. and E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Hungarian National Research, Development and Innovation Office (K123836, K134214), 2017-1.2.1 NKP-2017-00002 (NAP-2; Chronic Pain Research Group), EFOP-3.6.1.-16-2016-0004 and GINOP 2.3.2-15-2016-00050 "PEPSYS". This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The work was supported by the ÚNKP-20-5 and ÚNKP-20-3-I. New National Excellence Program of the Ministry for Innovation and Technology. Supported by the PTE ÁOK-KA No:2019/KA-2019-31. The project was supported by the European Union, co-financed by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004. This project is supported by Eötvös Loránd Research Network.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article and Supplementary Material.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Logu, F.; Nassini, R.; Materazzi, S.; Carvalho, G.M.; Nosi, D.; Rossi, D.D.; Marone, I.M.; Ferreira, J.; Li Puma, S.; Benemei, S.; et al. Schwann cell TRPA1 mediates neuroinflammation that sustains macrophage-dependent neuropathic pain in mice. *Nat. Commun.* **2017**, *8*, 1–16. [CrossRef] [PubMed]
2. Paulsen, C.E.; Armache, J.P.; Gao, Y.; Cheng, Y.; Julius, D. Structure of the TRPA1 ion channel suggests regulatory mechanisms. *Nature* **2015**, *520*, 511–517. [CrossRef] [PubMed]
3. Liu, C.; Reese, R.; Vu, S.; Rougé, L.; Shields, S.D.; Kakiuchi-Kiyota, S.; Chen, H.; Johnson, K.; Shi, Y.P.; Chernov-Rogan, T.; et al. A Non-covalent Ligand Reveals Biased Agonism of the TRPA1 Ion Channel. *Neuron* **2020**, *109*, 273–284. [CrossRef] [PubMed]

4. Takaya, J.; Mio, K.; Shiraishi, T.; Kurokawa, T.; Otsuka, S.; Mori, Y.; Uesugi, M. A Potent and Site-Selective Agonist of TRPA1. *J. Am. Chem. Soc.* **2015**, *137*, 15859–15864. [CrossRef]
5. Pozsgai, G.; Bátai, I.Z.; Pintér, E. Effects of sulfide and polysulfides transmitted by direct or signal transduction-mediated activation of TRPA1 channels. *Br. J. Pharmacol.* **2019**, *176*, 628–645. [CrossRef] [PubMed]
6. Suo, Y.; Wang, Z.; Zubcevic, L.; Hsu, A.L.; He, Q.; Borgnia, M.J.; Ji, R.R.; Lee, S.Y. Structural Insights into Electrophile Irritant Sensing by the Human TRPA1 Channel. *Neuron* **2020**, *105*, 882–894. [CrossRef]
7. Zhao, J.; Lin King, J.V.; Paulsen, C.E.; Cheng, Y.; Julius, D. Irritant-evoked activation and calcium modulation of the TRPA1 receptor. *Nature* **2020**, *585*, 141–145. [CrossRef]
8. Chernov-Rogan, T.; Gianti, E.; Liu, C.; Villemure, E.; Cridland, A.P.; Hu, X.; Ballini, E.; Lange, W.; Deisemann, H.; Li, T.; et al. TRPA1 modulation by piperidine carboxamides suggests an evolutionarily conserved binding site and gating mechanism. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 26008–26019. [CrossRef] [PubMed]
9. Tseng, W.C.; Pryde, D.C.; Yoger, K.E.; Padilla, K.M.; Antonio, B.M.; Han, S.; Shanmugasundaram, V.; Gerlach, A.C. TRPA1 ankyrin repeat six interacts with a small molecule inhibitor chemotype. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 12301–12306. [CrossRef]
10. Wang, Y.Y.; Chang, R.B.; Waters, H.N.; McKemy, D.D.; Liman, E.R. The nociceptor ion channel TRPA1 is potentiated and inactivated by permeating calcium ions. *J. Biol. Chem.* **2008**, *283*, 32691–32703. [CrossRef] [PubMed]
11. Ábrányi-Balogh, P.; Petri, L.; Imre, T.; Szijj, P.; Scarpino, A.; Hrast, M.; Mitrović, A.; Fonovič, U.P.; Németh, K.; Barreteau, H.; et al. A road map for prioritizing warheads for cysteine targeting covalent inhibitors. *Eur. J. Med. Chem.* **2018**, *160*, 94–107. [CrossRef] [PubMed]
12. Petri, L.; Egyed, A.; Bajusz, D.; Imre, T.; Hetényi, A.; Martinek, T.; Ábrányi-Balogh, P.; Keserű, G.M. An electrophilic warhead library for mapping the reactivity and accessibility of tractable cysteines in protein kinases. *Eur. J. Med. Chem.* **2020**, *207*, 1–9. [CrossRef] [PubMed]
13. Corbeil, C.R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449. [CrossRef]
14. Pottel, J.; Therrien, E.; Gleason, J.L.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 6. Development and application to the docking of HDACs and other zinc metalloenzymes inhibitors. *J. Chem. Inf. Model.* **2014**, *54*, 254–265. [CrossRef] [PubMed]
15. Therrien, E.; Englebienne, P.; Arrowsmith, A.G.; Mendoza-Sanchez, R.; Corbeil, C.R.; Weill, N.; Campagna-Slater, V.; Moitessier, N. Integrating medicinal chemistry, organic/combinatorial chemistry, and computational chemistry for the discovery of selective estrogen receptor modulatorswith FORECASTER, a novel platform for drug discovery. *J. Chem. Inf. Model.* **2012**, *52*, 210–224. [CrossRef]
16. Bálint, M.; Horváth, I.; Mészáros, N.; Hetényi, C. Towards Unraveling the Histone Code by Fragment Blind Docking. *Int. J. Mol. Sci.* **2019**, *20*, 422. [CrossRef]
17. Kevener, H.E.; Zhao, W.; Ball, D.M.; Babaoglu, K.; Qi, J.; White, S.W.; Lee, R.E. Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. *J. Chem. Inf. Model.* **2009**, *49*, 444–460. [CrossRef] [PubMed]
18. Castro-Alvarez, A.; Costa, A.M.; Vilarrasa, J. The Performance of several docking programs at reproducing protein-macrolide-like crystal structures. *Molecules* **2017**, *22*, 136. [CrossRef]
19. Mena-Ulecia, K.; Tiznado, W.; Caballero, J. Study of the differential activity of thrombin inhibitors using docking, QSAR, molecular dynamics, and MM-GBSA. *PLoS ONE* **2015**, *10*, 1–21.
20. Ramírez, D.; Caballero, J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* **2018**, *23*, 1038. [CrossRef]
21. Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356. [CrossRef]
22. Bálint, M.; Jeszenoi, N.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Systematic exploration of multiple drug binding sites. *J. Cheminform.* **2017**, *9*, 65–79. [CrossRef]
23. Sotriffer, C. Docking of Covalent Ligands: Challenges and Approaches. *Mol. Inform.* **2018**, *37*, 1–12. [CrossRef]
24. Kumalo, H.M.; Bhakat, S.; Soliman, M.E.S. Theory and applications of covalent docking in drug discovery: Merits and pitfalls. *Molecules* **2015**, *20*, 1984–2000. [CrossRef]
25. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated docking using a Lamarckian Genetic Algorithm and empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662. [CrossRef]
26. Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. An extensible and systematic force field, ESFF, for molecular modeling of organic, inorganic, and organometallic systems. *J. Comput. Chem.* **2003**, *24*, 1059–1076. [CrossRef] [PubMed]
27. *CDiscoVer, 98.0*; Accelrys, Inc.: San Diego, CA, USA, 2001.
28. Macpherson, L.J.; Dubin, A.E.; Evans, M.J.; Marr, F.; Schultz, P.G.; Cravatt, B.F.; Patapoutian, A. Noxious compounds activate TRPA1 ion channels through covalent modification of cysteines. *Nature* **2007**, *445*, 541–545. [CrossRef] [PubMed]
29. Warren, L.D. *The PyMOL Molecular Graphics System*; Version 2.0; Schrödinger, LLC.: New York, NY, USA, 2002.
30. Stewart, J.J.P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213. [CrossRef]
31. Stewart, J.J.P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1–32. [CrossRef]

32. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228. [CrossRef]
33. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel. *J. Cheminform.* **2011**, *3*, 1–14.
34. Moitessier, N.; Therrien, E.; Hanessian, S. A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic β-secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885–5894. [CrossRef]
35. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef]
36. *Schrödinger Release 2020-4: Maestro*; Schrödinger, LLC.: New York, NY, USA, 2021.
37. Dupradeau, F.Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. tools: Advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821–7839. [CrossRef] [PubMed]
38. Schmidt, M.W.; Baldridge, K.K.; Boatz, J.A.; Elbert, S.T.; Gordon, M.S.; Jensen, J.H.; Koseki, S.; Matsunaga, N.; Nguyen, K.A.; Su, S.; et al. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347–1363. [CrossRef]
39. Sousa Da Silva, A.W.; Vranken, W.F. ACPYPE—AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 1–8. [CrossRef]
40. Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260. [CrossRef]
41. Zsidó, B.Z.; Börzsei, R.; Szél, V.; Hetényi, C. Determination of Ligand Binding Modes in Hydrated Viral Ion Channels to Foster Drug Design and Repositioning. *J. Chem. Inf. Model.* **2021**, *8*, 4011–4022. [CrossRef]
42. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef]
43. Zsidó, B.Z.; Hetényi, C. Molecular structure, binding affinity, and biological activity in the epigenome. *Int. J. Mol. Sci.* **2020**, *21*, 4143. [CrossRef]
44. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; De Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. [CrossRef]
45. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.E.; Berendsen, H.J.C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718. [CrossRef]
46. Fliszár-Nyúl, E.; Faisal, Z.; Mohos, V.; Derdák, D.; Lemli, B.; Kálai, T.; Sár, C.; Zsidó, B.Z.; Hetényi, C.; Horváth, Á.I.; et al. Interaction of SZV 1287, a novel oxime analgesic drug candidate, and its metabolites with serum albumin. *J. Mol. Liq.* **2021**, *333*, 1–10. [CrossRef]
47. Zsidó, B.Z.; Balog, M.; Erős, N.; Poór, M.; Mohos, V.; Fliszár-Nyúl, E.; Hetényi, C.; Nagane, M.; Hideg, K.; Kálai, T.; et al. Synthesis of spin-labelled bergamottin: A potent CYP3A4 inhibitor with antiproliferative activity. *Int. J. Mol. Sci.* **2020**, *21*, 508. [CrossRef] [PubMed]
48. Mohos, V.; Fliszár-Nyúl, E.; Ungvári, O.; Bakos, É.; Kuffa, K.; Bencsik, T.; Zsidó, B.Z.; Hetényi, C.; Telbisz, Á.; Özvegy-Laczka, C.; et al. Effects of chrysin and its major conjugated metabolites chrysin-7-sulfate and chrysin-7-glucuronide on cytochrome P450 enzymes and on OATP, P-gp, BCRP, and MRP2 transporters. *Drug Metab. Dispos.* **2020**, *48*, 1064–1073. [CrossRef] [PubMed]
49. Mohos, V.; Fliszár-Nyúl, E.; Lemli, B.; Zsidó, B.Z.; Hetényi, C.; Mladěnka, P.; Horký, P.; Pour, M.; Poór, M. Testing the pharmacokinetic interactions of 24 colonic flavonoid metabolites with human serum albumin and cytochrome P450 enzymes. *Biomolecules* **2020**, *10*, 409. [CrossRef] [PubMed]
50. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
51. Huey, R.; Morris, G.M.; Olson, A.J.; Goodsell, D.S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–1152. [CrossRef] [PubMed]
52. Hetényi, C.; Van Der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett.* **2006**, *580*, 1447–1450. [CrossRef]

D7

International Journal of
Molecular Sciences

Article

Binding Networks Identify Targetable Protein Pockets for Mechanism-Based Drug Design

Mónika Bálint [1,†], Balázs Zoltán Zsidó [1,†], David van der Spoel [2] and Csaba Hetényi [1,*]

[1] Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12., 7624 Pécs, Hungary; monibalint18@gmail.com (M.B.); zsido.balazs@pte.hu (B.Z.Z.)
[2] Department of Cell and Molecular Biology, Uppsala University, Box 596, SE-75124 Uppsala, Sweden; david.vanderspoel@icm.uu.se
[*] Correspondence: hetenyi.csaba@pte.hu
[†] These authors contributed equally to this work.

Abstract: The human genome codes only a few thousand druggable proteins, mainly receptors and enzymes. While this pool of available drug targets is limited, there is an untapped potential for discovering new drug-binding mechanisms and modes. For example, enzymes with long binding cavities offer numerous prerequisite binding sites that may be visited by an inhibitor during migration from a bulk solution to the destination site. Drug design can use these prerequisite sites as new structural targets. However, identifying these ephemeral sites is challenging. Here, we introduce a new method called NetBinder for the systematic identification and classification of prerequisite binding sites at atomic resolution. NetBinder is based on atomistic simulations of the full inhibitor binding process and provides a networking framework on which to select the most important binding modes and uncover the entire binding mechanism, including previously undiscovered events. NetBinder was validated by a study of the binding mechanism of blebbistatin (a potent inhibitor) to myosin 2 (a promising target for cancer chemotherapy). Myosin 2 is a good test enzyme because, like other potential targets, it has a long internal binding cavity that provides blebbistatin with numerous potential prerequisite binding sites. The mechanism proposed by NetBinder of myosin 2 structural changes during blebbistatin binding shows excellent agreement with experimentally determined binding sites and structural changes. While NetBinder was tested on myosin 2, it may easily be adopted to other proteins with long internal cavities, such as G-protein-coupled receptors or ion channels, the most popular current drug targets. NetBinder provides a new paradigm for drug design by a network-based elucidation of binding mechanisms at an atomic resolution.

Keywords: ligand; mechanism; pathway; dynamics; channel

check for updates

Citation: Bálint, M.; Zsidó, B.Z.; van der Spoel, D.; Hetényi, C. Binding Networks Identify Targetable Protein Pockets for Mechanism-Based Drug Design. *Int. J. Mol. Sci.* **2022**, *23*, 7313. https://doi.org/10.3390/ijms23137313

Academic Editor: Edmond Dik Lung Ma

Received: 10 June 2022
Accepted: 28 June 2022
Published: 30 June 2022

1. Introduction

Uncovering the mechanism(s) by which a drug binds to its target is of primary importance in drug design. To date, established experimental methods such as X-ray crystallography [1,2] and cryo-electron microscopy [3–5] have been used to capture the atomic resolution structure of a drug bound with its target (called the binding mode). However, these techniques usually do not supply the entire binding mechanism or the intermediate interactions required (the prerequisite binding modes, or PMs), which are often difficult to capture experimentally [6]. Detecting intermediates is especially difficult with targets such as myosin 2 that have long binding cavities. The widely debated ligand recognition, conformational selection, and induced fit mechanisms for ligand binding [7] suggest that the identification of PMs is crucial for a comprehensive understanding of the process.

Recently, molecular dynamics (MD) has emerged as a suitable approach for identifying PMs of specific drugs binding to specific targets [6,8–12]. MD can generate appropriate samples of target–ligand complex structures, allowing conformational flexibility and explicit solvent effects [13,14]. Here, we present an MD-based approach to the study of myosin 2 (a motor protein with a crucial role in eukaryotic motility) and one of its well-known inhibitors. In myosin 2, an ATP bound to the head or motor region is hydrolyzed, which causes conformational changes in the neck region, and this movement is then transferred to actin microfilaments. Myosin 2 is important in muscle contraction, cytokinesis [15,16], the shape formation of cells [17], force generation in cell dynamism [18,19] mitochondrial fission [20], neurite retraction, outgrowth [21], and glioma invasion of the brain [22]. Because of its role in numerous physiological processes, particularly in cellular multiplication and differentiation, myosin 2 has been targeted in several drug design projects [23–25], including investigations of cures for breast cancer [26,27] and pancreatic adenocarcinoma [28].

In the past decades, a non-competitive inhibitor of myosin, S-blebbistatin (BS) [29] and derivatives thereof [23,30,31] have been used to increase our understanding of the role myosin plays in fundamental biological processes [22,32,33]. BS was characterized as a selective [34] inhibitor of non-muscle myosin 2, with its inhibitory effect attributed to blocking phosphate release in the force-producing step, which consequently stabilizes the myosin-ADP-P$_i$ intermediate [18,24,29] through Switch 1 and the P-loop [4] (Figure 1). A hydrophobic pocket of myosin 2 was also discovered [24] that binds BS (here called "destination", D in Figure 1). Pocket D is located on the far edge of the long cavity between the bulk interface of the actin-binding cleft and the nucleotide-binding pocket (pocket N). This binding cavity offers long binding pathways up to 20 Å for BS that may associate with several temporary prerequisite binding sites during its migration from BK to pocket D.

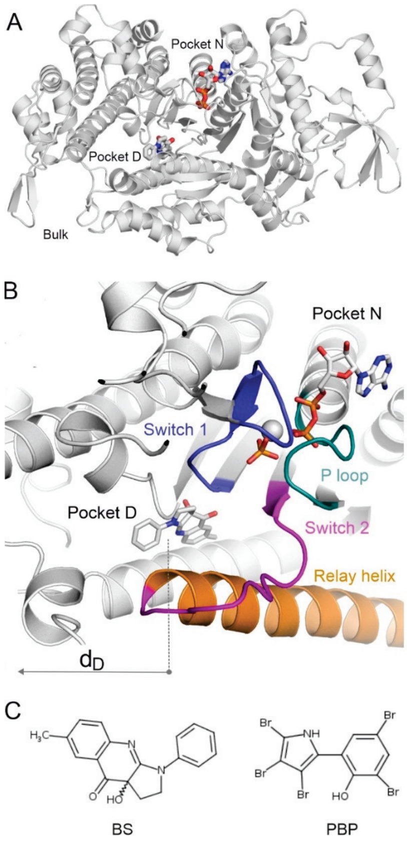


Figure 1. (**A**) Myosin 2 is shown in grey cartoon. The destination pocket (Pocket D) and nucleotide-binding pocket (Pocket N) are highlighted by the experimental binding positions of BS and ADP

(grey sticks), respectively. The figure was prepared using the holoenzyme structure (PDB code 1yv3). (**B**) A close-up of the far edge of the binding tunnel inside myosin 2 including pockets D and N with key structural elements in different colors. The switches, the P-loop, and the relay helix are important elements of the enzyme mechanism and also indicate the location of pocket D. The distance of center of mass (d_D, arrow) of a binding mode of BS is measured from the center of pocket D. (**C**) The Lewis structures of BS and PBP.

In the present study, we introduce a new strategy, NetBinder, to investigate the binding mechanism in the BS-myosin 2 system. NetBinder uses network theory to link the systematically identified PMs and to reconstruct the complete binding mechanism at an atomic resolution. In this manner, we are able to elucidate the complete inhibitory mechanism of BS when binding to the myosin 2 system.

2. Results and Discussion

2.1. Systematic Mapping of Binding Modes

Mapping the binding pathways of BS up to pocket D requires systematically identifying all possible PMs on the target (internal) surface. The final binding modes of known inhibitors were determined by X-ray crystallography [22,24], and the fast computational docking method (Methods Section) was verified to reproduce these experimental binding conformations correctly (Figure S1 and Table S1). However, it has been shown that a simple series of fast docking calculations cannot deliver a fully systematic mapping [14,35] of all possible binding modes. Therefore, the present NetBinder strategy (Figure 2) applies a systematic search technique called Wrap ‘n’ Shake [35], and PMs were detected by wrapping the entire inner surface of the binding cavity of apo myosin 2 (Figure 1) in numerous copies of BS. The wrapping process resulted in a monolayer of 16 docked conformations of BS covering the entire surface of the binding cavity of myosin 2 (Figure 2, Figure S2). The 16 corresponding complexes formed by the docked BS and myosin 2 molecules were equipped with structural water molecules [14] and further challenged in the shaking steps of sixteen 1 µs long MD calculations in simulation boxes filled with explicit water molecules (Methods, Table S2).

Shaking accelerated the dissociation of weakly bound ligand conformations [35] by thermal motions of the explicit water bath and target side chains. In the present study, a ligand copy was considered dissociated if the distance (d_D, Figure 1) between its center of mass and that of the destination BS conformation (in pocket D [24]) became larger than 30 Å. Shaking also allowed an extensive scanning of uncovered segments of the cavity, producing more than 5000 bound conformations for BS, collected in a pool.

After clustering the pooled contents (Methods Section), 23 PMs were distilled and ranked according to their interaction energies (E_{inter}) with the myosin 2 target. Because conformations with low E_{inter} were preferred during clustering (Methods Section), PMs were evenly distributed along the entire cavity (Figure 3A), covering the full range of d_D between PM_1 and PM_{23}. The plots of E_{inter} values of the pooled conformations and PMs as a function of d_D are shown in Figure 3B and Table S3. An energy slope was observed in the plot, that is, E_{inter} significantly decreased with a decrease in d_D. For the PM data points, a linear correlation was calculated for the energy slopes, resulting in a remarkably large squared correlation coefficient (r^2 of 0.7). This finding is in line with the “energy funnel” concept presented in numerous studies [36–39], which assumes that a ligand adopts binding positions with decreasing E_{inter} values when approaching the destination (pocket D) in the target molecule. The energy slope obtained is a two-dimensional cross-section of the energy funnel along variable d_D, representing the position of BS during the binding process.

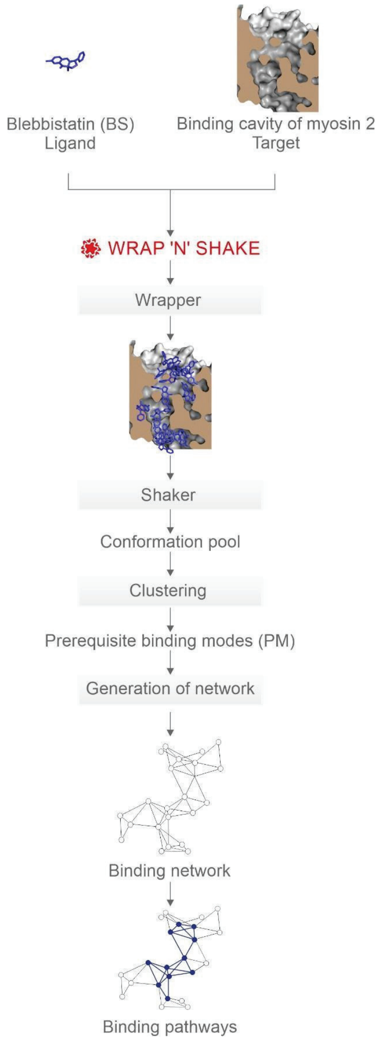


Figure 2. The NetBinder strategy with BS as a ligand and myosin 2 as a target.

There is a "parking lot" (PL, Figure 3C) region of the tunnel gathering the majority of docked conformations centered at a d_D of 10 Å (Figure 3B). The PL is practically the next main binding region after the bulk opening (Bulk in Figure 3A) of the tunnel. The PL was shown experimentally (22) to bind halogenated molecules such as pentabromopseudilin (PBP, Figures 1 and 3). The crystallographic binding conformation of PBP overlaps with PM_2 and PM_4 in the center of PL (Figure 3C) verifying that the PL is a relevant binding region of BS. Thus, the PL does not differentiate between such ligand conformations but rather serves as a large storage depot before their last steps to pocket D. In PM_1, BS has a low E_{inter} of −50.1 kcal/mol at a d_D of 7 Å (Figure 3B), close to that of pocket D (−46.1 kcal/mol, d_D = 0 Å per def.), and therefore, it has a good chance to enter pocket D in one step from PM_1 (see Section 3 for further discussion). The above findings can be of general importance in mechanism-based drug design. While PL serves as a relatively large storage place, the proximal region at PM_1 is a narrower transient place during ligand navigation. It may be expected that the ligand can enter pocket D from PM_1 in a forward step without returning and using bypasses through other PMs backwards. In the search for new druggable sites, the identification of such parking and proximal PMs can be equally important.

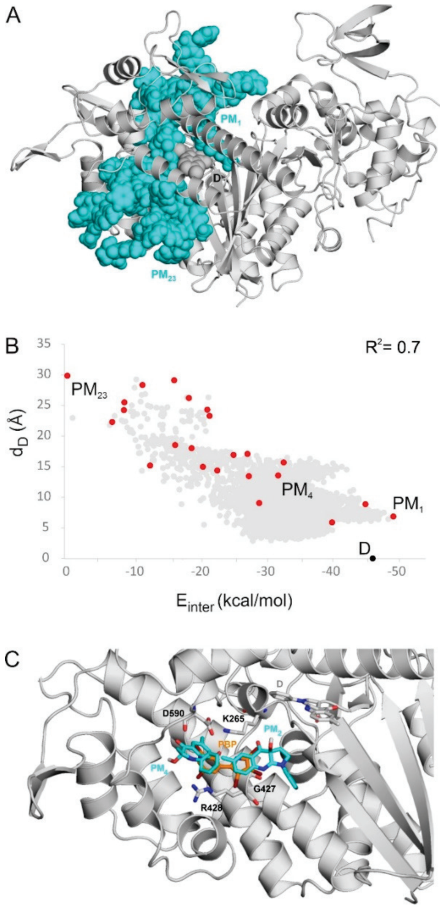


Figure 3. (**A**) The BS PMs (teal spheres) cover the entire binding cavity of myosin 2. The myosin 2 protein is shown as a grey cartoon. PMs with the smallest (PM$_1$) and largest (PM$_{23}$) d_D values are labelled. BS in pocket D (from PDB 1yv3) is shown as grey spheres. (**B**) Interaction energies (E_{inter}) calculated between the PMs of BS and myosin 2 correlate with d_D (energy slope in red dots). A similar trend can be observed for the raw BS conformations of the pools (grey dots). The PMs are ordered by the distances of their centers of mass, measured from that of the crystallographic destination (D) binding mode. (**C**) The overlapping binding position of PBP (from PDB 2jhr) and PMs 2 and 4 from the present study. BS bound to pocket D (from PDB 1yv3) is highlighted as grey sticks. PM2, 4, and PBP are shown as teal and orange sticks, respectively. Surrounding amino acids that participate in the binding of PBP, according to the 2jhr structure [25],′ are shown as grey sticks and are labelled accordingly. Non-polar (C-connected) H atoms are not shown for the calculated ligand molecules (PM) in the figures for clarity.

2.2. *Binding Pathways from Binding Networks*

Beyond the knowledge of the single structural snapshot of pocket D provided by Ref. [24], the determination of PMs (Section 1) was necessary to draw the possible binding pathways of BS. The NetBinder strategy approaches the binding process as a networking problem and produces a representative binding pathway (Figure 4A) based on the corresponding network (Figure 4B). Network science has been successfully applied in structural chemistry [40–45], and we evaluated whether a network representation of the binding events would simplify the elucidation of the binding mechanism of BS. In the NetBinder network approach, the PMs correspond to nodes and edges, representing segments of ligand pathways by definition. All nodes were considered in light of two attributes, namely d_D and the number of edges leading from that node. The first two attributes were adopted from the corresponding PMs of energy slopes (Figure 3B), while the last one was simply counted after the construction of the graph (Figure 4A, Table S4). Nodes with more than four edges were considered hubs, and interconnected hubs form the backbone of a network. These hubs represent the busiest PMs in terms of ligand binding and constitute a binding pathway of a ligand. A graphic representation of the binding network of BS (Figure 4A) holds all nodes connected by edges and was produced by the conversion of the three-dimensional positions of the PMs (Figure 4B) into a two-dimensional connectivity list (Figure 4A, Methods).

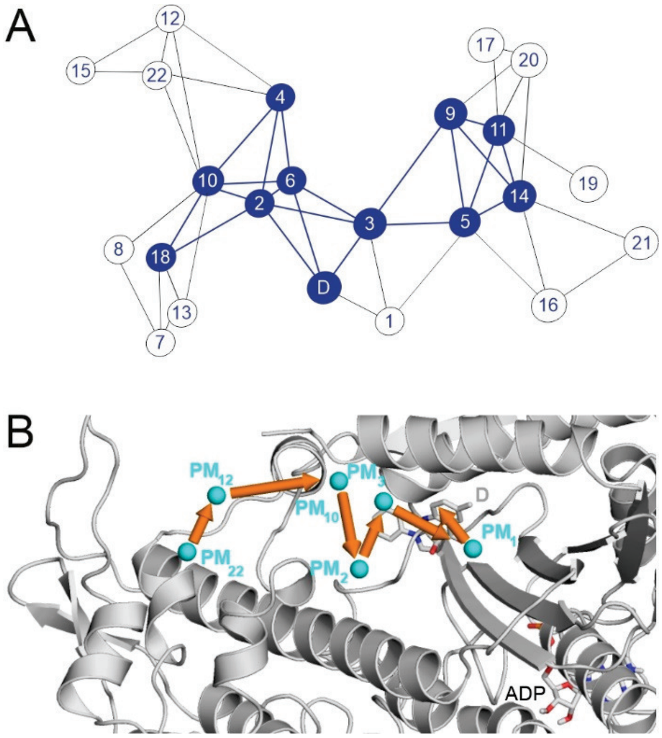


Figure 4. (**A**) Binding network of PMs as nodes for BS. Simple nodes are empty circles; hubs are blue full circle plates. Edges are black lines, and backbone edges are highlighted as blue lines. (**B**) A suggested binding pathway of BS, see also Movie S1. The myosin protein is shown as a grey cartoon. ADP molecule is shown with grey all-atom representation sticks. PMs are shown as blue spheres, and their movement is highlighted with orange arrows. The experimental binding mode (D, superimposed from PDB 1yv3) is shown with grey sticks.

BS has a complex binding graph with 23 nodes, 47 edges, and 10 hubs (Figure 4A), suggesting various pathways for binding. The central hub PM_3, has six connections, and the connectivity of the nodes is especially noticeable around the destination (i.e, at PMs with low d_D values) and is a direct consequence of the presence of proximal PMs with low E_{inter} values (see also Figure 3B). The network has a massive backbone of ten hubs connected to each other that might serve as an excellent "binding highway" (Figure 4B), and it is clearly distinguishable from sub-nets of peripheral nodes, which can be thought of as anchoring regions or dead ends of ligand migration. BS enters the tunnel via anchoring PMs (PM_{22}, PM_{18}, PM_{15}, PM_{12}, and PM_{10}) close to the opening to the bulk, and then it enters the PL. There is a variety of possible pathways at the stage of PL. However, some of the PL nodes (PM_6, PM_4, and PM_2) seem essential to arrive at pocket D. Similarly, the next node, PM_3, is a key hub guiding BS "correctly" towards the destination, as PM_3 has one direct connection to D and three connections that are only two steps away from D. Two of these indirect links go through the PL nodes PM_2 and PM_6, while the third approaches D via PM_1, which has the lowest E_{inter} of all PMs (see also Figure 3B). The remaining four hubs (PM_5, PM_9, PM_{11}, and PM_{14}) probably belong to a separate dissociative dead end (Figure 4A). The above findings are summarized in a representative binding pathway of BS (Figure 5, Movie S1) based on a massive network backbone with key proximal hubs PM_3 and PM_1. This graph-based binding pathway describes the above networking between individual PMs during the full binding process.

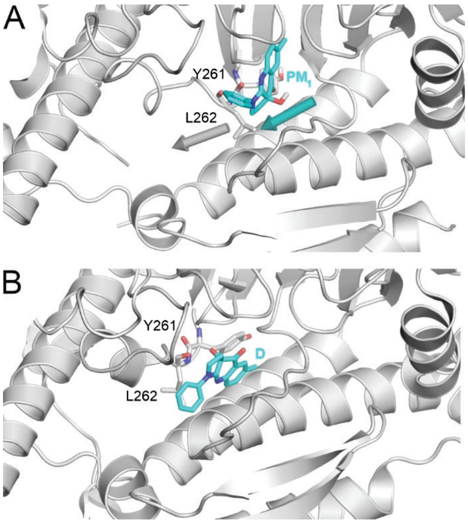


Figure 5. (**A**) In the apo myosin structure (top), L262 blocks the entrance of BS from PM_1 (d_D = 6.9 Å) towards the holo conformation bound to pocket D (**B**). Myosin is shown with grey cartoon. The important amino acids are highlighted with grey all-atom representation sticks and are labelled accordingly. BS is shown with teal all-atom representation sticks. Arrows indicate the movement of L262 and BS.

2.3. Final Test: Docking to the Destination Pocket

In the previous sections, NetBinder determined the most important PMs from among the large pools of possible binding conformations. Based on these PMs, a networking approach was used to produce a complete binding mechanism of BS in the myosin 2 tunnel. The binding graph resulted in a representative binding pathway of BS (Figure 4B), leading to PM_1, which was hypothesized to be a key proximal PM of the lowest E_{inter} (Figure 3B). A validation of PM_1 remains as a final test of NetBinder. For this, it was investigated if the final crystallographic conformation (D) of BS could be obtained starting from PM_1 using fully flexible MD simulations without any biasing restraints.

It is known [24] that for BS to enter into pocket D requires a conformational change to myosin 2, as pocket D is closed by the side chain L262 in the apo form used in the previous sections. During the binding of BS, L262 changes its position (Figure 5), which opens the entrance to the pocket [24]. Because the binding affinity of BS to myosin 2 is rather low (IC_{50}, Figure S2) and the above conformational change is also time-consuming, it was expected that a single MD simulation would not dock BS to the destination site. Thus, terminal docking from PM_1 was attempted in 12 repeated MD simulations with a maximal length of 1 µs each. Two of these simulations resulted in exactly the crystallographic destination position of BS that was determined experimentally [24], with d_D values of 0.9 and 1.5 Å, respectively (Table S5). The atomic level fit of the calculated and crystallographic conformations (Figure S3) demonstrates the precision of MD-based docking and also verified that PM_1 is indeed a key proximal PM, directly leading to the final crystallographic binding mode D of BS.

The fully flexible MD simulations (Methods Section) also allowed the detection of (real time) changes in myosin 2 conformation following the docking process at the atomic level. The MD simulation resulting in the best-matching final BS conformation (d_D = 0.9 Å) was selected for a detailed analysis and is featured in Figure 6 and Movie S2. The corresponding structural changes during BS docking are listed in Table 1, which also compares the final stage of the simulation with the experimentally determined values.

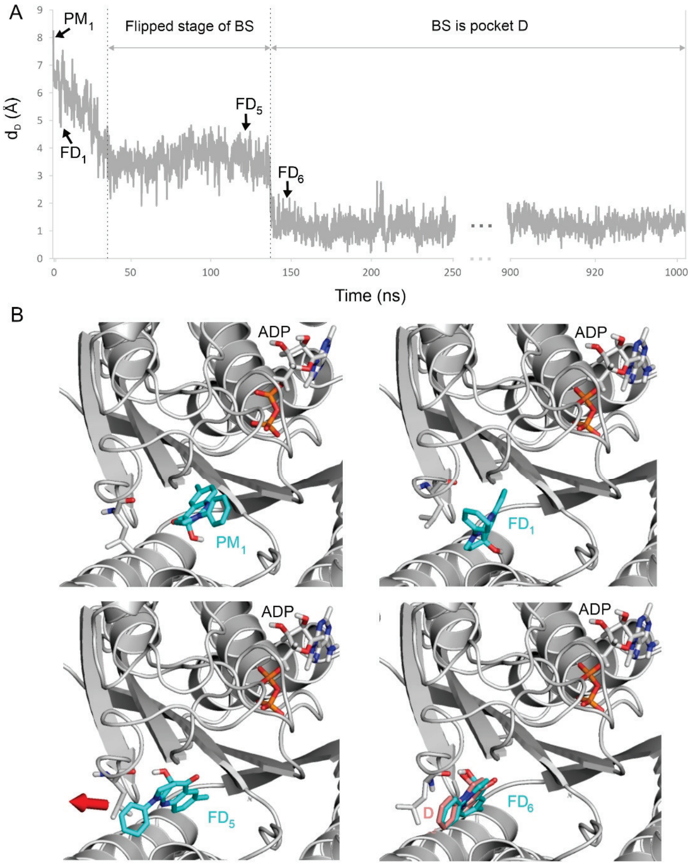


Figure 6. (**A**) Final docking (FD) of BS starting from PM_1 (t = 0 ns) to the destination pocket, represented by a continuous decrease in d_D during the 1 µs simulation. (**B**) Main FD events of the binding process are highlighted (see also Movie S2 for all six final docking steps) as well as the corresponding snapshots of BS and the surrounding amino acids of myosin 2. First, BS undergoes flipping, where its phenyl ring (in FD_1) turns towards L262 to form a hydrophobic interaction in position FD_5. Then, the inward movement of L262 pulls (red arrow) BS towards the destination binding mode into FD_6, which agrees well with reference D (pink sticks) from PDB 1yv3.

Table 1. Structural changes during final docking of BS into the destination site.

Structural Change	Starting Time (ns)	Distance (Å)		
		0 ns	1000 ns	Experimental
Formation of H-bond (G240 . . . BS)	40	8.3 [a]	3.2 [a]	2.8 [a]
Formation of salt bridge (E459 . . . R238)	71	8.3 [b]	4.0 [b]	4.2 [b]
Movement of BS	137	8.0 [c]	0.9 [c]	0 [c]
Formation of H-bond (L262 . . . BS)	137	6.6 [d]	2.5 [d]	2.5 [d]
Flipping of L262	137	3.4 [c]	1.8 [c]	0 [c]
Flipping of Y261	143	1.9 [a]	1.4 [a]	0 [a]
Flipping of S456	447	4.87 [a]	1.9 [a]	0 [a]

[a] Distance between NH (G240)—OH (BS); [b] Distance between CZ (R238)—CD (E459); [c] $d_{D;}$ [d] Distance between O (L262)—OH (BS).

The docking of BS with myosin 2 starts from an open conformation of the cleft between loops Switch 1 and Switch 2 [46], as represented by the apo structure pre-recovery-stroke conformation [46] used in this study. As BS moves from the PM_1 starting position, the cleft closes, and a salt bridge forms spontaneously between E459 and R238, which has also been observed in previous computational studies [47,48]. During this process, Switch 2 is pulled by BS into position via a hydrogen bond between its hydroxyl group and the side chain of E459. As BS further proceeds towards the binding site (i.e., as d_D decreases), the BS-E459 hydrogen bond disappears and, between 40 and 137 ns, BS flips 90° (Figure 6), which is boosted by an interaction between the carbonyl group of BS and the amide group of Q637, and the carbonyl and hydroxyl groups of BS interacting with E467. The flipped stage is an important prerequisite for BS to fit into its destination pocket.

Another important driving factor is the hydrophobic interaction of BS with L262, which also moves with BS and makes room in the destination site (Figure 6, Movie S2), which is consistent with its previously reported structural role, and its final position (d_D = 1.8 Å) is consistent with its experimentally determined position D [24] (Table 1). The positions of both BS and L262 stabilize after 138 ns for the rest of the simulation. Similar agreements with the experimental position can be observed at residues Y261 and S456 as well. Y261 is involved in a π-stacking between the phenyl group of BS, while S456 is important in myosin 2 isoform selectivity [24]. BS forms hydrogen bonds with G240 and L262, which can also be seen as the MD simulation stabilizes at close to experimental values (Table 1).

Thus, the above MD docking calculations verified the prediction of NetBinder on PM_1 and showed that it is a prerequisite binding mode towards crystallographic ligand conformation D.

3. Materials and Methods

Target preparation. Atomic coordinates of the apo myosin 2 target structure were obtained from the Protein Databank [46] (PDB code 1mmd). Atoms of amino acids missing in the target structure were inserted with Swiss-PdbViewer [49]. Missing terminal and non-terminal amino acids and acetyl and amide capping groups were added with the Schrödinger Maestro program package v. 9.6 [50]. The ADP-P_i-Mg^{2+} complex was used because it is the intermediate stage of ATP hydrolysis stabilized after BS binds to myosin 2 [18,29]. An ADP-beryllium trifluoride-Mg^{2+} complex was modified to ADP-P_i-Mg^{2+} by positioning the phosphate ion in the place of the beryllium trifluoride ion (see parametrization of non-amino acid ligands below). The constructed target was then minimized, allowing full flexibility on the heavy atoms (see shaker/energy minimization below). After the minimization steps, target molecule inputs for docking (pdbqt files) were prepared with AutoDock Tools. A united atom representation for hydrogen atoms in non-polar covalent bonds and Gasteiger–Marsili partial charges [51] were applied to the input files.

Ligand preparation. The atomic coordinates of the BS and PBP ligands were extracted from PDB structures 1yv3 and 2jhr, respectively. The pKa values of ligand molecules were calculated using the pKa plug-in in Marvin Sketch, v 6.3.0 [52]. Hydrogen atoms were added according to the correct protonation state at pH 7. Energy minimization was performed on hydrogenated structures using the semi-empirical quantum chemistry program package MOPAC [53]. Geometry optimization with MOPAC was carried out with a gradient of 0.001 kcalmol^{-1}Å^{-1}, and force calculations were carried out with PM3 parameterization. In all cases, the force constant matrices were definitely positive. The minimized ligand molecules were prepared for docking to the targets as described above.

Wrapper. For the BS ligand, the wrapper method [35] was applied to the binding cavity of myosin 2 (instead of the entire surface) by performing 20 consecutive blind docking cycles, which were enough to increase the target–ligand interaction energy close to 0 kcal/mol. In each blind docking cycle, AutoGrid 4.2 was used to generate the grid maps, with boundaries set to cover the whole binding cavity of myosin 2 (for visualization, see Figure 2) using a box of 130 × 100 × 100 grid points centered on the destination BS conformation ([24] PDB 1yv3). The docking cycles were carried out with the AutoDock 4.2 [54] package using the Lamarckian genetic algorithm (LGA). One hundred docked ligand conformations were obtained in each cycle. The docking parameters were used as described in a previous study [55]. Wrapper ended with a trimming step using a cut-off (i.e., a maximum distance between the binding cavity amino acids and the docked conformation) of 3.5 Å. After trimming the BS docking results, only the ligands interacting with the binding cavity amino acids (Table S6) were retained. These filtering steps reduced the number docked conformations (from a starting value of 87) to 16 conformations, and these were investigated further by molecular dynamics simulations after the prediction of interface water molecules.

Prediction of interface water molecules. Interfacial water molecules play an important role in the dissociating weak binder conformations and improving target–ligand complex structures [35]. Appropriate interface water positions were calculated by MobyWat [56] using the M3 protocol as described previously [42] and also described recently in the HydroDock [14] protocol for docked binding modes. Complexes with the predicted interface water molecules were re-minimized by a two-step minimization algorithm described in the energy minimization section.

Parametrization of non-amino acid ligands. The parameterization of non-amino acid ligands (ADP, P_i, and BS) was necessary because the AMBER99SB-ILDN force field [57] does not include molecular mechanics parameters for the ligands used in our study. The charge calculation was performed on the R.E.D. Server [58] for the optimized structure (see ligand preparation) with RESP-A1 [58] charge fitting compatible with AMBER99SB-ILDN force fields. The calculations were performed with the Gaussian09 software [59] using the HF/6-31G* split valence basis set [60].

Energy minimization. Target structures were energy-minimized using a two-step protocol including a steepest descent and a conjugated gradient step. The calculations were performed by the GROMACS 5.0.6 [61] software package with the AMBER99SB-ILDN force field [57] and TIP3P explicit water model [62]. The target structure was placed in the center of a cubic box with the distance between the box and the solute atoms set to 10 Å. The simulation box was filled with water molecules and counter-ions to neutralize the total charge of the system. The particle mesh Ewald method was used for long-range electrostatics. The van der Waals and Coulomb cut-offs were set to 11 Å. The convergence threshold of the first step (steepest descent) was set to 10^3 kJ mol^{-1}nm^{-2}. In the second step (conjugant gradient) of minimization it was set to 10 kJmol^{-1}nm^{-2}. The final structures obtained from the energy minimization were extracted for further calculation with wrapper (see wrapper) or subjected to MD simulations (see shaker).

The molecular dynamics (MD) calculations of various lengths detailed in shaker and final docking were performed with the GROMACS 5.0.6 software package [61] using the AMBER99SB-ILDN force field [57] and the TIP3P explicit water model [62]. The energy-minimized structures were subjected to NPT MD simulations at a temperature of 300 K. For temperature coupling, the velocity rescale algorithm was adopted. Pressure was coupled to the Parrinello–Rahman algorithm with a coupling time constant of 0.5 ps, a compressibility of 4.5×10^{-5} bar^{-1}, and a reference pressure of 1 bar. A particle mesh Ewald summation was used for long range electrostatics. Van der Waals and Coulomb interactions had a cut-off of 11 Å.

Shaker. Parallel MD runs of a maximum 1 μs each were performed on the 16 BS ligand–target complex structures that were obtained from wrapper and energy-minimized as described above. The MD simulations were performed as described in the molecular dynamics section with the following specific settings. Position and distance restraints were applied in the parallel MD runs as detailed below. Position restraints were applied with a force constant of 10^3 kJmol^{-1}nm^{-2} during the whole MD simulation on the backbone C_α atoms of the protein and the heavy atoms of the co-factor (ADP-P_i-Mg^{2+}) and its surrounding amino acids (N127, Y135, K185, T186, and N233). Distance restraints were generated between the atom pairs (Table S7). A simulation was terminated if ligand dissociation (d_D > 30 Å) was observed. The length of the simulations and the d_D values calculated for the last frame of the parallel MD runs are detailed in Table S2. After trimming, 6952 frames were obtained for BS (the "conformation pools").

Calculation of intermolecular interaction energy ($\mathbf{E_{inter}}$). E_{inter} was calculated between myosin 2 target and BS ligand molecules using the Lennard-Jones parameters of the Amber force field [57] in Equation (1).

$$\begin{aligned} \mathbf{E_{inter}} &= \sum_{\mathbf{i,j}}^{\mathbf{N_T N_L}} \left[\frac{\mathbf{A_{ij}}}{\mathbf{r_{ij}^{12}}} - \frac{\mathbf{B_{ij}}}{\mathbf{r_{ij}^{6}}} \right] \\ \mathbf{A_{ij}} &= \boldsymbol{\varepsilon}_{\mathbf{ij}} \mathbf{R_{ij}^{12}} \\ \mathbf{B_{ij}} &= 2\boldsymbol{\varepsilon}_{\mathbf{ij}} \mathbf{R_{ij}^{6}} \\ \mathbf{R_{ij}} &= \mathbf{R_i} + \mathbf{R_j} \\ \boldsymbol{\varepsilon}_{\mathbf{ij}} &= \sqrt{\boldsymbol{\varepsilon}_{\mathbf{i}}\ \boldsymbol{\varepsilon}_{\mathbf{j}}} \end{aligned} \tag{1}$$

where ε_{ij} is the potential well depth at equilibrium between the ith (ligand) and jth (target) atoms, R_{ij} is the inter-nuclear distance at equilibrium between the ith (ligand) and jth (target) atoms, N_T is the number of target atoms, and N_L is the number of ligand atoms.

Clustering. The conformation pools were forwarded from the clustering process and E_{inter} for all frames of the conformation pools (the calculation of intermolecular interaction energy). The BS conformations were clustered and ranked by their E_{inter} values and by the closest distance between each heavy atom of the ligand (d_{min}). The BS conformation with the lowest E_{inter} value among the cumulated BS copies was selected to represent Cluster 1 (PM_1). The BS conformation of the second lowest E_{inter} was considered a new Cluster 2 (PM_2) if $d_{min} > d_{rnk}$, where d_{rnk} was a ranking tolerance (a distance cut-off of separation of clusters (PMs) from each other) set to 1.75 Å. If $d_{min} \leq d_{rnk}$ for that cluster, then the ligand conformation of the second lowest E_{inter} was placed in Cluster 1. All subsequent clusters were evaluated by this method, and the resulting representative conformations were evenly spread over the myosin 2 cavity without contacting each other. This clustering technique manifests in the shift in the red dots in Figure 3B towards the best binder cluster representative as measured by E_{inter} rather than towards the conformation closer to the destination, pocket D (in other words, the d_D of the red dots does not necessarily approach zero). The 1.75 Å distance cut-off was used so that the cluster representatives would not overlap in order to systematically and evenly cover the binding cavity. After clustering, 23 conformations were obtained for BS (Figure 3B, Table S3). For reference, d_D values between the PMs and the DC conformations were also calculated.

Calculation of distances between the centers of mass. The distances between the centers of mass (d_D) of two BS conformations (simulated and reference destination) were calculated for each MD simulation frame and used to eliminate MD frames where a BS copy was dissociated from the myosin 2 surface (Table S3). The same d_D was also used for the calculation of Figure 3B and the network evaluations of the conformation pools. The d_{PM} values were also calculated between the centers of mass of the PMs obtained after clustering, and these values were used to generate the binding graphs (Table S4).

Binding network. The minimal d_{PM} distance between two PMs was taken from the distance matrix (previous point) with a cut-off of d_D >12 Å set between PMs (the length of BS is about 12 Å). By this cut-off, BS PM_{23} had no connections to any neighboring PMs. Hence, this PM was not included in the graph generation. Second, graphs were generated with the NetDraw mode of the program MobyWat [42]. This mode of MobyWat was initially used to generate water–water or water–target interaction networks by calculating the distance between heavy atoms. In our case, instead of the distance between oxygen atoms of water molecules, the d_D of each PM was calculated to be used as an input. Additionally, in the B-factor column of our input PDB file, the d_D measure between the PM and the D center of mass was used. NetDraw's source code was modified to allow a maximum of ten edges for each node instead of the default maximum of four edges (a legacy from NetDraw's original use for water networks). Gephi 0.9.2 [63] and CorelDraw were used to visualize and re-draw the graphs according to the structural information of the PM positions and the edges generated by NetDraw.

The final docking of BS was performed by twelve simulated annealing, fully flexible MD runs of 1 µs each (Table S5) using the same MD parameters as described above, except that simulation annealing was implemented with the temperature scheme presented in Table S8. The target was fully flexible in these runs, no position restraints were applied at all, except that the cofactor and frames were exported every 0.1 ns, resulting in 10^4 frames after 1 µs of simulation. The d_D of BS was calculated for each frame during the 1 µs (Figure 6). For the d_D of the BS conformation in the final frame, see Table S5.

4. Conclusions

Experimental structure determination techniques provide invaluable information of the atomic resolution binding modes of drugs to their macromolecular targets. While some techniques for determining binding dynamics have been suggested [64], complete drug mechanisms cannot be produced routinely, and a systematic method for detecting prerequisite binding modes (PMs) has been lacking. Based on the static structures from experiments, theoretical methods have provided the missing binding dynamics, and some PMs of drugs have been found [6,10,12,35] that should be considered "footsteps" on a drug's pathway towards its destination on the target. However, the connections between these PMs have hitherto not been uncovered. NetBinder solves this problem by creating networks of PMs and extracting binding pathways. NetBinder, combined with fully flexible MD simulations for the final docking stages, can supply complete mechanisms for drugs that target a long tunnel-shaped binding channel, such as that of myosin 2. Similarly shaped binding channels often occur in transporters [65], important receptors such as muscarinic [66], enzymes such as cyclooxygenase-1 and -2 [67], and transmembrane viral ion channels, such as that of the influenza A [68,69] and SARS-CoV-2 viruses [14,70], and it is our hope that the NetBinder strategy presented here will be adapted to aid in the investigation of such targets. The present work also offers a starting kit of new tools for the identification and classification of PMs, binding sites (parking and proximal), and networking elements (hub and backbone) in mechanism-based drug design.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23137313/s1.

Author Contributions: M.B. and B.Z.Z. performed the research and wrote the manuscript; D.v.d.S. created the concepts and wrote the manuscript; C.H. supervised and designed the research, created the concepts, and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Hungarian National Research, Development and Innovation Office (K123836). Project no. TKP2021-EGA-16 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the EGA 16 funding scheme. Project no. TKP2021-EGA-13 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the EGA-13 funding scheme and the Lorand Eotvos research network. This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The work was supported by ÚNKP-21-5 and ÚNKP-21-3-II from the New National Excellence Program of the Ministry for Innovation and Technology and by PTE ÁOK-KA 2021/KA-2021-39. The project has been supported by the European Union, co-financed by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Files are provided online at: https://zenodo.org/record/6536287#.YnzSsFRByHt. Accessed on 27 June 2022.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research have been achieved using the DECI resource Archer2 based in the UK at the National Supercomputing Service with support from the PRACE aisbl.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blundell, T.L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002**, *1*, 45–54. [CrossRef] [PubMed]
2. Hui, R.; Edwards, A. High-throughput protein crystallization. *J. Struct. Biol.* **2003**, *142*, 154–161. [CrossRef]
3. Cheng, Y.; Glaeser, R.M.; Nogales, E. How Cryo-EM Became so Hot. *Cell* **2017**, *171*, 1229–1231. [CrossRef]
4. Houdusse, A.; Sweeney, H.L. How Myosin Generates Force on Actin Filaments. *Trends Biochem. Sci.* **2016**, *41*, 989–997. [CrossRef]
5. Jiang, W.; Tang, L. Atomic cryo-EM structures of viruses. *Curr. Opin. Struct. Biol.* **2017**, *46*, 122–129. [CrossRef] [PubMed]
6. Zsidó, B.Z.; Börzsei, R.; Pintér, E.; Hetényi, C. Prerequisite binding modes determine the dynamics of action of covalent agonists of ion channel trpa1. *Pharmaceuticals* **2021**, *14*, 988. [CrossRef] [PubMed]
7. Vogt, A.D.; Di Cera, E. Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. *Biochemistry* **2012**, *51*, 5894–5902. [CrossRef] [PubMed]
8. Balint, M.; Jeszenoi, N.; Horvath, I.; Abraham, I.M.; Hetenyi, C. Dynamic changes in binding interaction networks of sex steroids establish their non-classical effects. *Sci. Rep.* **2017**, *7*, 14847. [CrossRef]
9. Decherchi, S.; Berteotti, A.; Bottegoni, G.; Rocchia, W.; Cavalli, A. The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun.* **2015**, *6*, 6155. [CrossRef]
10. Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10184–10189. [CrossRef]
11. Dror, R.O.; Pan, A.C.; Arlow, D.H.; Borhani, D.W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D.E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13118–13123. [CrossRef] [PubMed]
12. Shan, Y.; Kim, E.T.; Eastwood, M.P.; Dror, R.O.; Seeliger, M.A.; Shaw, D.E. How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* **2011**, *133*, 9181–9183. [CrossRef] [PubMed]
13. Zsidó, B.Z.; Hetényi, C. The role of water in ligand binding. *Curr. Opin. Struct. Biol.* **2021**, *67*, 1–8. [CrossRef]
14. Zsidó, B.Z.; Börzsei, R.; Szél, V.; Hetényi, C. Determination of Ligand Binding Modes in Hydrated Viral Ion Channels to Foster Drug Design and Repositioning. *J. Chem. Inf. Model.* **2021**, *61*, 4011–4022. [CrossRef] [PubMed]
15. Straight, A.F.; Cheung, A.; Limouze, J.; Chen, I.; Westwood, N.J.; Sellers, J.R.; Mitchison, T.J. Dissecting temporal and spatial control of cytokinesis with a myosin II inhibitor. *Science* **2003**, *299*, 1743–1747. [CrossRef]
16. Ma, X.; Kovacs, M.; Conti, M.A.; Wang, A.; Zhang, Y.; Sellers, J.R.; Adelstein, R.S. Nonmuscle myosin II exerts tension but does not translocate actin in vertebrate cytokinesis. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 4509–4514. [CrossRef]
17. Mogilner, A.; Keren, K. The Shape of Motile Cells. *Curr. Biol.* **2009**, *19*, R762–R771. [CrossRef]
18. Takacs, B.; Billington, N.; Gyimesi, M.; Kintses, B.; Malnasi-Csizmadia, A.; Knight, P.J.; Kovacs, M. Myosin complexed with ADP and blebbistatin reversibly adopts a conformation resembling the start point of the working stroke. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 6799–6804. [CrossRef]
19. Wakatsuki, T. Mechanics of cell spreading: Role of myosin II. *J. Cell Sci.* **2003**, *116*, 1617–1625. [CrossRef]
20. Korobova, F.; Gauvin, T.J.; Higgs, H.N. Report A Role for Myosin II in Mammalian Mitochondrial Fission. *Curr. Biol.* **2014**, *24*, 409–414. [CrossRef]
21. Wylie, S.R.; Chantler, P.D. Myosin IIA Drives Neurite Retraction. *Mol. Biol. Cell* **2003**, *14*, 4654–4666. [CrossRef] [PubMed]

22. Beadle, C.; Assanah, M.C.; Monzo, P.; Vallee, R.; Rosenfeld, S.S.; Canoll, P. The Role of Myosin II in Glioma Invasion of the Brain. *Mol. Biol. Cell* **2008**, *19*, 3357–3368. [CrossRef] [PubMed]
23. Lucas-Lopez, C.; Allingham, J.S.; Lebl, T.; Lawson, C.P.A.T.; Brenk, R.; Sellers, J.R.; Rayment, I.; Westwood, N.J. The small molecule tool (S)-(−)-blebbistatin: Novel insights of relevance to myosin inhibitor design. *Org. Biomol. Chem.* **2008**, *6*, 2076–2084. [CrossRef] [PubMed]
24. Allingham, J.S.; Smith, R.; Rayment, I. The structural basis of blebbistatin inhibition and specificity for myosin II. *Nat. Struct. Mol. Biol.* **2005**, *12*, 378–379. [CrossRef]
25. Fedorov, R.; Böhl, M.; Tsiavaliaris, G.; Hartmann, F.K.; Taft, M.H.; Baruch, P.; Brenner, B.; Martin, R.; Knölker, H.J.; Gutzeit, H.O.; et al. The mechanism of pentabromopseudilin inhibition of myosin motor activity. *Nat. Struct. Mol. Biol.* **2009**, *16*, 80–88. [CrossRef]
26. Paszek, M.J.; Dufort, C.C.; Rossier, O.; Bainer, R.; Cassereau, L.; Rubashkin, M.G.; Magbanua, M.J.; Kurt, S. The cancer glycocalyx mechanically primes integrin-mediated growth and survival. *Nature* **2015**, *511*, 319–325. [CrossRef]
27. Derycke, L.; Stove, C.; Wever, O.D.E.; Dollé, L.; Colpaert, N.; Depypere, H.; Michalski, J.; Bracke, M. The role of non-muscle myosin IIA in aggregation and invasion of human MCF-7 breast cancer cells. *Int. J. Dev. Biol.* **2011**, *55*, 835–840. [CrossRef]
28. Duxbury, M.S.; Ashley, S.W.; Whang, E.E. Inhibition of pancreatic adenocarcinoma cellular invasiveness by blebbistatin: A novel myosin II inhibitor. *Biochem. Biophys. Res. Commun.* **2004**, *313*, 992–997. [CrossRef]
29. Kovacs, M.; Toth, J.; Hetenyi, C.; Malnasi-Csizmadia, A.; Sellers, J.R. Mechanism of blebbistatin inhibition of myosin II. *J. Biol. Chem.* **2004**, *279*, 35557–35563. [CrossRef]
30. Kepiro, M.; Varkuti, B.H.; Vegner, L.; Voros, G.; Hegyi, G.; Varga, M.; Malnasi-Csizmadia, A. Para-nitroblebbistatin, the non-cytotoxic and photostable myosin II inhibitor. *Angew. Chem. Int. Ed.* **2014**, *53*, 8211–8215. [CrossRef]
31. Rauscher, A.; Gyimesi, M.; Kovacs, M.; Malnasi-Csizmadia, A. Targeting Myosin by Blebbistatin Derivatives: Optimization and Pharmacological Potential. *Trends Biochem. Sci.* **2018**, *43*, 700–713. [CrossRef] [PubMed]
32. Chan, C.J.; Ekpenyong, A.E.; Golfier, S.; Li, W.; Chalut, K.J.; Otto, O.; Elgeti, J.; Guck, J.; Lautenschläger, F. Myosin II activity softens cells in suspension. *Biophys. J.* **2015**, *108*, 1856–1869. [CrossRef] [PubMed]
33. Sayyad, W.A.; Amin, L.; Fabris, P.; Ercolini, E.; Torre, V. The role of myosin-II in force generation of DRG filopodia and lamellipodia. *Sci. Rep.* **2015**, *5*, 7842. [CrossRef]
34. Eddinger, T.J.; Meer, D.P.; Miner, A.S.; Joel, M.; Rovner, A.S.; Ratz, P.H. Potent Inhibition of Arterial Smooth Muscle Tonic Contractions by the Selective Myosin II Inhibitor, Blebbistatin. *Med. Econ.* **2013**, *90*, 22–24. [CrossRef] [PubMed]
35. Balint, M.; Jeszenoi, N.; Horvath, I.; van der Spoel, D.; Hetenyi, C. Systematic exploration of multiple drug binding sites. *J. Cheminform.* **2017**, *9*, 65. [CrossRef] [PubMed]
36. Aldeghi, M.; Heifetz, A.; Bodkin, M.J.; Knapp, S.; Biggin, P.C. Predictions of ligand selectivity from absolute binding free energy calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957. [CrossRef]
37. Chu, W.T.; Wang, J. Energy landscape topography reveals the underlying link between binding specificity and activity of enzymes. *Sci. Rep.* **2016**, *6*, 27808. [CrossRef] [PubMed]
38. Moraca, F.; Amato, J.; Ortuso, F.; Artese, A.; Pagano, B.; Novellino, E.; Alcaro, S.; Parrinello, M.; Limongelli, V. Ligand binding to telomeric G-quadruplex DNA investigated by funnel-metadynamics simulations. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E2136–E2145. [CrossRef]
39. Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **2017**, *9*, 1005–1011. [CrossRef]
40. Csermely, P.; Korcsmáros, T.; Kiss, H.J.M.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol. Ther.* **2013**, *138*, 333–408. [CrossRef]
41. Hulovatyy, Y.; Chen, H.; Milenković, T. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics* **2015**, *31*, i171–i180. [CrossRef] [PubMed]
42. Jeszenoi, N.; Balint, M.; Horvath, I.; Van Der Spoel, D.; Hetenyi, C. Exploration of Interfacial Hydration Networks of Target-Ligand Complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148–158. [CrossRef] [PubMed]
43. Brysbaert, G.; Blossey, R.; Lensink, M.F. The inclusion of water molecules in residue interaction networks identifies additional central residues. *Front. Mol. Biosci.* **2018**, *5*, 88. [CrossRef] [PubMed]
44. Brysbaert, G.; Lorgouilloux, K.; Vranken, W.F.; Lensink, M.F. RINspector: A Cytoscape app for centrality analyses and DynaMine flexibility prediction. *Bioinformatics* **2018**, *34*, 294–296. [CrossRef]
45. Kunstmann, S.; Gohlke, U.; Broeker, N.K.; Roske, Y.; Heinemann, U.; Santer, M.; Barbirz, S. Solvent Networks Tune Thermodynamics of Oligosaccharide Complex Formation in an Extended Protein Binding Site. *J. Am. Chem. Soc.* **2018**, *140*, 10447–10455. [CrossRef]
46. Fisher, A.J.; Smith, C.A.; Thoden, J.B.; Smith, R.; Sutoh, K.; Holden, H.M.; Rayment, I. X-ray Structures of the Myosin Motor Domain of Dictyostelium discoideum Complexed with MgADP.BeFx and MgADP.AlF4-. *Biochemistry* **1995**, *34*, 8960–8972. [CrossRef]
47. Fischer, S.; Windshügel, B.; Horak, D.; Holmes, K.C.; Smith, J.C. Structural mechanism of the recovery stroke in the myosin molecular motor. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6873–6878. [CrossRef]
48. Koppole, S.; Smith, J.C.; Fischer, S. The Structural Coupling between ATPase Activation and Recovery Stroke in the Myosin II Motor. *Structure* **2007**, *15*, 825–837. [CrossRef]

49. Guex, N.; Peitsch, M.C.; Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **2009**, *30*, 162–173. [CrossRef]
50. LLC. *Schrödinger Release 2019-3: Maestro, Schrödinger*; LLC: New York, NY, USA, 2019. Available online: https://www.schrodinger.com/maestro. (accessed on 27 June 2022).
51. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228. [CrossRef]
52. Chemaxon. *Marvin Sketch*; v 6.3.0; Chemaxon: Budapest, Hungary, 2014.
53. Stewart, J.J.P. *MOPAC2009, 2009*; Steward Computational Chemistry: Colorado Springs, CO, USA, 2008.
54. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *28*, 73–86. [CrossRef] [PubMed]
55. Hetényi, C.; Van Der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett.* **2006**, *580*, 1447–1450. [CrossRef] [PubMed]
56. Jeszenoi, N.; Horváth, I.; Bálint, M.; Van Der Spoel, D.; Hetényi, C. Mobility-based prediction of hydration structures of protein surfaces. *Bioinformatics* **2015**, *31*, 1959–1965. [CrossRef]
57. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78*, 1950–1958. [CrossRef] [PubMed]
58. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J.C.; Cieplak, P.; Dupradeau, F.Y. RED Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **2011**, *39*, 511–517. [CrossRef] [PubMed]
59. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; et al. *Gaussian*; Version 09; Gaussian, Inc.: Wallingford, CT, USA, 2009.
60. Krishnan, R.; Binkley, J.S.; Seeger, R.; Pople, J.A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654. [CrossRef]
61. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
62. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
63. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In Proceedings of the International AAAI Conference on Web and Social Media, San Jose, CA, USA, 17–20 May 2009.
64. Boutet, S.; Lomb, L.; Williams, G.J.; Barends, T.R.; Aquila, A.; Doak, R.B.; Weierstall, U.; DePonte, D.P.; Steinbrener, J.; Shoeman, R.L.; et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science* **2012**, *337*, 362–364. [CrossRef]
65. Choinowski, T.; Hauser, H.; Piontek, K. Structure of sterol carrier protein 2 at 1.8 Å resolution reveals a hydrophobic tunnel suitable for lipid binding. *Biochemistry* **2000**, *39*, 1897–1902. [CrossRef]
66. Logothetis, D.E.; Kurachi, Y.; Galper, J.; Neer, E.J.; Clapham, D.E. The βγ subunits of GTP-binding proteins activate the muscarinic K+ channel in heart. *Nature* **1987**, *325*, 321–326. [CrossRef] [PubMed]
67. Thuresson, E.D.; Lakkides, K.M.; Rieke, C.J.; Sun, Y.; Wingerd, B.A.; Micielli, R.; Mulichak, A.M.; Malkowski, M.G.; Garavito, R.M.; Smith, W.L. Prostaglandin endoperoxide H synthase-1: The functions of cyclooxygenase active site residues in the binding, positioning, and oxygenation of arachidonic acid. *J. Biol. Chem.* **2001**, *276*, 10347–10357. [CrossRef] [PubMed]
68. Acharya, R.; Carnevale, V.; Fiorin, G.; Levine, B.G.; Polishchuk, A.L.; Balannik, V.; Samish, I.; Lamb, R.A.; Pinto, L.H.; DeGrado, W.F.; et al. Structure and mechanism of proton transport through the transmembrane tetrameric M2 protein bundle of the influenza A virus. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 15075–15080. [CrossRef]
69. Thomaston, J.L.; Polizzi, N.F.; Konstantinidi, A.; Wang, J.; Kolocouris, A.; Degrado, W.F. Inhibitors of the M2 Proton Channel Engage and Disrupt Transmembrane Networks of Hydrogen-Bonded Waters. *J. Am. Chem. Soc.* **2018**, *140*, 15219–15226. [CrossRef] [PubMed]
70. Mandala, V.; McKay, M.; Shcherbakov, A.; Dregni, A.; Kolocouris, A.; Hong, M. Structure and Drug Binding of the SARS-CoV-2 Envelope Protein in Phospholipid Bilayers. *Nat. Struct. Mol. Biol.* **2020**, *27*, 1202–1208. [CrossRef] [PubMed]

D8

International Journal of
Molecular Sciences

MDPI

Review

Molecular Structure, Binding Affinity, and Biological Activity in the Epigenome

Balázs Zoltán Zsidó and Csaba Hetényi *

Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary; zsido.balazs@pte.hu
* Correspondence: hetenyi.csaba@pte.hu; Tel.: +36-72-536-000 (ext. 31649)

Received: 28 May 2020; Accepted: 8 June 2020; Published: 10 June 2020

Abstract: Development of valid structure–activity relationships (SARs) is a key to the elucidation of pathomechanisms of epigenetic diseases and the development of efficient, new drugs. The present review is based on selected methodologies and applications supplying molecular structure, binding affinity and biological activity data for the development of new SARs. An emphasis is placed on emerging trends and permanent challenges of new discoveries of SARs in the context of proteins as epigenetic drug targets. The review gives a brief overview and classification of the molecular background of epigenetic changes, and surveys both experimental and theoretical approaches in the field. Besides the results of sophisticated, cutting edge techniques such as cryo-electron microscopy, protein crystallography, and isothermal titration calorimetry, examples of frequently used assays and fast screening techniques are also selected. The review features how different experimental methods and theoretical approaches complement each other and result in valid SARs of the epigenome.

Keywords: histone code; DNA; RNA; post-translational modification; in vitro; in vivo; molecular modeling; docking; high throughput screening; drug design; molecular dynamics; cancer; peptide

1. Molecular Background of the Epigenome

According to Waddington, epigenetics is “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” [1,2]. Riggs further specified epigenetics as “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence” [3]. Following Waddington’s proposal, Holliday [4] also refers to a switch mechanism early in development that results in a random, yet permanent and successively heritable activation of some chromosomes and deactivation of others. This evolutionary chain of definitions of epigenetics is the hallmark of a rapidly developing and indispensable approach which “provides hope that we are more than just the sequence of our genes” [5].

Epigenetics explains distinct aspects of ontogenesis in normal physiology as well as pathophysiological effects of various diseases resulted by our lifestyles and the environment that might be inheritable [6]. The effect of lifestyle factors such as nightshift working, physical activity, stressful experiences, polyphenols and phytoestrogens in food, on epigenetic modifications has been reviewed [7]. Epigenetic regulation is important in learning, memory and neurogenesis, and it plays a role in related diseases, such as depression and schizophrenia [8]. Epigenetic changes also play a role in neurological, immunological and viral diseases [9]. Cancer is one of the most frequently studied diseases in general and in epigenetics, as well. Epigenetic alterations interfere with tumor progenitor genes, increasing the likelihood of cancer and worsening its prognosis [10–12]. Feinberg’s study [13] highlights a specific disease, Beckwith-Wiedemann Syndrome, which is caused by epigenetic defects that are specifically linked to cancer risk in affected patients. This opens up the possibility of accepting epigenetic alterations as cause, rather than consequence of cancer.

To capture the epigenetic mechanisms of developmental biology it is necessary to unravel how the genetic program unfolds or is modified in the case of diseases at the level of nucleosomes. This goal can be achieved by the development of structure–activity relationships based on intermolecular interactions of bio-macromolecules directing cell cycle, transcription, translation and cellular signaling pathways [14–16]. In this sense, precise understanding of epigenetic regulation requires atomic level determination of interactions in nucleosomes between histone proteins and DNA [17,18], readers, and writers affecting gene expression in the brain [19]. Figure 1 sketches the afore-mentioned levels of epigenetic regulation. With this molecular background in mind, epigenetics can be also considered as "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" [20].

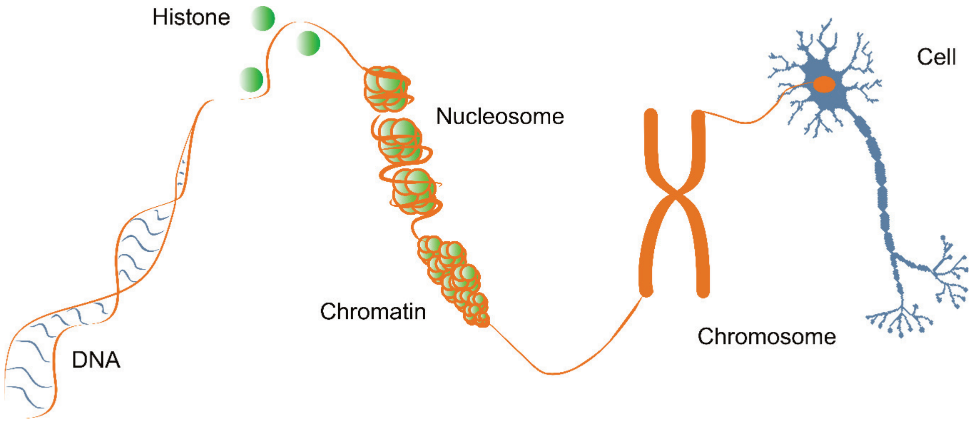


Figure 1. Structural background of the epigenome with a schematic illustration of major organizational levels. A neuron was selected to represent the cellular level as "precise epigenetic regulation may be critical for neuronal homeostasis" [21].

From a structural viewpoint, epigenetic regulatory mechanisms can be classified into the following categories according to the participant bio-macromolecules [22].

Category 1. Covalent modifications of DNA play a crucial role in processes for transfer of the genetic code, like in transcription. Such modifications have been linked [23] to specific types of cancers via enzymes such as methyltransferases, acetyltransferases, and kinases. For example, DNA methylation is associated with diabetes and cancer [24]. Methylation often occurs on cytosine and is carried out by DNA methyltransferases (DNMT), DNMT3A, DNMT3B and DNMT1. This results in gene repression by modifying the recognition sites and histone binding of DNA binding proteins. The hypermethylation of TRPA1 gene occurs in people with post-herpetic neuralgia and lower back pain, and is also associated with pain symptoms, burning sensations and a decreased heat pain threshold [25]. Acetylation of DNA is also important in the pathomechanism of certain types of cancer. DNA acetylation is controlled by two enzymatic families: (1) the histone lysine acetyltransferases (KAT) and (2) histone deacetylases (HDACs).

Category 2. Covalent modifications of histones. The core histone proteins H2B, H2A, H3 and H4 are essential constituents of the chromatin. Two copies of each histone are assembled into an octamer and a DNA super-helix of ca. 146 base pairs are organized around it forming the nucleosome (Figure 1), the elementary unit of the chromatin [26–29]. Nucleosomes are connected by linker DNA, and histone H1, which induces a compact structure upon binding [30] to finally yield a high-level structure of supercoiled helices building up the chromosomes [29]. Histones, except for histone H1, have long peptide tails passing through the DNA wrap of the nucleosomes (Figures 1 and 2), between the turns of the coiled DNA. A wide range of structural elements extends from the histone fold domain motifs, that are structurally conserved regions found near the C-terminus in every core histone, responsible for organizing the histones into heterodimers. These structural elements play an important role in protein–protein interactions in epigenetics [29]. The N-terminal amino acids of histones also play a significant role in the interference between the DNA superhelix and neighboring compounds [29] and hold numerous PTMs [31] (Figure 2).

A great array of PTMs of the histones creates the ’histone code’ [27,28], completing the information of the genetic code [24,27]. The histone tails pass through the DNA supercoil and their PTMs are accessible for a direct or enzyme-mediated readout [32]. Besides the effector (reader) proteins, there are also writers, erasers [33,34] and remodelers [21] working in the heart of PTM machinery of the histone code (Figure 2). While readers recognize the PTMs, writers add, and erasers delete them, respectively.

Abundant PTMs including methylation, acetylation, phosphorylation and ubiquitination mostly appear on the N-terminal linear tails of the histones. For example, lysine residues can be methylated or acetylated, and a new study [35] shows, that their lactylation is also possible, directly stimulating gene transcription from chromatin.

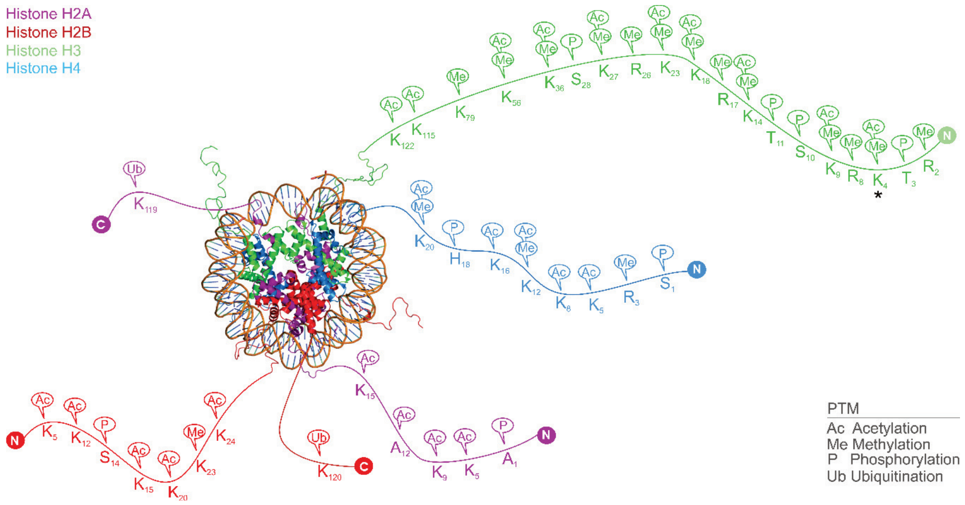


Figure 2. The three-dimensional structure of a nucleosome combined with a schematic representation of post-translational modifications (PTMs) on the histone tails assembled from recent articles [31,36–46]. The nucleosome structure was rendered in top view by PyMol [47] using PDB structure 1kx5 [48]. Histone proteins are shown in cartoon representation as wrapped by the DNA double helix. The N- and C-terminal tails of histone proteins pass through the cylinder of the supercoiled DNA and are available for reader proteins recognizing the PTMs, key components of the histone code system.

Different histone PTMs play various roles in normal physiology and disease pathomechanisms. PTMs have a wide variety of functions [49], by directly controlling nucleosome stability they inflict DNA repair and transcription and even influence nucleosome structure. For example, di-methylation of the 4th lysine of the histone H3 tail (H3K4me2, the location is marked with an asterisk in Figure 2) results in transcriptional activation of protein WDR5, which plays an essential role in vertebrate development [50]. (Notably, the above-abridged form of histone PTMs will be used throughout this manuscript. The abridgment includes the type of histone “H3” in the asterisk-marked example, the type and serial number of amino acid “K3” holding the PTM, and the type and count of PTM “me2”). The lack of WDR5 function results in delay of ontogenesis, by four stages of development [51]. At the same time, histone methylation is involved in the development of cancer [24] and Huntington’s disease [24]. Various enzymes modulate this unique histone code during condensation, such as histone acetylases (HAT), HDACs, histone methylases, and other histone-modifying enzymes. Similarly to histone PTMs involved in (patho)physiology, their reader, writer and eraser enzymes also play an important role in maintaining physiological functions, and in disease pathomechanisms, creating a tempting target for drug design [34,52–55]. Histone acetylation plays an important role in regulating gene activity, through influencing the stability of the chromatin [36] and is also important in diabetes, asthma, and cancer [24,56]. De-acetylation maintains immuno-physiological pathways of host defense. Accordingly, HDAC inhibitors increase susceptibility to various pathogens in vivo [53]. Histone methylation and acetylation also partake in gene expression (silencing or promotion) of cyclin-dependent kinase 5 (Cdk5) gene. The expression of neuronal protein Cdk5 is increased upon chronic cocaine administration [57] and the Cdk5-zinc finger protein transcription factors can bi-directionally regulate Cdk5 gene expression with the enrichment of their respective histone modifications. Histone H3K9/14ac increases cocaine-induced locomotor behavior, while H3K9me2 attenuates it [58].

Category 3. Small non-protein coding RNAs or microRNAs (miRNAs) and long non-coding RNAs (lncRNAs). The miRNAs are responsible for the direct destruction or translational repression of their target RNAs, actually being functionally equivalent to small interfering RNAs (siRNAs) [59], whose function is to promote the degradation of mRNAs or inhibit their translation. Altered miRNA expression partakes in various cancer pathomechanisms, through silencing of tumor suppressor genes. A decrease in miRNA precursor family (miR) miR-101 expression leads to increased H3K27 trimethylation, which is a silencing mutation. A decreased miR-29 expression leads to an increase in the activity of DNMT3A and DNMT3B [24], both mechanisms result in tumor suppressor gene silencing. These methyltransferases are frequently up-regulated in lung cancer, and associated with poor prognosis [60]. The lncRNAs are involved in epigenetic regulation by mediating chromatin modification and DNA methylation. They also play a role in transcriptional regulation through modifying protein-DNA interactions by binding to transcription factors to facilitate their interaction with DNA to repress or activate mRNA, and post-transcriptional regulation by mRNA processing, as well as direct protein interactions to regulate protein (post-)translational modifications [61].

Category 4. Transcription factors are proteins binding DNA and regulating gene expression. They can form functional communities called transcription factor networks that regulate particular genes. For example, tripartite motif-containing protein 24 (TRIM24) is node of protein interactions, a promiscuous protein, with forty-four interacting partners, has a wide variety of functions, including as a ubiquitinase, a histone reader and a co-regulator of nuclear receptor-regulated transcription [62]. TRIM24 negatively regulates p53, a tumor suppressor, interacts with NRs, and directly associates with chromatin via its plant homeodomain (PHD)-bromodomain. As TRIM24 is a node of such an extended network, it has to be regulated precisely in order to avoid severe diseases, its knockout causes hepatocellular carcinoma, yet its overexpression leads to a poor prognostic breast cancer [62]. Nuclear receptors (NRs) are also important transcription factors that regulate gene expression upon binding to the specific ligand [63]. This receptor family includes intracellular steroid hormone receptors, among others. For example, estrogens are steroid hormones that act on nuclear receptors, namely human estrogen receptor α and β (hERα,β). These receptors act as ligand-activated transcription factors, upon estrogen binding, the receptors dimerize and bind to estrogen response elements (EREs), located at the promoter site of transcriptionally active genes [63–65]. Interestingly, not every gene contains an ERE sequence that is regulated by ERs, which necessitates distinct modes of endocrine action. They can modulate the function of other transcription factors, through protein–protein interactions, as non-genomic actions, moreover orphan nuclear hormone receptor SF-1, can serve as a direct binding site for hERα, but not hERβ [64].

Category 5. Complexes of chromatin remodeling and co-regulators. Covalent modification of DNA (Category 1), like methylation is fundamental in dynamic chromatin remodeling mechanisms [66,67]. Histone PTMs (Category 2) regulate transcription via controlling transcription factor (Category 4) accessibility [68]. The activity of transcription factors can be further modulated by hundreds of their own PTMs [69]. Histone PTMs can take their effects by influencing the overall structure of the chromatin via direct regulation of inter-nucleosomal contacts and controlling higher-order chromatin folding. They can also recruit specific chromatin modifiers [70,71] or remodeling enzymes that use the energy derived from ATP hydrolysis [70]. For example, acetylation of histone H2AX, member of the histone H2A family, is carried out by KAT5 (also known as Tip60) at the position H2AXK5, promoting H2AXK119 ubiquitination and enhancing chromatin remodeling [72].

Co-regulators are proteins that interact with transcription factors (nuclear receptors), to activate (co-activators) or repress (co-repressors) gene activity [63]. Co-activators participate in the regulation of a chromatin remodeling process when the condensed DNA becomes accessible for transcription. The co-regulators of the reverse process are the co-repressors. Co-regulators adopt various mechanisms of action. For example, they can play a role in the regulation of nuclear receptors, potentiating the activity of the receptor by switching between inactive and active states [63]. Leucine-rich motifs are frequent structural features of co-regulator molecules interacting with the ligand-binding domains of nuclear receptors. For example, the proline-, glutamic acid-, leucine-rich protein 1 (PELP1) [73] is a potential oncogene that interacts with ER, modulating its genomic and non-genomic functions, and its expression is misregulated in breast, endometrium and ovarian cancer progression [73]. Apart from being a co-activator for ER, PELP1 exerts its function as a co-repressor through association with HDAC2 and via deacetylation activity, suppresses histone acetylation and masks core histones from histone acetyltransferase mediated acetylation [74].

Histone readers and writers interact with (altered) histones, as was introduced in Category 2. For example, the switching defective/sucrose non-fermenting (SWI/SNF) and chromodomain, helicase, DNA binding (CHD) families partake in chromatin remodeling by interacting with the altered histone residues [72]. The SWI/SNF proteins have multiple bromodomains, enabling them to recognize and bind acetylated histone residues [72], and also have ATPase domains, typical of chromatin remodeling factors. The CHD proteins consist of tandem chromodomain and ATPase domains incorporated in a protein complex, called nucleosome remodeling deacetylase (NURD), which shows HDAC and chromatin remodeling properties [72].

Histone writer and eraser proteins can also function as co-regulators [75]. For example, histone acetyltransferases can weaken the interactions between the positively charged lysine side-chains of histones and the negatively charged DNA backbone phosphate groups by attaching an acetyl group and eliminating the positive charge, functioning as a co-activator [70]. The weakened interaction between the histone core octamer and the DNA backbone leads to destabilization of the local chromatin structure, which favors transcriptional activation [70]. On the contrary, HDACs, which remove the acetyl group, leave the lysine side-chain with a positive charge. In this way, they reinforce the local chromatin architecture, and are predominantly transcriptional co-repressors [70].

Histone writer and eraser proteins are often parts of large multi-protein complexes, and the composition of the complexes can determine the function of the histone writer or eraser [70]. Repressor element-1 silencing transcriptional factor (REST) has a co-repressor protein CoREST. If lysine-specific-demethylase 1 (LSD1) is complexed with CoREST, it demethylates H3K4me1/2, acting as a co-repressor, and if in complex with androgen receptor it demethylates H3K9, acting as a co-activator [70]. In contrast to histone acetyltransferases, histone demethylases show a greater substrate specificity, for example LSD1 requires a positively charged N atom, resulting a substrate specificity to H3K4me1/2 [70], interestingly the demethylation of H3K4me3 requires a jumonji domain, with a radical attack mechanism [70].

Histone writers can be also subjected to mutations, pathologic elevation or decrease in expression. Methyltransferase, acetyltransferase and kinase enzymes recruit additional chromatin modifiers and remodeler enzymes. Mutations of such enzymes frequently occur in diseases. For example, DNMT3A enzyme is mutated in myeloproliferative diseases and myelodysplasic syndromes [76]. Genes of KDM5A, and KDM5C code lysine-specific demethylase enzymes. KDM5A is mutated in acute myeloid leukemia [23] and plays a role in breast cancer formation [77,78]. KDM5C is mutated in renal carcinoma [23] and plays a role in acute myeloid leukemia [77]. The KAT3A enzyme is mutated in acute myeloid leukemia, acute lymphoid leukemia and transitional cell carcinoma of the urinary bladder, KAT3B is mutated in colorectal, breast and pancreatic carcinomas [23,78].

Category 6. Proteins with multiple functions. Regarding the extensive intertwined nature of epigenetic regulatory mechanisms, it is fairly common that some regulatory proteins play multiple roles in the epigenome, discussed in the previous Categories, respectively. An example of TRIM24 was mentioned in Category 4. Another example, the bromodomain PHD finger transcription factor (BPTF), is a nucleosome-remodeling factor subunit protein, which functions as a transcription factor. If amplified, it is prognostic for primary breast cancer [79]. At the same time, BPTF PHD finger is a histone reader sensitive to the state of methylation of the histone tail [80], which underlines that the activity of transcription factors is modulated by PTMs. Finally, we mention a protein coded in the alpha thalassemia/mental retardation X-linked (ARTX) gene. The protein has an N-terminal ATRX-DNMT3-DNMT3L (ADD) domain that binds histone H3 tail, and a C-terminal domain that is an ATP-dependent chromatin remodeling domain. The ADD domain has a PHD and a GATA zinc finger, the latter type of domains named after the specific binding of a DNA sequence. Interestingly, the ADD domain recognizes H3K4me3 with an atypical binding pocket at the interface between the GATA and PHD fingers [81]. Binding of ADD to histone H3 is facilitated by the recognition of methylated H3K9me3 (−12.2 kcal/mol), with an almost doubled binding enthalpy if compared to unmethylated H3K9me0 (−6.1 kcal/mol), indicating that H3K9me3 recognition is an enthalpy driven process [81] (see also Section 2.2.1). Unusually, the positive charge of the trimethyllysine is not accepted by an aromatic cage, but rather only one aromatic sidechain during the recognition by ADD [81] (see also Section 2.1.2). DNMT3L protein recognizes unmethylated histone H3 tail, and induces DNA methylation by DNMT3A2, establishing methylation patterns for heritable silencing and inactivation of the X chromosome in females [82].

In the previous paragraphs, the molecular background of epigenetic regulatory mechanisms was briefly sketched. The discussion was limited to only a segment of the most important molecules and a few examples. Exploration of the full proteome and interactome of the epigenetic universe seems a fairly demanding mission. However, various experimental and theoretical approaches have been adopted to answer this challenge. The next Sections survey recent approaches and selected contributions to the development of structure–activity relationships (SARs) of the epigenome.

2. Experimental Approaches

Exploration of molecular pathomechanisms of diseases of epigenetic origin and the discovery of new drugs require the determination of molecular structure, binding, and activity. Such experimental measurements are primary resources of new data for building SARs, and are also used for validation of computational approaches [83] of structural biology and drug design (Section 3).

2.1. Molecular Structure

The determination of three-dimensional structures of biomacromolecules of the epigenome is necessary for the precise description of their interactions and function at the atomic level. The technical breakthrough and first protein structures solved by X-ray crystallography date back to the previous century [84,85]. The technique requires expression, purification, and crystallization of biomacromolecules at a relatively large quantity and works typically on globular structures [86] neatly packed in the crystal lattice. Nuclear magnetic resonance (NMR) spectroscopy has started to supply structures for the Protein Databank (PDB, [87]) some decades ago. Beyond a static snapshot, NMR techniques also provide atomic resolution details on molecular dynamics of various systems including intrinsically disordered proteins [88,89]. However, the maximal measurable system size in NMR (ca. 35 kDa) is smaller than that in X-ray crystallography. Since the Nobel prize in 2017 [90], cryo-electron microscopy has been highlighted as an indispensable source of atomic resolution structures of the largest biological units.

2.1.1. Trends

To satisfy the above-mentioned need for establishing new SARs, numerous biomolecular structures of epigenetic importance have recently been deposited in the PDB. A quick search of the PDB results in more than four thousand entries, and more than half of these entries were deposited in the past decade. The corresponding statistics are presented in Figure 3 with a general overview of the trends of experimental structure determination in the epigenome. The statistics are based on the counts of PDB structures relevant to Categories i–iii in Section 1. In general, X-ray crystallography is the oldest and most wide-spread technique, and it has a leading role (Figure 3A) in the determination of structures of the epigenome, as well. If considering the types of biomacromolecules, histone-containing structures form the most abundant group. The number of such entries shows a dynamic increase (Figure 3B) in the past ten years. This trend reflects the growing efforts on solving the "histone code" and exploration of the effects of PTMs (Category ii in Section 1).

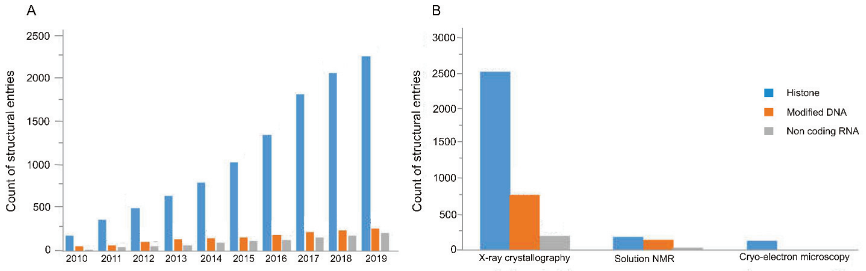


Figure 3. Trends of experimental structure determination of representative macromolecules of the epigenome. (**A**) The count of structural entries in the Protein Databank (PDB) per year (cumulative plot). (**B**) Distribution of entries of 2019 in (**A**) grouped by the main experimental techniques. The plots are based on a search in the PDB using key words 'histone', 'modified DNA' and 'non-coding RNA', which also involves 'siRNA', 'miRNA', 'lncRNA' were used in PDB. Accession date: 6 April 2020.

The above trends of statistical figures are reflected in the progress of structure determination of important biological units such as nucleosomes. As it was discussed (Figures 1 and 2), nucleosomes are the core units of the chromatin, and central scenes of the epigenome. Thus, the determination of their atomic resolution structure is of utmost importance. The first X-ray crystallographic measurements of the nucleosome date back to 1984 and confirmed the disk-like shape of the core particle at a 7 Å resolution [91]. There was constant progress towards the atomic level with a resolution of 2.8 Å in 1997 [29]. In 2002, the crystallographic structure was solved at 1.9 Å [48] with the whole histone H3 protein (PDB code 1kx5, Figure 2).

While the nucleosome was solved by X-ray crystallography, the determination of important functional assemblies, such as nucleosome-reader complex structures remained extremely challenging, and necessitated the use of cryo-electron microscopy [92–94] (next paragraph). In solution NMR studies [80,95], the terminal peptide tails of histones have mostly been captured in their complexes with reader proteins. Similarly, X-ray crystallographic entries contain only part of the nucleosome-reader complexes. In many cases, structures of only the reader-bound terminal peptide tails (Figure 2) of histones have been captured [94,96]. Atomic level assignation of the DNA segments and interacting histone core sequences is often missing too.

Following the new trends of recent years, nucleosome structures have also been determined by cryo-electron microscopy [93,94]. Although cryo-electron microscopy is still not as wide-spread as X-ray crystallography (Figure 3), it helps to overcome size and shape limitations [97,98], and has received a spotlight in the past decades. In the epigenome, cryo-electron microscopy has a remarkable role in the determination of multi-molecular units, such as the above-mentioned nucleosome-reader complexes. Determination of full structure of these complexes is of particular importance for exploration of the effect of PTMs on the nucleosome and development of SARs unraveling the histone code. Cryo-electron microscopy provides pioneering examples for the solution of full nucleosome-reader complexes.

For example, in a recent study of Wagner et al. [94] the whole triad of the DNA–histone–reader complex was solved using cryo-electron microscopy (Figure 4). The structure contains the switch/sucrose non fermentable (SWI/SNF) chromatin structure remodeling complex with a subunit called nuclear protein STH1/NPS1, which is multi-functional as a histone H4 reader, as well. The interaction of the nucleosome core histone octamer, the DNA double helix wrapped around it, the protruding histone H4 tail and the reader protein are all visible, providing indispensable details for the development of SARs. The first electron microscopic map supplied the whole complex at a 15 Å resolution, and further refinements were possible for the nucleosome. Parts of the whole complex were also rigid body fitted from other PDB structures, which included both X-ray crystallographic and cryo-EM structures. This study [94] applies complementary experimental and theoretical methods such as map alignment, rigid body fitting, homology modeling and real space refinement with secondary structure restraints to solve the complex of more than one thousand kDa molecular weight.

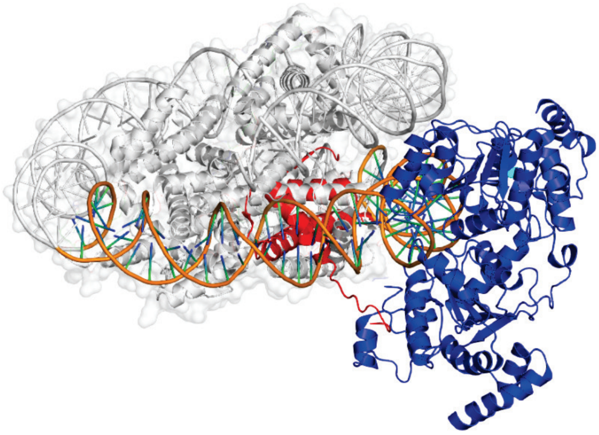


Figure 4. The ternary complex of STH1/NPS1 nuclear protein (blue cartoon), a histone reader, the DNA (orange cartoon), wrapped around the nucleosome and histone H4 (red cartoon), that is buried in the nucleosome. The non-interacting parts are represented in grey cartoon. The figure was rendered by PyMol [47] using PDB structure 6tda [94].

Beyond the static structures discussed in the previous paragraphs, development of up-to-date SARs necessitates the exploration of molecular dynamics of biomolecules of the epigenome. Recent studies [99,100] highlight the necessity of such information even at the level of molecular design. For probing structural dynamics of the nucleosome there are appropriate experimental methods like fluorescence resonance energy transfer (FRET) and nuclear magnetic resonance (NMR) spectroscopy [101]. FRET is well suited for the investigation of dramatic conformational and compositional changes. For example, FRET grants access to the measurement of nucleosome unwrapping equilibrium [101]. The equilibrium occurs between fully wrapped and partially unwrapped states of the nucleosome, with the binding of site-specific DNA binding proteins, the equilibrium shifts towards the unwrapped state, explaining the increasing accessibility of the DNA [102].

Apart from dramatic conformational changes, subtler changes of molecular conformations also occur as a part of nucleosome dynamics. NMR allows the measurement of protein dynamics and interactions at atomic level, even is disordered regions and transient complexes. NMR experiments probe how macromolecules shift between conformational sub-states in solution [103,104]. A special type of NMR, namely methyl-transverse relaxation optimized NMR (methyl-TROSY) is more suitable for the investigation of subtle dynamic changes, which is more typical for the histone tail-reader protein interactome [101]. The NMR linewidths of the base imino protons of DNA provide an informative insight into base pair opening dynamics [105]. A broader line shows reduced base pair stability and increased base pair opening rates, for example, oxidation of guanine led to line broadening of guanine base imino protons, and methylation of cytosine resulted in imino proton line narrowing [105], indicating that cytosine methylation stabilizes the DNA duplex. Solid state NMR can determine the binding sites on the nucleosome surface, and demonstrate the dynamic nature of the N and C terminal tails of histones within the core octamer [106]. These terminal ends are DNA bound and rich in PTMs, their structural dynamics is exploitable by histone reader, writer and eraser proteins [106]. Solid-state NMR does not have an intrinsic size limit, larger chromatin substrates can also be accessed, and is a complementing method to cryo-EM and X-ray crystallographic settings, when the plasticity of histones is thought to play a role or smaller proteins are observed [106].

Approaches combining different techniques such as small angle X-ray scattering, solution NMR spectroscopy and molecular dynamics were successfully applied to study the ubiquitin-like, containing PHD and RING finger domains, 1 (UHRF1) protein and its tandem Tudor domain–plant homeodomain (TTD-PHD) histone reader module [107]. UHRF1 is expressed in various cancers, being a promising target in antitumor therapy, a known small molecule acts by binding to its TTD. The study identified a novel antagonistic approach to UHRF1 function, through the allosteric disruption of the co-operative binding mode of its TTD-PHD module [107].

2.1.2. Challenges

In addition to the large size of nucleosomal assemblies discussed in the previous Sections, structure determination faces other challenges due to complexity of the histone code system, conformational and functional diversity of the epigenome. Both experimental and theoretical (Section 3) approaches face the challenges described in detail in the forthcoming Sections.

The Size and Complexity of the Histone Code System

The histone code originates from PTMs on amino acids of histone proteins (Category ii, Section 1). It is even possible, that every single amino acid of a histone tail has a specific meaning and place in a peculiar vocabulary [36]. The code system has astronomical proportions if considering the large number amino acids and types of modifying groups involved (Figure 2). The number of possible codes can be illustrated using a specific case of methylation of H3 lysine residues. Histone H3 is known to be methylated at nine lysines, K4 [31,36–40], K9 [31,37–40], K14 [31,38], K18 [38,46], K23 [31,38,40], K27 [31,36–40], K36 [31,37–40], K56 [38,45] and K79 [31,37–40,46] (Figure 2). A single lysine side-chain can accept a maximum of three methyl groups, and there is the non-methylated, native amino acid resulting in four possible marks per residue. This means 49 (262,144) possible variations, only for lysine methylation of histone H3 not including, e.g., lysine ubiquitination, acetylation, arginine methylation, PTMs on serine, tyrosine, and other histones. Thus, the amount of PTMs of the histone code is almost uncountable, involved in many, if not all, DNA-templated processes [36,108]. Recent studies usually accumulate more variations than older works, highlighting that exploration of new PTMs is still an evolving field of epigenetics.

Besides its enormous size, the code system is further complicated by the yet unpredictable distribution of the different PTM types. Some amino acids like K4 even tend to accept multiple modification marks, resulting in different binding schemes. There are PTMs, missing from some histone types. Moreover, there are different histone reader, writer and eraser proteins, that can assess these altered amino acids in a wide variety of conformational possibilities. Lysine methylation, acetylation and ubiquitination appear from the 4th position up to the 123rd position on the histone chain, arginine methylation mostly occurs on the lower positions of the tail, while serine and tyrosine phosphorylation is typically closer to the N-terminal end in histone H3 and further in H2AX [40].

The complexity of the histone code is further increased by networking and cross-talk of the codes [109]. Some networking modifications enhance, while others inhibit the functions of others [40,110].

One study [70] proposes five mechanisms of PTM cross-talk, to which an extra level of complexity is ascribed over the histone code, for fine tuning of the overall control of the chromatin structure. Lysine residues might be target of various modifications, such as acetylation, methylation or ubiquitination and these agents might compete with each other. In Saccharomyces cerevisiae, methylation of H3K4 is dependent on the ubiquitination of H2BK123. The phosphorylation of H3S10 disrupts the binding of heterochromatin protein 1 (HP1) to H3K9me2/3, which would occur in the absence of the phosphorylation. In yeast, the FK506-binding protein 4 (scFpr4) proline isomerase catalyzes the interconversion of the peptide bond of H3P38, which interferes with the methylating ability of histone-lysine N-methyltransferase, H3 lysine-36 specific (scSet2) on H3K36. Finally, PHD finger protein 8 (PHF8) binds to H3K4me3 with its PHD finger, and this interaction is stronger when H3K9ac and H3K14ac occur at the same time [70]. Apart from these subtle mechanisms, PTMs also play a role in chromatin remodeling by altering the physico-chemical properties of the nucleosome. As it was discussed in Section 1, acetylation marks on histones H3 and H4 weaken DNA-histone interactions, enhances formation of accessible DNA, and transcriptional activation [41,70,111]. This phenomenon is mostly additive, with the more acetyl marks on the histone, the DNA becomes more accessible. Simultaneous acetylation of H4K79 and H3K122 has an amplified destabilizing effect on the nucleosome when compared to a single acetylation mark [41]. This means that additional PTMs, acting as a network, help each other to pose regulatory effects on the nucleosome.

The above-discussed extremely large size and high complexity of the histone code system is based on numerous corresponding complexes of the participant macromolecules (DNA, histones, effectors, etc.) at the atomic level. Experimental structure determination of such an infinite number of complexes would be an impossible undertaking, even with high throughput methods [112]. As experimental structure determination methods can clear up only tiny pathways in this jungle of the epigenome, involvement of fast, complementary theoretical approaches is necessary to speed up the exploration of new structures (Section 3).

Conformational Diversity and Water-Mediated Weak Interactions

Macromolecules of the epigenome, especially linear peptide tails of histones and RNAs [113] often adopt various binding conformations imposing further challenges on structure determination methods. Histone tails are linear structures, which are seldom compatible with X-ray crystallographic approaches [114] since these experimental methods are better at handling globular structures that can be crystallized [114,115]. Such linear peptides are better accessed by X-ray crystallography, when they are a part of a globular structure, like a nucleosome. In this case, the structure of the whole histone tail can be assigned [48]. Similar to histone tails, RNAs of the epigenome also pose a challenge for both X-ray crystallographic and solution NMR methods, due to their great flexibility. Size is also a limiting factor in their case [116,117]. The highly complex interactome (see previous Section) around the chromatin involves dynamic and flexible parts, exacerbating the difficulty of unraveling the machinery [112,118]. These difficulties often lead to experimental structures of complexes with N-terminal histone peptides of only 10–15 amino acids [81,95]. If the length of a histone tail exceeds 30 amino acids, and peptides of this length tend to loosely stick to the surface and give response signals that are non-specific for the original peptide [119]. In these cases, the N-terminal end of the peptides usually hang out to the bulk, and do not have any (specific) interactions with the partner protein (reader, Figure 5). This situation is experimentally challenging, as the peptide tails have dynamically changing positions, great flexibility, and conformational uncertainty as a result of their interactions with bulk water molecules in continuous thermal motion.

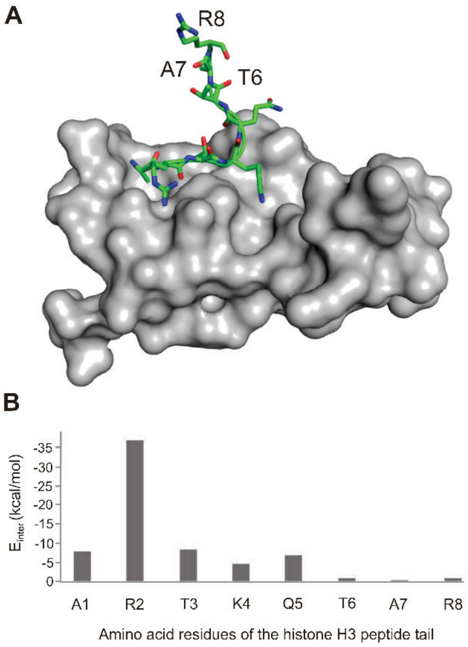


Figure 5. The structure (**A**) of the UHRF protein PHD finger (grey surface) in complex with histone H3 peptide tail (sticks, PDB ID 3sou). The figure was rendered by PyMol [47]. (**B**) Per residue peptide–protein interaction energies (Einter, bottom) were calculated after energy-minimization of the crystallographic complex. Einter values were calculated as a sum of Lennard-Jones and Coulomb interaction energies as described previously [99]. After the first 5 amino acid residues of the histone peptide, Einter diminishes as the last three residues interact with the bulk.

The histone H3–autoimmune regulator protein (AIRE) complex is a representative example of the above-mentioned situation. Most of the interactions take place between the first five amino acids of the N-terminal tail of the histone hampering the determination of the rest of the histone tail either experimentally or theoretically [99]. Histone recognition by a bromo-, homeo-, chromodomains or transcription factors involves only a shallow depression of the protein surface [91,96–98]. These flat binding surfaces often result in low interaction energies at most of the binding residues. As another example, the complex of the UHRF (see also Section 2.1.1) protein PHD finger and histone H3 N-terminal peptide tail is featured in Figure 5. The atomic resolution structure ([120], Figure 5A) and the corresponding [99] distribution of per-residue interaction energy values (Einter, Figure 5B) show that starting from the sixth amino acid, physical interactions tend to cease. This is a challenging situation in the investigation of histone tail binding to reader proteins both experimentally and theoretically. From an experimental perspective, the non-interacting part of the histone peptide tail moves dynamically in the bulk solvent, and it is hard to capture. On the other hand, fast computational docking approaches try to find the bound peptide position with the best possible interaction energy, creating non-existent interactions (mis-docked conformations, see also Section 3).

As a consequence of the shallow binding surfaces and few contact histone residues, the total binding affinities of (modified) histone tails are often limited to a micromolar range [121–125], indicating relatively weak complexes.

Histone–partner interactions are also affected by structural water molecules located in the binding interface. However, the determination of hydration structure is challenging in many cases [126,127]. There is often a water network formed in the interface increasing the complexity and stability of the interactions. Disruption of the hydration network can lead to complex instability. For example, the interaction of the histone reader death-associated protein 6 (DAXX) to histone H3.3 N-terminal peptide tail was investigated with a special highlight on interacting water molecules [128] and their networking [127]. It was found experimentally that if one water molecule was displaced from the interfacial hydration network by introducing an active site mutation, the binding affinity was reduced by 50%. A computational investigation [127] further analyzed the networking of interfacial water molecules. The complete interfacial hydration networks were produced, using a molecular dynamics (MD)-based determination of the complete hydration structure by MobyWat [126]. In the mutant structures, important water nodes changed their positions or disappeared from the static core of the hydration network of the wild type. In agreement with the experimental results [128], the networking study [127] found that in the mutant system, the static core of the interfacial hydration network has disintegrated into a dynamic hydration network, explaining the reduced binding affinity.

Functional Diversity of the Histone Code

Histone reader, writer and eraser proteins are often promiscuous [62,70], their substrate specificity may depend on the complex they participate in [70]. Histone reader proteins are structurally diverse including plant homeo-, chromo-, bromodomains, Tudor, ADD, WD40 and PWWP modules [40]. On the other hand, a histone holding a PTM variation (a code) can also interact with multiple readers. The PTMs of the H3K4 residue is recognized by fourteen different reader proteins [40], including PHD finger containing proteins, recombination activating gene protein 2 (RAG2), inhibitor of growth protein 2 (ING2), BPTF (see in Category 6 of Section 1), AIRE, Tudor domain containing protein, SAGA complex associated factor 29 (Sgf29) and chromo domain containing proteins Jumonji domain containing 2A (JMJD2A) and chromodomain helicase DNA-binding (CHD) [129,130]. Another example is H3K9me3, which is regarded as a general transcriptional repressive mark [40], influencing a wide variety of cellular functions. Therefore, associating a distinct function with a PTM is challenging.

The functional diversity of the histone code is further increased by chromatin-associated protein complexes often containing multiple domains with different functions, as the same protein complex can include both a reader and a writer domain. For example, nucleosome acetyltransferase for H4 (NuA4) and Saccharomyces cerevisiae reduced potassium dependency 3 small (Rpd3S) protein, a HDAC share the same chromodomain containing subunit Esa1p-Associated Factor (Eaf3). Whereas Eaf3 is a histone reader domain [40] identically present in both NuA4 and Rpd3S complexes, NuA4 also has a histone writer domain, and Rpd3S contains a histone eraser domain, as well.

Investigations of the effects of histone PTMs are complicated by their different accessibility [40] as a free peptide or under physiological conditions, embedded in the nucleosome. For example, in the H3K79me-reader interaction, flanking residues of H3K79 take different positions in their nucleosome-bound and free states. When wrapped in the nucleosome, there are structural constraints of the flanking residues, hindering the recognition of histone peptides by reader proteins [40].

Lysine acetylation PTMs are recognized by bromodomains with wide pockets, and by tandem PHD fingers with shallow binding pocket. Other residues surrounding these PTMs tend to form less characteristic contacts with the surface, resulting in a decreased substrate specificity [40]. On the other hand, recognition of lysine methylation by histone (de)methylases require higher substrate specificity [40]. Addition of a methyl mark to the lysine residue results in a positive charge, and increases hydrophobicity at the same time, which can be recognized by an aromatic cage (Figure 6C). Thus, binding surfaces of lysine methylation marks are similar to each-other [40]. At the same time, the non-methylated state of a lysine residues also acts as a coding variant as methylated lysine residues are not recognized by readers specific for non-methylated lysines, and vice versa [40].

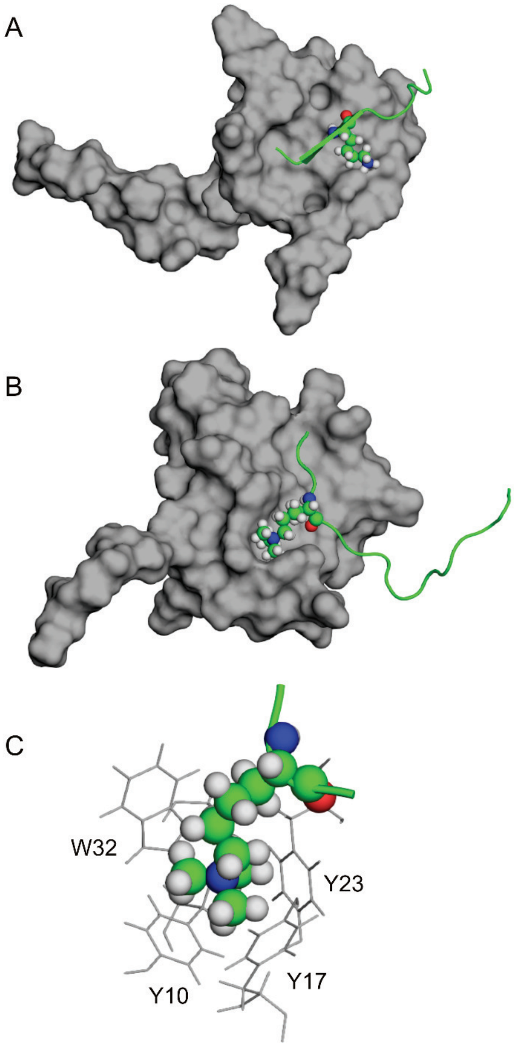


Figure 6. Binding of non-methylated (**A**) and tri-methylated (**B**,**C**) histone peptide tails (cartoon coils) to PHD fingers (grey surface). (**A**) The non-modified histone peptide tail with the H3K4me0 residue (spheres) binds the shallow surface of AIRE PHD finger (PDB ID 2ke1). (**B**) The histone peptide tail tri-methylated at H3K4me3 (spheres) binds to the aromatic cage of BPTF PHD finger (PDB ID 2fuu). (**C**) Close-up of the aromatic cage (BPTF PHD finger residues are in grey sticks) in complex with the tri-methylated lysine residue (H3K4me3, spheres). The figure was rendered by PyMol [47].

For example, the PHD finger of AIRE protein binds a non-methylated lysine residue at position H3K4me0 [130] (Figure 6A). With three methyl groups attached to the K4 side-chain (H3K4me3), the binding completely diminishes according to ITC measurements [130]. In the case of AIRE PHD finger the recognition occurs on a flat binding surface of the protein, and this PHD finger does not have an aromatic cage. On the contrary, the PHD finger of the BPTF protein has an aromatic cage (Figure 6B,C), which can recognize the positively charged, trimethylated lysine residue of H3K4me3 PTM [80] (see another example in Category 6 of Section 1). This example nicely illustrates the binding diversity of histones, as the same PTM (H3K4me3) has radically different binding affinity to different readers.

2.2. Binding Affinity and Biological Activity

Epigenetic events manifest themselves as (patho)physiological activities at a systemic level. Structural results reviewed in the previous Sections demonstrated that formation of complexes of two or more molecules provides the background of such activities at an atomic level. Complex formation assumes that the partners have several intermolecular interactions (Figure 5) and high binding affinity to each other. Whereas structural description of intermolecular interactions is rather challenging and costly (Section 2.1.2) at an atomic level, measurement of binding affinity or in vitro activity is often less demanding, especially in the cases of routine assays and kits briefly mentioned in the next Sections. However, the large size of the epigenetic interactome, especially of histone PTMs (Section 2.1.2) indicates that there are hardly enough resources to measure all corresponding in vitro affinities. The highest level, in vivo activity measurements are again rather expensive, require special conditions of animal keeping and often limited by ethical concerns, as well.

2.2.1. Binding Affinity

Isothermal titration calorimetry (ITC) measurements are often performed [131–134] to gain an insight into the effect of histone PTMs and different states of methylation or acetylation on the side of the ligand. Besides investigation of PTMs, ITC can supply binding thermodynamics parameters for any binary system [98] via the measurement of the generated or absorbed heat during the titration of the solution of one partner with the other. ITC is a gold standard for determination of the full thermodynamics profile of complex formation including binding free energy (ΔG), enthalpy (ΔH), entropy (ΔS), and the stoichiometry (n) of the complex. Equilibrium dissociation constant (Kd) can be obviously calculated from ΔG and temperature data, as well.

ITC is a primary tool for finding new ligands and optimization of lead compounds in drug design. Appropriate (large negative) ΔG of ligand to a target molecule is a necessary, yet not sufficient requirement for pharmacological efficacy. Different targets might require distinct thermodynamic binding profiles to show biological effect upon interaction with their ligands [135]. During ligand optimization enthalpy and entropy-based approaches are applied [136–139] during the early stage of the optimization. An enthalpy excess can be introduced by additional hydrogen bonds to the interaction, while entropic optimization typically occurs during the later stages by for instance rigidifying the ligand in a bound conformation [135]. HIV reverse transcriptase inhibitors (e.g., etravirine) were subjected to entropic optimization, to avoid viral resistance upon mutational changes, by allowing a high residual mobility to be able to acquire multiple binding modes [135,137,138]. The design of high-affinity adaptive inhibitors can be achieved through engineering their vital interactions for affinity and specificity with conserved regions of the target. In addition, at moieties of the ligand that will most likely face rapidly mutating sites of the virus, flexible asymmetric mutations are introduced [138]. In these cutting-edge strategies, ITC is an indispensable technique, providing both ΔH and ΔS components of the overall binding affinity (ΔG).

In the context of epigenetics, ITC is an excellent tool to study the interactions of wild type (see also in Category 6 of Section 1) and mutated readers such as the PHD finger of AIRE and different PTM states of the N-terminal histone tail [95,130,140]. ITC investigations of such studies provide the full binding thermodynamics profile (ΔG, ΔH, ΔS, n), and allow measurement of effects of PTMs. For example, AIRE does not have an aromatic cage to accept H3K4 methylated lysine binding (Figure 6A); in agreement with this, H3K4me0 bound with the greatest affinity, and trimethylation of H3K4me3 caused a lack of binding to the AIRE reader protein [130]. A common way to perform a mutagenesis assay is to change the residues of interest to inert alanine residues (alanine scan), either on the side of the ligand or the target. For example, if the binding affinity of H3K4me0 to AIRE-PHD1 is compared to the binding affinity after the mutation of an aspartic acid residue to an alanine at the binding site of the target, the measured ΔH_o of binding of the same H3K4me0 is halved [130]. Beyond local changes on ligand binding, mutations of amino acids may alter function of the target protein by changing its overall integrity and global folding, as well.

In another study, [55], the binding of H3K4me3 was also investigated by ITC analysis. Binding of H3K4me3 and H3c4me3 (where c refers to the neutral carba analog) to five specific H3K4me3 readers, the PHD domains of JARID1A, BPTF, TAF3, the Tudor domains of the Royal Family of SGF29 and JMJD2A was studied. All the readers have aromatic cages for specific trimethylated lysine binding, but all have a different architecture of the typical motif. It was shown, that H3K4me3, which is positively charged binds 2-33-fold stronger, than the neutral H3c4me3 to readers that contain a Trp residue in their aromatic cage [55]. Interestingly, the association of H3K4me3 is more favorable enthalpically, but less favorable entropically, compared to the association of H3c4me3, in the same aromatic cages [55]. The two histone peptides bind with indistinguishable thermodynamics of associations to half-aromatic cages, indicating little to no contribution from cation-π interactions to the binding [55]. Aromatic cages containing tryptophan residues show stronger cation-π interactions binding to quaternary ammonium ions, if compared with aromatic cages containing phenylalanine and tyrosine residues [55].

ITC is also applicable for the thermodynamic analysis of large complexes of the epigenome. For example, binding of aprataxin and polynucleotide kinase like factor (APLF) to histone dimers and tetramers H2A-H2B and (H3-H4)2 was investigated in a study [141]. It was found that both histone systems bind the APLF reader protein with micromolar affinity, both are enthalpically favorable interactions, yet the binding of (H3-H4)2 is entropically unfavorable, which might explain the difference in their Kd values. ITC did not detect additional lower affinity binding modes, resulting in a stoichiometry of $n = 1$ in both cases [141].

Surface plasmon resonance (SPR) techniques are essential for high-throughput probing of biomacromolecular interactions. Kd is the main outcome of SPR which generally correlates well with the Kd from ITC measurements [119]. In contrast with ITC, the SPR measurements also provide reaction kinetic information of association and dissociation rates which can be useful in estimation of the kinetic stability of a drug-target complex [142]. Application of SPR yielded excellent comprehensive studies. For example, 125 types of modified histone (H1-H4) peptides with different PTMs were investigated [119] in combination with 8 histone reader proteins, resulting in one thousand pairs of interactions. It was discovered that KDM5A (also known as JARID1A, see in Category 5 of Section 1) interacts with H3K4me3 specifically [119]. KDM5A also interacts with human estrogen receptor and plays a role in osteogenesis. The study also showed that heterochromatin protein 1, important in the DNA repair after UV-induced damage, interacts with H3K9me3 [119]. Another study [143] used 204 proteins from either the Royal Family, PHD, bromodomains or CW domains and subjected them to SPR investigations with three specific histone modifications (H3K4me3, H4K5acK8ac, H3S10ph). The results were confirmed by ITC. It was found that the Tudor domain of echinoderm microtubule-associated protein-like 1 (EML1) binds to H3K3me3 with a greater affinity than to H3K36me3 [143]. EML1 is associated with Usher syndromes, which is a disease that eventually progresses in the whole brain. SPR was also used [50] to investigate the dependence of histone H3 binding to WDR5 on the methylation state of H3K4. The peptides were immobilized in the analyses, wild-type and mutant proteins were also assessed. H3K4me2 shows the strongest binding to WDR5, mono- and tri-methylated peptides bind seven and eight-fold weaker. Interestingly, H3K4me2 has both the smallest association and dissociation rate, if compared with the otherwise methylated H3K4 peptides. The small k_{on} rate indicates a slower approach to equilibrium and k_{off} corresponds to a sluggish decay of bound peptide signal when dissociating. This might be explained by an extended intracomplex interaction formed by H3K4me2 if compared with other modified H3K4 peptides, involving a hydrogen-bonding network between the ligand, water molecules and a backbone amino acid residue of the target [50]. For all histones, a micromolar Kd was measured, indicating a relatively weak binding affinity, compared to for example, strong, small molecule inhibitors [50].

Fluorescence spectroscopy is a versatile and sensitive method, used for investigations of a wide range of interactions including histone binding [95,130] and nucleic acid modifying enzymes [144]. Fluorescence spectroscopic determination of Kd of histone–reader complexes often completes PTM studies to affirm the findings of ITC measurements [95,130].

The methyl-CpG-binding domain protein 3 (MBD3), a DNA methylation reader was investigated in living cells, under hypoxia and decitabine treatment [145] by fluorescence correlation spectroscopy. The changes in the environment alter the fluorescence of the reporter (green fluorescent protein), which is suitable for detecting conformational changes at the timescale of milliseconds [144]. By monitoring the dynamics of the MBD3 protein, a fast diffusion in the nucleosome was observed, showing a form of demethylation that is independent of DNA replication [145]. Fluorescence correlation spectroscopy also contributed to identifying hypoxia sensitive cells and the real time follow-up of demethylation, which occurs in context with the hypoxia [145].

Fluorescence resonance energy transfer (FRET) assays are also popular tools to study, e.g., estrogen receptors and their coactivators [146], DNA bending by charge variant bZIP proteins [147], and transcription factor binding kinetics to nucleosomes and DNA [148].

Besides determination of atomic resolution structures (Section 2) NMR techniques are important in ligand-based binding assays and hit generation [149]. Such measurements are based on the spectral differences of hits and non-binding ligands. Target immobilized NMR screening can use the same target sample for a considerable number of ligands [149]. A study [141] of APLF nicely shows how NMR can supply both binding and structural information in the same experiment. APLF is a DNA repair factor with histone chaperone activity of its acidic domain, which uses two aromatic side chain anchors towards histones H2A and H2B. NMR titration experiments with stoichiometric addition of reagent solution (APLF acidic domain) to H2A and H2B were performed and peak intensity ratios and residue-specific chemical shift perturbations were collected from N-TROSY spectra. 2D NMR line shape analyses of the data resulted in Kd and also kinetic (k_{off}) values, and a distinction was possible after fixing the initial values, to investigate the formation of a secondary complex, a secondary binding event with lower affinity [141]. With the calculated chemical shift perturbations, key residues of the complexes were also identified, allowing assignation of the structural origin of the binding affinities.

Inhibition assays are common, fast methods for estimation of binding affinities of inhibitors to enzymes. Such assays often produce the half maximal inhibitory concentration (IC_{50}) values which can be related to the thermodynamic inhibition constant (Ki) [150]. Thus, IC_{50} is system-dependent, and not directly applicable instead of Ki (Kd). It can be applied for fast comparison or screening of a series of ligands according to their inhibitory effect on the same target enzyme. However, this level of information is often sufficient for further investigations or drawing conclusions. For example, a small molecule inhibitor was tested in vitro in mantle cell lymphoma models [151]. Protein arginine methyltransferase 5 (PRMT5) is overexpressed in patient samples with mantle cell lymphoma. Small nuclear ribonucleoprotein Sm D3 (SmD3), is a protein involved in RNA splicing, and is methylated by PRMT5. Observing the methylation of SmD3 with biochemical assays, PRMT5 enzyme activity was measured. The study [151] suggests that observed antiproliferative effects were a direct result of PRMT5 inhibition by the small molecular inhibitor that showed an IC_{50} value of 22 nM. In general, lysine methylation is a better studied area in cancer pathogenesis than arginine methylation; however, this study [151] offers an insight into arginine methylation in cancer, and represents a validated chemical probe for further studying.

In many epigenetic studies, a combination of various experimental methods is applied resulting in a more complete and reliable picture of the interactome. For example, a study combined FRET and ITC to measure selective inhibition of HDAC isoforms [152]. Results of a set of experimental methods (in vitro binding assays, NMR binding, fluorescence titration assays, ITC, expression analysis and chromatin immunoprecipitation) provides a solid basis for further, structural investigations of, e.g., the effects of histone methylations [55,130] with theoretical approaches (Section 3).

2.2.2. Biological Activity

There is a wide range of methodologies for measuring the activities of biomacromolecules and their ligand partners. Whereas binding affinity information (Section 2.2.1) is very important for molecular engineering, activity measurements provide high level tests of the new molecule. Thus, affinity and activity data complement each other, and in many studies, both types of measurements are present for the same system. In general, in vitro tests precede in vivo investigations, as the latter ones are rather costly, and therefore, they are applied mostly on a thoroughly screened, narrow set of compounds.

In Vitro Activity

In a recent review [153], various in vitro methodologies including fluorescent, electrochemical, and surface-enhanced Raman spectroscopy-based assays were featured for investigations of histone PTMs and histone modifying enzymes. Fluorescent assay can measure e.g., the de-acetylation of peptides by HDAC enzyme [154]. The activity of HAT, adenovirus E1A-associated protein (p300) can be detected by monitoring coenzyme A formation in an electrochemical assay [155]. The activity of histone demethylase 1 enzyme can be also measured by detection of product concentrations in surface enhanced Raman spectroscopy-based assays [156]. In the next paragraphs, selected assays with some recent applications are also reviewed.

In vitro DNA methylation assay measures the enzymatic activity of DNMT3A protein (see Categories 3, 5 and 6 of Section 1). After methylation of DNA by DNMT3A, the isotopically labeled methyl groups can be detected. The assay was applied for the investigation of the effects of histone lysine residue methylation on DNMT3A enzyme autoactivation in an interesting study [157]. Relative enzyme activities were measured, in the presence and absence of histone peptides H3K4me0, H3K4me3, and the catalytic domain (CD) of the DNMT3A enzyme. In the presence of histone H3K4me3 the DNMT3A protein preferred an autoinhibitory conformation, in which its ADD (Section 1) does not bind to the histone tail. The binding of histone H3K4me0 to the ADD domain of DNMT3A induced a conformational switch favoring the active form of the enzyme [157], allowing the formation of the DNA-CD interaction (Figure 7).

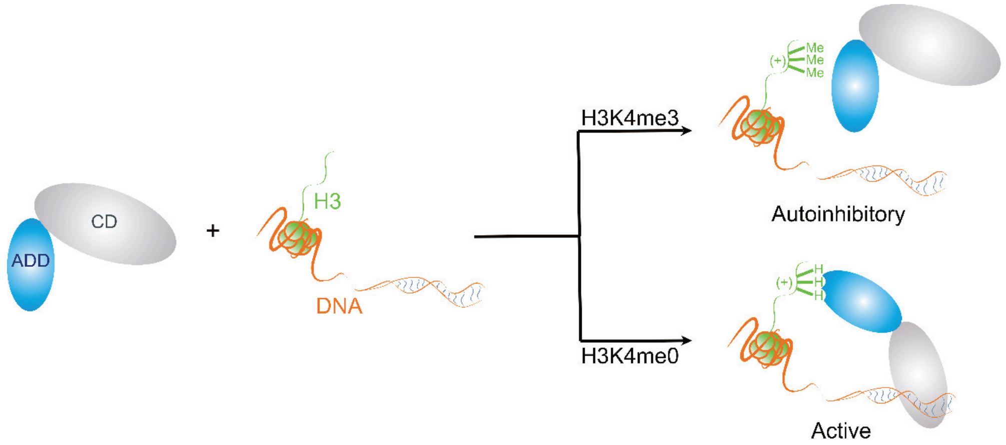


Figure 7. A schematic representation of the effect of H3K4me3 trimethylation on DNMT3A enzyme activity. In the presence of H3K4me3 PTM, the DNMT3A enzyme is in an autoinhibitory form. In the case of the non-methylated H3K4me0, the enzyme interacts with the DNA in an active form where the arrangements of its ADD and CD domains is different from that in the autoinhibitory form.

An activity assay can also be applied to uncover the mechanism of the enzyme involved. The histone-lysine N-methyltransferase SUV39H1, uses H3K9me1 as a substrate to create H3K9me3 [158]. This methylation leads to a condensed state of the chromatin (see Category 5 of Section 1), and the exact mechanism that lies beneath this phenomenon was investigated in the study [158]. The enzyme has a CD and an N-terminal end, which are also important for its function. First, in its free form the enzyme samples chromatin through its CD. Then recognition of H3K9me3 allosterically activates a chromatin binding motif to anchor the enzyme with the likely involvement of its N-terminal segment, promoting H3K9 methylation. An active site mutation resulted in the inactivity of SUV39H1. Disabling the active site of the CD of the enzyme led to a disruption of promotion of H3K9 methylation through the N-terminal segment. Accordingly, the enzyme mutant lacking the N-terminal end showed lower activity [158]. The addition of H3K9me3 peptide strongly enhanced enzyme activity, by a tenfold increase in maximum velocity, and a fourfold reduction of the substrate concentration required to reach half-maximal rate [158]. This finding underlined the allosteric activation by H3K9me3. The promotion of H3K9me3 was called spreading in the study [158], and it resulted in a spatial closure of the nucleosomes. In this study [158], causal relationship was established with respect to how a PTM is transferred as a function.

Colorimetric assays, sulforhodamine B (SRB), mitochondrial metabolic activity (MTT) and crystal violet (CV) are used to determine cell viability [159]. The SRB, MTT, CV and LDH assays were applied to measure the cytotoxicity of apicidin on human pancreatic cancer cell lines, Capan-1 and Panc-1 [159]. Apicidin is a HDAC inhibitor. HDACs catalyze the deacetylation of primarily lysine residues at the N-terminal tails of histones [159]. The HDAC enzyme family takes part in chromatin remodeling and modification of gene expression (see in Categories 1, 2 and 5 of Section 1), and in the consequent pathogenesis of various malignant diseases. Pancreatic cancer cell lines were cultured, grown and plated before the experiments. Apicidin, a HDAC inhibitor was used at different concentrations for incubation with cell lines. The effects of short duration and longer exposure were investigated, and non-treated cells were used as control. EC50 values of apicidin were measured by these assays, and a dose dependent cytotoxicity was detected after 24 h treatment. An increase in cytotoxicity and decrease in cell viability was observed after treatment with 100 nM or higher doses of apicidin [159]. Moreover, in pancreatic cancer cell lines, apicidin showed an initial antiproliferative effect before the onset of cytotoxicity [159].

MTT cell viability assay was used to measure the effectiveness of tamoxifen and anacardic acid on MCF-7 and T47D breast cancer cell lines [160]. In an effort to unravel disease pathomechanism, and find possible therapeutic targets, epigenetic-related markers were screened, including oxidative forms of DNA-methylation, histone modification and methyl-binding domains to identify H4K12ac and H3K27ac as potential epigenetic therapeutic targets [160]. Anacardic acid, a HAT (see Section 1) inhibitor reduces the levels of acetylated H4K12ac and H3K27ac [160]. Then, it was combined with tamoxifen, a widely used agent in the treatment of breast cancer. The cell lines were cultured and incubated, as a pre-treatment for MTT cell viability assays [160]. The assays were then performed by addition of anacardic acid and tamoxifen, and a second incubation time was introduced before analysis. A solvent, not containing any cells were used as background [160]. The combination of tamoxifen and anacardic acid resulted in a marked inhibition in cancer cell viability with an additional loss of FRET efficiency between ERα and histone acetylation marks. Such combined epigenetic and hormone receptor mediated pathomechanism of breast cancer results raises the possibility of a combined treatment targeting multiple pathways of the disease [160].

Similarly, cell viability was measured by MTT assay, after treatment with a HDAC inhibitor, on pediatric embryonal cell lines [161]. HDACs are often used to treat various malignant diseases (see in Categories 1, 2 and 5 of Section 1), screening of compound libraries for HDAC activity is an emerging area of drug development [161]. A potent novel agent, HKI46F08 was tested on pediatric neuroblastoma and medulloblastoma cancer cell lines [161]. Its EC50 value was in the range of 0.1–4 µmol/L. Furthermore, HKI46F08 induced cell differentiation and apoptosis, overall being a promising agent in pediatric malignant diseases [161].

A cell proliferation assay, based on the measurement of optical density was used to study the effects of a small molecule inhibitor on gastric cancer [162] and gastrointestinal stromal tumor cell lines [163]. C646, a HAT inhibitor (see Section 1), inhibits the enzymatic activities of p300 and CREB binding protein (CBP). These HATs control the acetylation of histone H3, their inhibition exerts antineoplastic effects on various cancer cell lines [162,163]. In two studies [162,163], it was tested whether their inhibition provides beneficial effects against two types of malignant diseases of the gastrointestinal tract. In the first study [162], CBP and p300 enzymes were overexpressed in five gastric cancer cell lines. The control was a normal gastric cell line. C646 was also added to all of the cell lines. Optical densities were measured after incubation, and cell proliferation was calculated by dividing the optical density of the active set with the optical density of the control set. Higher doses (> 10 µmol/L) of C646 resulted in a stronger inhibition of cell proliferation on gastric cancer cell lines, than on normal gastric epithelial cells. In addition, it increased the number of apoptotic cells in gastric cancer cell lines and reduced migration and invasion potential [162]. C646 treatment reduced the acetylation of histone H3 in both gastric carcinoma cells and normal gastric epithelial cells. In the second study [163], the same protocol was repeated for gastrointestinal stromal tumor cell lines, with the introduction of an add-on treatment with imatinib. It was found, that alone 15 µmol/L C646 caused a marked decrease in cell proliferation of gastrointestinal stromal tumor cell lines, which result was further improved when combined with 500 nmol/L imatinib [163].

In Vivo Activity

In vivo activity tests cover investigations on living animals with various approaches. For example, in vivo enzyme activity tests can be performed by magnetic resonance (MRI)-based methods, microdialysis and fluorescence imaging [164] among others. In vivo MRI-based methods can apply microinjection of contrast agents cleaved by a specific enzyme to map the gene expression of transgenic animals, and improve our knowledge of mRNA expression, inheritance patterns and plasmid gene expression [165]. A selection of further in vivo studies on the epigenome is detailed in the forthcoming paragraphs.

In vivo chromatin immunoprecipitation (ChiP) assays are frequently used in recent works. ChiP is performed by cross-linking DNA and associated proteins, then fragmented DNA segments associated with proteins are extracted from the debris by protein-specific antibodies. The DNA segments are then purified, and their sequences are determined. With this approach, locations in the genome associated with specific histone PTMs can be screened. For example, ChiP assays identified that AIRE forms complexes with small fractions of H3K4me0, but not with H3K4me3. Furthermore, the specific promoter regions of DNA interacting with AIRE were found [130]. Promoter regions, where mostly H3K4me0 is expressed, like the insulin promoter region interact with AIRE. At the same time, regions that lack H3K4me0 but rich in H3K4me3, like the glyceraldehyde-3-phosphate dehydrogenase promoter region does not interact with AIRE [130].

ChiP-sequencing (ChiP-seq) combines ChiP assays with parallel DNA sequencing, similarly, to map DNA binding sites of proteins. After the ChiP assay, all DNA fragment sequences are determined in parallel, for a genome-wide analysis. In a ChiP-sequencing study [166], nuclei were extracted from midbrain dopamine producing neurons (mDA) of adult mice to create ChiP-seq libraries. The presence of repressive and permissive histone PTMs, H3K27me3, H3K9me3 and H3K4me3 around transcription start sites were screened to gain a picture on how the equilibrium state of the chromatin correlates with gene expression rates [166]. Occurrence of H3K4me3 was associated with high expression if compared with the total average gene expression level even when co-occurring with repressive modifications. The distribution of other histone modifications also correlated with gene expression as chromatin regions rich in H3K27me3 and H3K9me3 corresponded to lower than average gene expression. The simultaneous presence of H3K27me3 and H3K9me3 were associated with terminal repression of the gene expression. This chromatin equilibria regulation is maintained during transition from neuronal progenitor cells (NPCs) to mDA. The already H3K27me3 enriched genes not only maintain their repressed states in equilibrium, but even gain additional H3K9me3 marks upon transition from NPCs to mDA [166].

The xenograft tests are applied in tumor growth studies in vivo, where cancer cell lines are transferred into animals, then control and treated groups are formed to study the effect of a drug on tumor size. For example, a drug named corin was tested in a melanoma xenograft mice model [54] for tumor growth modifying effects [54] through targeting epigenetic pathways. A therapeutic target of special interest, that contains a HDAC enzyme (see in Categories 1, 2 and 5 of Section 1), namely CoREST complex (see in Category 5 of Section 1), consisting of REST co-repressor 1 protein (CoREST), HDAC1 or HDAC2 and LSD 1 enzymes. Corin was tested as a dual action LSD1 (see in Category 5 of Section 1)/HDAC inhibitor targeting the CoREST complex. Corin showed metabolic stability and proved to be well-tolerable in mice. Mice were divided into vehicle and corin treated groups. Following euthanasia of the animals, tumors were collected and measured. Corin showed a marked reducing effect on tumor growth compared to vehicle. Tumor cells extracted from these mice showed an elevated H3K9ac acetylation and H3K4me2 dimethylation in corin-treated mice, compared to vehicle-administered mice [54]. This observation correlates well with the HDAC and demethylase inhibitor functions of the drug, corin.

In another xenograft study [160], sixty-day releasing 17ß-estradiol pellets were subcutaneously inserted into the shoulders of eleven to fifteen weeks-old female mice [160]. MCF7 cells (See Section 2.2.2) were also subcutaneously inoculated into the animals. After the tumor reached a certain size, the mice were divided into groups, and the treatment was initiated. The control group was injected with solvent, treatment groups were injected with tamoxifen and anacardic acid. Tumor xenograft volumes were then measured. The tumor growth was reduced compared to control groups [160], this is in good agreement with the in vitro results of the same study [160], detailed in Section 2.2.2. The combined treatment with tamoxifen and anacardic acid inhibited ER-regulated gene transcription. Anacardic acid alone showed a reduction in H4K12ac occupancy near growth regulation by estrogen in breast cancer 1 (GREB1) transcription starting site, tamoxifen alone did not exhibit this effect [160].

In tumorigenesis studies, the growth of a tumor is also induced in animals, like in xenograft studies, but with the intention to study the pathomechanism of a certain type of tumor, or the pathologic pathway induced by an agent. The contribution of epigenetic changes to the carcinogenicity of potassium dichromate (further referred to as CrVI) was investigated in a study [167]. CrVI is a known genotoxic carcinogen. CrVI-transformed cells from human lung cancer tissues and CrVI-exposed human bronchial epithelial cell lines were injected into female nude mice [167]. Chronic CrVI exposure increased histone-lysine methyltransferase expression, and consequently repressive H3 methylation marks, playing a causal role in the carcinogenicity of CrVI. Gene knockdown or pharmacological inhibition of the histone-lysine methyltransferase diminished this effect [167].

Among invasive sampling methods microdialysis is often used for continuous measurements of unbound analyte concentrations of the extracellular fluid. For example, dopamine levels were measured in mice brain after alcohol administration, to investigate the effect of alcohol on histone acetylation patterns [168]. Microdialysis was performed by inserting a dialysis probe into the brain tissue of mice above the nucleus accumbens. The probe was perfused with artificial cerebrospinal fluid at a constant rate. Baseline samples were taken to measure the baseline neurotransmitter levels. Animals were injected either with ethanol or saline and samples were collected every 20 min through the microdialysis probe [168]. It was found [168], that ethanol administration provokes similar prolonged dopamine response in both adolescent and adult rats, but basal dopamine levels were higher in ethanol-treated adolescent rats, than in similarly treated adult rats [168]. Finally, ethanol administration changed the histone H3 and H4 acetylation of adolescent rats in the nucleus accumbens, striatum and frontal cortex. The study [168] concluded that epigenetic changes might contribute to the increased vulnerability of adolescent rats to alcohol addiction.

3. Theoretical Calculations of Molecular Structure and Binding Affinity

The previous paragraphs of Section 2.1.2 on experimental methods highlighted the most important challenges of determination of molecular structures in the epigenome. It was explained how the high number of variations meet large biomolecular system sizes, resulting in an extraordinary problem to be tackled. Such complexity of the epigenetic interactome clearly shows the limits of experimental approaches. In this situation, the use of theoretical approaches is inevitable to enhance the production of new structural information and also to predict the strength of corresponding molecular interactions. Although some of the theoretical methods require an advanced computational infrastructure, the cost of such facilities is still moderate if compared with that of experimental studies. Moreover, due to the general need and spread of information technologies in all fields of society, their development obviously shows an increasing trend. This has a positive feedback effect on the scientific applications often resulting in higher benefit-cost ratios in both the software and the hardware components of these technologies. Beyond a complementary use of theoretical approaches, the forthcoming paragraphs also highlight their advances over physical experiments in problematic cases where measurements are not available or reached their natural limits. The survey of the forthcoming paragraphs includes examples at various levels of theory.

Knowledge-based approaches are trained on sets of experimental molecular structures and their physico-chemical background is restricted to basic principles only. They often depend on comprehensive databases and internet services such as the Protein Databank [87], the Basic Local Alignment Search Tool (BLAST, [169]) or FASTA [170]. Knowledge-based methods provide fast results, often implemented in on-line servers, and do not require extensive computational infrastructure. However, the reliability of their results is limited by the training data set and the state-of-art of the databases and servers working in the background. Molecular mechanics (MM) methods are generally published as standalone tools based on more sophisticated physical chemistry, but still working by the laws of classical physics [171]. MM methods allow not only local search and fast optimization of the structures, but also extensive conformational sampling and global search in molecular dynamics (MD), Monte-Carlo or genetic algorithms. Such features are of particular importance during investigations of structural (Section 3.1) and energetical (Section 3.2) properties of molecular interactions, also accounting for interface flexibility, or predicting protein side chain conformations, and so forth [172–174].

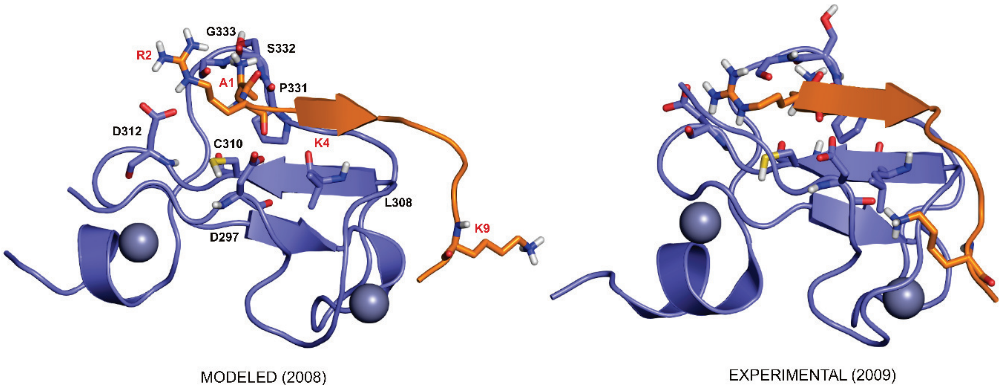
3.1. Molecular Structure

There are two main goals of theoretical methods on structural calculations. They produce either an atomic-resolution structure of a single macromolecule, such as a protein target for drug design, or a complex structure of two or more partners involved in protein–ligand, protein–protein, protein–DNA or other interactions. Beyond production of such static structures (snapshots), recent molecular dynamics investigations often produce a series of molecular geometries presenting the evolution of the systems. This feature is of particular importance for the exploration of induced effects and drug binding mechanism. Accordingly, the next paragraphs will feature selected results of static (Section 3.1.1) and dynamic (Section 3.1.2) methods, as well.

3.1.1. Static Methods

Among knowledge-based approaches, homology modeling is a primary structure prediction method of proteins and their complexes. It is a quick technique based on the assumption that similar sequences fold into similar structures. The technique requires access to on-line databases holding protein sequences [175], the above-mentioned sequence comparison algorithms (BLAST, FASTA) for selection of a template protein available in the Protein Databank. An acceptable homology model requires a large sequence identity between the modeled (target) and template structures [176]. The number of known protein sequences is higher than that of determined protein structures [177]. As protein structure is primary information for target-based drug design, homology modeling is often involved in such projects with epigenetic targets [52,178–192]. The homology-modeled protein targets can be used in virtual screening of chemical libraries to find lead molecules. For example, in a study [188], a 3D structure of KDM5A (see in Category 5 of Introduction and Section 2.2.1) jumonji domain was built by homology modeling. For building KDM5A jumonji domain the program MODELLER [193] was used, with four templates, which either contain a jumonji domain or a similar structure. Template proteins lysine specific demethylase 4C jumonji domain [194] and 2-oxoglutarate oxygenase [194] both had a ca. 40% amino acid sequence identity to the jumonji domain of KDM5A. Five models were generated from each template, and then the models were subjected to different MM minimization steps, to select the most appropriate model for the screening process. Then molecular docking-based virtual screening was performed on compound libraries, and structure activity relationship analysis was carried out, to identify novel potent inhibitors of the enzyme. KDM5A functions as a transcriptional repressor (see in Category 5 of Introduction), through the demethylation of H3K4me3 [188]. Dysregulation of KDM5A is involved in the pathomechanism of various human malignant diseases, such as breast cancer and acute myeloid leukemia [77,78,188]. In addition, it was shown, that KDM5A is involved in the drug resistance of anti-cancer drugs [188]. The hit compound identified in this study [188] showed an in vitro IC_{50} value of 0.22 µM on the KDM5A enzyme, a promising starting point for further investigations.

Besides homology modeling of single targets, combination of existing structures can also lead to the solution of target–ligand complexes. For example, the complex of AIRE PHD finger and histone peptide H3K4me0 was modeled [130] by superimposing the histone complex of NURF BPTF PHD finger (see in Section 1, [80]) and the apo structure of AIRE PHD finger [195] by the program Lsqman [196]. Linear disordered C and N terminal parts of AIRE PHD finger and the NURF BPTF PHD finger were then removed, and the remaining AIRE PHD finger was blocked by acetyl and amide groups. The system was refined by MM energy minimizations with GROMACS [197]. One year later, the solution NMR structure of the complex was also captured [95]. Comparison of the modeled and experimental structures (Figure 8) show that the binding mode of the H3 peptide ligand perfectly matches in the two structures including the three β-strands in antiparallel organization and important side-chain interactions between the histone peptide residues H3R2, H3K4, H3T6 and backbone residues of the AIRE PHD finger C310, L308 and G306, respectively. Furthermore, backbone carbonyl oxygen atoms of residues P331-G333 anchored the N-terminal end of the histone peptide tail through intermolecular hydrogen bonds. Hydrophobic interactions occurred between the methyl group of H3A1 and the pyrrolidine ring of P331, and the methylene groups of H3K4 and L308. Finally, two salt bridges were formed between the side chains of H3R2 and D312 and H3K4 and D297 [130].



MODELED (2008)

EXPERIMENTAL (2009)

Figure 8. Comparison of modeled [130] and experimental [95] structures of histone H3K4me0 (orange cartoon) in complex with AIRE PHD finger (navy blue cartoon). Zinc ions and key residues are shown as navy blue spheres and labelled sticks, respectively. There is a good agreement between the modeled and the experimental structures both capturing the antiparallel ß-strand as binding conformation of histone H3K4me0.

Calculation of structures of target–ligand complexes is an important goal of MM- or knowledge-based computational docking approaches, as well. Initially, fast docking programs were introduced to select small-molecule drug candidates during rapid screening of ligand libraries (see previous paragraph on KDMA5 target for an example). The technique is also useful for predictions on the target side during investigation of the effects of amino acid mutations involved in ligand binding [198–200]. Beyond small molecule drug candidates, there is a rising interest in peptides as ligands. However, flexibility of both the target and the peptide ligand [201–204] is challenging for fast docking methods [205] which often neglect conformation flexibility to reduce computational cost [100].

Numerous servers have become available based on fast docking approaches [202,204–214]. Some of these are designated for peptide ligands [207,209,210,212,214,215]. For example, the FlexPepDock [207] server was used to dock H3K4me3 peptide on transcription factor 19 (TCF19) PHD [216]. The homology model of TCF19 PHD was created on the basis of the PHD finger of the Jumonji/ARID domain containing protein 1A, this template has a 50% amino acid sequence identity with the TCF19 PHD. The molecular mechanism behind the function of TCF19 was explored, and it was found, that TCF19 PHD selectively interacts with histone H3K4me3 mark, and recruits the co-repressor complex NuRD (Section 1), to regulate gluconeogenic gene expression in HepG2 cells [216].

While the field of fast docking methods is rapidly developing, the above problems of peptide ligands have not been solved yet. In particular, the docking of histone peptides to their targets (readers) is still problematic and rarely addressed. To address this problem, a fragment blind docking [99] strategy was introduced and tested for docking of 7–13 amino acid long histone N-terminal tail peptides. Selected epigenetic proteins were targeted (references to the Sections indicate the places of further explanations on each protein) including AIRE (Section 2.1.2), ATRX-ADD (Section 1), DNMT3L (Section 1), KAT (Section 1) and Set domain containing protein (Section 2.1.2). The strategy applied Wrap'n'Shake [100], a blind docking method wrapping the target surfaces with a monolayer of copies of dipeptide ligand fragments. Then the full peptide ligands were reconstructed by linking the fragments. With this strategy, good agreements were achieved with experimental structures for the N-terminal part of histone peptides. For example, the N-terminal ARTK peptide of H3 showed a low root mean squared deviation (RMSD of 1.3 Å) from the experimental conformation [99]. Notably, the N-terminal segment has primary role in histone interactions and structure determination of H3 complexes is often restricted to this region (see also Section 2.1.2).

The results of virtual docking screens are often piped into in vitro assays for final selection of the top candidates. For example, the discovery of novel DNMT3A (Section 1) inhibitors was aided by structure-based virtual screening and in vitro DNMT3A inhibition assays [217]. DNMT3A is responsible for the methylation of cytosine at C5 position [217]. DNA hypermerthylation of specific genes contributes to cancer initiation and progression [217]. Specifically, DNMT3A mutations are associated with haematological malignancies [217]. In a study, over 77,000 commercially available molecules were subjected to virtual screening via molecular docking for an X-ray crystallographic DNMT3A structure [217]. The molecules were docked onto the S-adenosyl-l-homocysteine site of the enzyme, with Glide [218] DOCK [219] docking programs. The top ranked molecules were evaluated by AMBER [220] scoring, and the remaining 1000 molecules were merged into one file for cluster analysis [217]. 107 molecules were then evaluated by in vitro DNMT3A inhibition assays, and two compounds displayed significant in vitro inhibitory activity with IC_{50} values of ca. 40 µM [217].

Similarly, docking-based virtual screening was performed on SPECS database, to identify novel non-nucleoside DNMT1 inhibitor compounds [221]. DNMT1 is the most abundant among DNMTs, yet non-nucleoside inhibitors are lacking against DNMT1 [221]. Non-nucleoside inhibitors do not show as many side effects, as nucleoside analogs, unfortunately they also show lower potencies than nucleoside analogs [221]. An X-ray crystallographic structure was used, and similarly, its S-adenosyl-l-homocysteine binding site was searched, and after discarding ligands with unfavorable physicochemical properties, over 110,000 compounds were screened against DNMT1 [221]. After scoring by Glide [218], 51 compounds remained for further evaluation in biochemical assays. A compound, DC_05 showed remarkable selectivity for DNMT1 isoform, with an IC_{50} value of 10.3 µM [221]. Afterwards, similarity-based analog searching was performed, with DC_05 as a lead compound, and two even more potent agents were found, DC_501 and DC_517 with IC_{50} values of 2.5 and 1.7 µM, respectively [221].

3.1.2. Dynamic Methods

Besides production of a static snapshot of a single conformation (Section 3.1.1), uncovering interaction dynamics is another key to epigenetic regulation. MD can produce a time series of conformations of proteins or any molecular assemblies. It can be used to check the long-term stability of folding and complexes of large molecules even on a ms scale. Explicit solvent MD simulations allow very precise calculations accurately modeling real solutions.

MD simulations can be used to study the effect of PTMs on the stability of histone-reader complexes. For example, the effect of methylation of H3K4 on its binding strength to AIRE PHD finger was investigated using GROMACS [222] software package with GROMOS96 [223] force field [130]. Instead of the H3K4me0 peptide (Figure 8) H3K4me3 with trimethylated lysine side chains was applied and neutralizing counter ions and 7800 explicit single point charge waters were used in rectangular simulation boxes [130]. Short MD simulations were performed on four different complexes including AIRE PHD finger in complex with four possible PTM variants of the K4 amino acid, respectively. It was found that the H3K4me0 variant had stable complex with the first β-strand of AIRE PHD finger, creating an antiparallel β-sheet, stabilized by the interactions listed in Figure 8 of Section 3.1.1 [130]. Two salt bridges were formed between histone H3R2 and D312 and H3K4 and D297 residues [130], which is crucial for complex stability. Increasing the number of methyl groups (H3K4me, H3K4me2) destabilized the complex. In the case of H3K4me3, the complex could not be stabilized, and the two partners quickly dissociated in the MD runs. The bulky trimethylamino group of H3K4me3 could not participate in the above salt bridge hindered by several possible clashes with the target surface [130].

The effects of interaction networks of water molecules on complex stability can be also investigated by MD. An example of the ternary complex of DAXX protein and histones H3.3 and H4 was described (Section 2.1.2) in detail previously. In another example, MD simulations with GROMACS and MobyWat [127] were applied for the calculation of interfacial hydration network of ATRX-ADD protein in complex with histone H3 tail, trimethylated at H3K9me3. The binding of ATRX-ADD (Section 1) to histone H3 tail is promoted by H3K9me3 PTM mark and inhibited by H3K4me3. After mutation of the trimethyllysine binding pocket of ATRX-ADD it cannot bind to histone H3K9me3, and pericentromeric heterochromatin, leading to apoptosis in neuroprogenitor cells and mental retardation syndrome [81]. The relatively large histone H3 tail interacts via a shallow binding interface, where its arginine and lysine side chains are open to interact with water molecules from the bulk solvent. MobyWat showed that 12.5% of the water molecules of the system has low mobility, and is involved in a static sub-network [127]. The formation of such static networks are essential for complex stability anchoring the N-terminal tail of the histone to the target molecule, and can also shield the target-ligand H-bonds from solute attacks [127].

Two bromodomains, CBP (Section 2.2.2) and bromodomain adjacent to zinc finger binding domain 2B (BAZ2B) [224] were also analyzed in MD simulations, and conserved water molecules were studied at the bottom of the acetyl-lysing binding site of the bromodomains. The movement of the ZA loop of the binding site of the bromodomains has an influence on the presence of conserved water molecules in the binding site. These water molecules are connected by hydrogen bonds and were all either present or absent along the simulation [224]. Co-solvents, DMSO and (m)ethanol were added to the system, and similar results were achieved, with available crystallographic conserved bromodomains with the same co-solvents [224]. In their most populated binding modes, the co-solvents accepted a hydrogen bond from the same asparagine residue that is involved in the binding of acetyl-lysine [224]. Upon reaching more buried binding modes, the co-solvents displaced the same structured water molecules during the MD simulations. It was concluded in the study [224], that during ligand design, only the structured water molecules, that do not exchange with bulk solvent should be kept in crystal structures, during docking runs [224], and the identified water molecules, displaced by (m)ethanol co-solvents, might be targeted by hydrophilic moieties of the ligand [224].

MD simulations also uncovered the structural background of substrate selectivity of lysine specific demethylase 4A (KDM4A or JMJD2A) [225]. JMJD2A is a histone demethylase, specific for di- and trimethylated H3K9 and H3K36. The expression of JMJD2A is increased in prostate cancer [78]. MD simulations with mono-, di-, and trimethylated H3K9 peptides and JMJD2A were performed. The JMJD2A enzyme has a Fe^{2+} ion in its active site, to which a water molecule is coordinated, which does not form any hydrogen bonds with its surrounding atoms [225]. This water molecule stayed coordinated to Fe^{2+} throughout the whole simulation (20 ns). In all three cases of the PTMs, the water molecule was located always between the Fe(II) and the methylammonium moiety [225]. In the case of the mono- and dimethylated peptides, water molecules occupied the place of the missing methyl groups. These water molecules play an important role in ligand orientation within the binding pocket of JMJD2A, for example a water molecule, that stayed close to the methylammonium heads of the ligands through the simulation, formed hydrogen bonds with serine and glycine residues of the protein [225]. Apart from water molecules, from a structural point of view, the binding of H3K9me3 was found to be favorable, because of the symmetry of the ligand, which leads to an adequate orientation of the methyl groups. The preferable orientation of the methyllysine head in the case of H3K9me2 results from the restriction of angular motion by surrounding asparagine and glycine residues. If the dimethyllysine was rotated one of the methyl groups would overlap with the atoms of the surrounding asparagine and glycine residues of JMJD2A. The energy barrier observed between the three minima of the torsion states of H3K9me2 methyllysine head prevented the head from a circular motion [225]. Furthermore, the H3R8 formed intramolecular hydrogen bonds with H3K9me2 and H3K9me3, this interaction has a favorable energy contribution to the ΔG of the ligand [225].

3.2. Binding Affinity

The relevance and experimental methods of the measurement of binding affinity were introduced in Section 2.2.1. The large number of molecular interactions in the epigenome (Section 2.1.2) necessitated the development of theoretical approaches for fast generation of binding affinity data. There are statistical and end-point methods [226] available for calculation of binding thermodynamics, mostly ΔG. The development of such structure-based approaches is a hot field of research due to their central importance in rational drug design.

3.2.1. Statistical Methods

The first group of methods uses sampling of a statistical ensemble of conformations of interacting molecules. MD (Section 3.1.2) is often used for production of such samples of billions of states of macromolecular systems also providing information for calculation of ΔS, as well [227]. Among the statistical methods, alchemical energy calculation methods involve the transformation of one ligand into another, or a non-interacting particle [226]. Pathway methods are somewhat computationally expensive and follow the whole path of the binding process a useful option in drug design [226].

Using alchemical atom-type mutations, thermodynamic integration technique was applied to calculate the ΔG of CpG DNA site with methyl-CpG binding domain protein 1 (MBD1), which binds to methylated sequences in DNA [228]. Via this binding event, MBD1 can influence transcription activity [228]. MD simulations uncovered the binding mechanism of MBD1 to a hemi-methylated DNA, where cytosine is only methylated on one DNA strand [228]. It was found that a hydrophobic path of MBD1 protein moves away from the demethylated cytosine, and this conformational change weakens the DNA-protein interaction [228]. During the binding process, bulk water enters the binding site at the interface, inducing the rearrangement of the hydrogen bond network and the loss of a crucial hydrogen bond, that would occur between methyl cytosine and a tyrosine residue of MBD1 [228]. On the other strand, due to these conformational changes, the hydrogens of the methyl group of the cytosine form hydrogen bonds with an arginine residue of MBD1 protein. In this way MD simulations contribute greatly to our knowledge on how methyl marks of the DNA is recognized in the epigenetic machinery [228]. The proposed mechanism was validated by experiments. The binding of MBD1 protein to fully methylated CpG DNA site is more favorable if compared with the unmethylated CpG DNA site [228].

In another alchemical paper, free energy perturbation was used [229] to quantify the interaction of methyl-lysine histone and its reader protein, lethal 3 malignant brain tumor like protein 1 (L3MBTL1). The calculated ΔG was validated by ITC measurements. It was found that an asparagine residue of L3MBTL1 protein acts as an anchor, and its mutation disables any measurable binding of histones [229]. Instead of histone peptides, probes were used in the calculations and experimental measurements, which were assumed to act similarly to histones. Interestingly, it was found that the addition of a methyl group to Nme0 (non-methylated amino moiety), or the removal of a methyl group from Nme3 results in an affinity gain (−4.73 kcal/mol and −2.11 kcal/mol, respectively) to the reader protein, while the addition of a methyl group to Nme1 (−0.3 kcal/mol) does not affect ΔG [229]. The atomic level background of this unusual phenomenon was investigated by MD simulations. It was found that Nme0 lacks all favorable van der Waals contributions; furthermore, its positive charge is shared between three hydrogen atoms, resulting in a considerable loss of electrostatic contribution to ΔG. Nme3 binding is penalized by steric repulsions, and positive energy contribution of non-polar terms, as well. Overall, Nme1 or 2 was concluded as a preferred PTM state of histone for binding to the L3MBTL1 protein [229].

Adaptive lambda square dynamics was applied [230], to calculate the impact of K14 acetylation on histone H3 conformation. It was found that H3K14ac results in a weaker interaction between the DNA and the histone H3 tail, and the acetyl mark enhances α-helix formation of the histone H3 tail. The favorable electrostatic interaction between H3K14ac and H3K18 leads to increased α-helix formation [230]. This results in a more compact tail conformation [230]. This compaction results in the unwrapping of the linker DNA from the nucleosome, and the exposure of the linker DNA [230], which enables DNA binding proteins (e.g., transcription factors, see Category 4 of Section 1), to bind to their target sequences [230].

The attach-pull-release method was used to calculate the binding free energy of seven small ligands to a bromodomain [231] (see also Section 3.1.2 on epigenetics relevance). During these investigations, the ligands were pulled off the bromodomain binding site, allowing its conformational relaxation. In this study [231], a conformational change in the bromodomain is revealed by MD simulations. In experimental apo crystal structures the bromodomain is in a closed state, which opens up in MD simulations after 20–60 ns run time [231]. In a loop, the two main chain asparagine residues undergo a transition of torsion angles, and other residues change only minimally. If a restraint on the torsion angle of one of the asparagine residues is applied, the conformational change does not occur [231]. The calculated ΔGs of the seven ligands were compared to experimental data from the literature. Additionally, various water models and ligand parameter set combinations were compared, both using the open and the closed states as the final apo state of the protein. Using the open state as the final apo state of the protein SPC/E [232] water model with GAFF force field [220,233] provided the best results compared to experimental data from the literature (RMSE 1.42 kcal/mol) [231]. The open conformation of the enzyme was found to be more favorable energetically [231]. As the transition of the apo protein to open state is thermodynamically favorable and promotes dissociation, keeping the protein in a closed state improves ΔGs [231]. This improvement in calculated binding free energies reduced the bias of computational results and led to a better agreement with experimental values. Interestingly, when the closed state of the protein was used as the final apo state the previously weakly performing TIP3P [234] and TIP4Pew [235] water models showed the best results (RMSE ranging from 1.14 to 1.61 kcal/mol).

3.2.2. End-Point Methods

End-point energy calculations are based on the initial free ligand and target, and the final complex structures [226]. Due to the small number of conformations end-point methods are computationally efficient and fast [226]. Molecular Mechanics Generalized Born Surface Area (MM GBSA) [236–245] and Molecular Mechanics Poisson Boltzmann Surface Area (MM PBSA) [239,241,242,246–250] methods are commonly applied, single-trajectory approaches. The conformations of the interacting partners in their complex are assumed to represent the unbound partners [226] leading to several approximations [140,251].

The performances of MM PBSA and absolute alchemical binding free energy calculation methods were compared [252] using 22 different targets of epigenetic importance. Most of the calculations were performed on the members of the bromodomain and extraterminal (BET) family, including BRD2, 3, 4 and BRDT proteins. [253,254]. BET proteins regulate the expression of key oncogenes and anti-apoptotic proteins, making them a promising target in epigenetic drug design against malignant diseases, inflammation and viral infections [252–254]. Acetylation of lysine residues on the N-terminal tail of histone is associated with an open chromatin conformation and therefore transcriptional activation [254] (see also Category 5 in Section 1). Bromodomains are the readers of the acetylated lysine residues, the therapeutic approaches, and their small molecule inhibitors are reviewed in the literature [254]. In the study [252], abundant small molecular inhibitors of bromodomains were used in ΔG calculations. The calculated energies were compared to experimental ΔGs, and a thorough statistical analysis was performed. Absolute binding free energy calculations outperformed the MM PBSA approach for the investigated bromodomain complexes [252].

Scoring function of docking programs are often based on end-point ΔG calculations and applied in epigenetic drug design [218,255,256]. Scoring values usually show low correlation with experimental ΔGs [255], and therefore, they are mostly applied for distinction between ligand candidates, relative comparison of the members of a docked ligand library. This problem can be also addressed by consensus scoring [255]. In this case, if a hit is identified, the majority of scoring function methods have to rank it as the top ranks to get accepted. As scoring functions show inconsistent performance on different receptors, the careful selection of a scoring function is important for virtual screening [255]. Scoring functions GoldScore, ChemPLP, ASP and CDOCKER_ENERGY were applied [255] to support the hit discovery of sirtuin 2 (SIRT2) inhibitor. SIRT2 is a nicotinamide adenine dinucleotide-dependent deacetylase, that plays a role in the pathomechanism of various diseases, including cancer and neurodegenerative diseases [255]. It has a wide variety of substrates, including histone H3K18ac and H3K56ac, and H4K16ac [257]. Interestingly, within its display of substrates, there is a histone writer, histone methyltransferase PR-Set7 [257]. PR-Se7 specifically mono-methylates H4K20me0, but if the enzyme is deacetylated by SIRT2, its localization on the chromatin is altered, decreasing its ability to methylate H4K20me [257]. In the study [255], a SIRT2 inhibitor with a new scaffold was identified, and subjected to structure activity relationship analysis, and finally four compounds were developed with IC_{50} values less than 10μM against SIRT2.

Determination of components of ΔG is important in thermodynamic optimization of drug candidates (Section 2.2.1). In particular, the optimization of ΔH can increase drug efficiency [137,138]. To help this trend of drug design, end-point quantum mechanical (QM) approaches have been developed for structure-based calculation of ΔH [258,259]. While QM calculations offer the highest possible theoretical accuracy, they are expensive and demanding in computational time. Thus, the program Fragmenter was developed for the reduction of system size by extraction of hydrated interfaces from target-ligand complexes [136]. The complexes were subjected to semi-empirical QM calculations with PM7 parameterization, and the calculated ΔHs were correlated with available experimental values. Interestingly, the study [136] found a simple scaling factor for conversion between calculated to experimental binding enthalpies. Among other protein-peptide systems, the method was used to calculate the ΔH of the AIRE PHD finger–histone H3 peptide system featured in Figure 8 [136].

4. Conclusions

The present review featured current trends and selected methodologies of exploration of molecular structure, binding affinity and pharmacological activity in the epigenome. The design of efficient epigenetic drugs requires valid SARs and simultaneous development of all three fields. In recent decades, cryo-electron microscopy has opened a new avenue in the determination of molecular structure of nucleosome-sized assemblies. At the same time, high-throughput determination of atomic resolution structure of large biomolecules has not been implemented routinely. Thus, the application of alternative crystallographic, and theoretical approaches remains inevitable. Measurement of binding affinity is an important intermediate step during optimization of pharmacological activity. There is a wide range of techniques available at different levels of sophistication. Depending on the project, assays can be applied for fast screening of drug candidates or assessment of the effects of epigenetic modifications. On the other hand, sophisticated techniques like isothermal titration calorimetry help the optimization of the lead compounds providing detailed information on binding thermodynamics. Similarly, available theoretical binding affinity calculators provide fast and/or precise solutions using molecular structures as starting points. Molecular dynamics also helps to uncover binding mechanisms and supplies statistical amount of molecular conformations for energy calculations. In vivo activity experiments are essential for the final decision on further development of drug candidates and also provide a feedback for re-investigation of the epigenetic background of a disease.

The reviewed methods have proved indispensable during the discovery of epigenetic drugs accepted for clinical use. Such U.S. Food and Drug Administration (FDA)-approved epigenetic modulating drugs include vorinostat, romidepsin, panobinostat, belinostat (HDAC inhibitors), azacitidine, decitabine (DNMT inhibitors), enasidenib and ivosidenibe (isocitrate dehidrogenase inhibitors) [260] and tazemetostat, a histone methyltransferase inhibitor [261,262]. Tazemetostat is a selective inhibitor of enhancer of zeste homolog 2 (EZH2), a histone methyltransferase, that trimethylates H3K27me3 [260–264]. Given that EZH2 is a transcriptional suppressor, histone methyltransferase, and transcriptional co-activator, it is involved in a wide variety of cellular processes, some of which are directly linked to cancer pathomechanisms [264], EZH2 is in the highlight of biotechnological and pharmaceutical companies. Tazemetostat is approved by FDA for the treatment of epithelioid sarcoma, malignant rhabdoid tumors, and integrase interactor 1 (INI1) negative tumors. Vorinostat [265–269], a HDAC inhibitor is used for the prevention of acute graft-versus-host disease, and the treatment of cutaneous T-cell lymphoma. The above-mentioned nine agents were approved between 2004 and 2018, highlighting the emerging role of epigenetics in current drug discovery and design. Further developments and spread of the above surveyed methods are essential but probably not sufficient criteria of future acceleration of the development of valid SARs and drug discovery in the epigenome. The invention of new computational technologies is necessary to handle the epigenetic SAR data universe and the improvement of their complementary applications with strong links to experiments is also inevitable.

Author Contributions: Conceptualization, B.Z.Z. and C.H.; resources, C.H.; writing—Original draft preparation, B.Z.Z. and C.H; writing—Review and editing, B.Z.Z. and C.H; visualization, B.Z.Z. and C.H.; supervision, C.H.; project administration, C.H.; funding acquisition, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Hungarian National Research, Development and Innovation Office (K123836), 2017-1.2.1-NKP-2017-00002 (NAP-2; Chronic Pain Research Group), EFOP-3.6.1.-16-2016-0004 and GINOP 2.3.2-15-2016-00050 "PEPSYS".

Acknowledgments: This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. The work was supported by the ÚNKP-19-4 New National Excellence Program of the Ministry for Innovation and Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

DNMT	DNA MethylTransferase
KAT	Lysine AcetylTransferase
HDAC	Histone DeACetylase
ITC	Isothermal Titration Calorimetry
SAR	Structure Activity Relationship
PHD	Plant HomeoDomain
MM	Molecular Mechanics
MD	Molecular Dynamics

References

1. Waddington, C.H. *The Strategy of the Genes*, 1st ed.; Routledge: New York, NY, USA, 1957; pp. 1–32.
2. Goldberg, A.D.; Allis, C.D.; Bernstein, E. Epigenetics: A landscape takes shape. *Cell* **2007**, *128*, 635–638. [CrossRef] [PubMed]
3. Riggs, A.D.; Martienssen, R.A.; Russo, V.E.A. Introduction. In *Epigenetic Mechanisms of Gene Regulation*, 1st ed.; Russo, V.E.A., Ed.; Cold Spring Harbor Laboratory Press: Huntington, NY, USA, 1996; pp. 1–14.
4. Holliday, R. Epigenetics: A historical overview. *Epigenetics* **2006**, *1*, 76–80. [CrossRef] [PubMed]
5. Heard, E.; Tishkoff, S.; Todd, J.A.; Vidal, M.; Wagner, G.P.; Wang, J.; Weigel, D.; Young, R. Ten years of genetics and genomics: What have we achieved and where are we heading? *Nat. Rev. Genet.* **2010**, *11*, 723–733. [CrossRef] [PubMed]
6. Monk, D.; Mackay, D.J.G.; Eggermann, T.; Maher, E.R.; Riccio, A. Genomic imprinting disorders: Lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.* **2019**, *20*, 235–248. [CrossRef] [PubMed]
7. Alegría-Torres, J.A.; Baccarelli, A.; Bollati, V. Epigenetics and lifestyle. *Epigenomics* **2011**, *3*, 267–277. [CrossRef] [PubMed]
8. Tsankova, N.; Renthal, W.; Kumar, A.; Nestler, E.J. Epigenetic regulation in psychiatric disorders. *Nat. Rev. Neurosci.* **2007**, *8*, 355–367. [CrossRef] [PubMed]
9. Berdasco, M.; Esteller, M. Clinical epigenetics: Seizing opportunities for translation. *Nat. Rev. Genet.* **2019**, *20*, 109–127. [CrossRef] [PubMed]
10. Feinberg, A.P.; Koldobskiy, M.A.; Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **2016**, *17*, 284–299. [CrossRef] [PubMed]
11. Lappalainen, T.; Greally, J.M. Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* **2017**, *18*, 441–451. [CrossRef] [PubMed]
12. Stricker, S.H.; Köferle, A.; Beck, S. From profiles to function in epigenomics. *Nat. Rev. Genet.* **2016**, *18*, 51–66. [CrossRef] [PubMed]
13. Feinberg, A.P.; Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **2004**, *4*, 143–153. [CrossRef] [PubMed]
14. Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding structural details to protein networks. *Nat. Methods* **2013**, *10*, 47–53. [CrossRef] [PubMed]
15. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. [CrossRef] [PubMed]
16. Bleicken, S.; Hantusch, A.; Das, K.K.; Frickey, T.; Garcia-Saez, A.J. Quantitative interactome of a membrane Bcl-2 network identifies a hierarchy of complexes for apoptosis regulation. *Nat. Commun.* **2017**, *8*, 73. [CrossRef] [PubMed]
17. Izzo, L.T.; Wellen, K.E. Histone lactylation links metabolism and gene regulation. *Nature* **2019**, *574*, 492–493. [CrossRef] [PubMed]
18. Gamblin, S.J.; Wilson, J.R. A key to unlock chromatin. *Nature* **2019**, *573*, 354–355. [CrossRef] [PubMed]
19. Mews, P.; Egervari, G.; Nativio, R.; Sidoli, S.; Donahue, G.; Lombroso, S.I.; Alexander, D.C.; Riesche, S.L.; Heller, E.A.; Nestler, E.J.; et al. Alcohol metabolism contributes to brain histone acetylation. *Nature* **2019**, *574*, 717–721. [CrossRef] [PubMed]
20. Bird, A. Perceptions of epigenetics. *Nature* **2007**, *447*, 396–398. [CrossRef] [PubMed]
21. Bjornsson, H.T. The mendelian disorders of the epigenetic machinery. *Genome Res.* **2015**, *25*, 1473–1481. [CrossRef] [PubMed]

22. Tough, D.F.; Tak, P.P.; Tarakhovsky, A.; Prinjha, R.K. Epigenetic drug discovery: Breaking through the immune barrier. *Nat. Rev. Drug Discov.* **2016**, *15*, 835–853. [CrossRef] [PubMed]
23. Dawson, M.A.; Kouzarides, T. Cancer epigenetics: From mechanism to therapy. *Cell* **2012**, *150*, 12–27. [CrossRef] [PubMed]
24. Kelly, T.K.; De Carvalho, D.D.; Jones, P.A. Epigenetic modifications as therapeutic targets. *Nat. Biotechnol.* **2010**, *28*, 1069–1078. [CrossRef] [PubMed]
25. Polli, A.; Godderis, L.; Ghosh, M.; Ickmans, K.; Nijs, J. Epigenetic and miRNA expression changes in people with pain: A systematic review. *J. Pain* **2019**. [CrossRef] [PubMed]
26. Richmond, T.J.; Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **2003**, *423*, 145–150. [CrossRef] [PubMed]
27. Lai, W.K.M.; Pugh, B.F. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 548–562. [CrossRef] [PubMed]
28. Kornberg, R.D.; Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **1999**, *98*, 285–294. [CrossRef]
29. Luger, K.; Mäder, A.W.; Richmond, R.K.; Sargent, D.F.; Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **1997**, *389*, 251–260. [CrossRef] [PubMed]
30. Bednar, J.; Garcia-Saez, I.; Boopathi, R.; Cutter, A.R.; Papai, G.; Reymer, A.; Syed, S.H.; Lone, I.N.; Tonchev, O.; Crucifix, C.; et al. Structure and dynamics of a 197 bp nucleosome in complex with linker histone H1. *Mol. Cell* **2017**, *66*, 384–397. [CrossRef] [PubMed]
31. Rodríguez-Paredes, M.; Esteller, M. Cancer epigenetics reaches mainstream oncology. *Nat. Med.* **2011**, *17*, 330–339. [CrossRef] [PubMed]
32. He, H.; Hu, Z.; Xiao, H.; Zhou, F.; Yang, B. The tale of histone modifications and its role in multiple sclerosis. *Hum. Genom.* **2018**, *12*, 31. [CrossRef] [PubMed]
33. Li, G.B.; Yang, L.L.; Yuan, Y.; Zou, J.; Cao, Y.; Yang, S.Y.; Xiang, R.; Xiang, M. Virtual screening in small molecule discovery for epigenetic targets. *Methods* **2015**, *71*, 158–166. [CrossRef] [PubMed]
34. Arrowsmith, C.H.; Bountra, C.; Fish, P.V.; Lee, K.; Schapira, M. Epigenetic protein families: A new frontier for drug discovery. *Nat. Rev. Drug Discov.* **2012**, *11*, 384–400. [CrossRef] [PubMed]
35. Zhang, H.; Emerson, D.J.; Gilgenast, T.G.; Titus, K.R.; Lan, Y.; Huang, P.; Zhang, D.; Wang, H.; Keller, C.A.; Giardine, B.; et al. Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **2019**, *576*, 158–162. [CrossRef] [PubMed]
36. Strahl, B.D.; Allis, C.D. The language of covalent histone modifications. *Nature* **2000**, *403*, 41–45. [CrossRef] [PubMed]
37. Bhaumik, S.R.; Smith, E.; Shilatifard, A. Covalent modifications of histones during development and disease pathogenesis. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1008–1016. [CrossRef] [PubMed]
38. Coetzee, N.; Sidoli, S.; Van Biljon, R.; Painter, H.; Llinás, M.; Garcia, B.A.; Birkholtz, L.M. Quantitative chromatin proteomics reveals a dynamic histone post-translational modification landscape that defines asexual and sexual *Plasmodium falciparum* parasites. *Sci. Rep.* **2017**, *7*, 607. [CrossRef] [PubMed]
39. Hee-Dae, K.; Call, T.S.M.; Ferguson, D. Drug addiction and histone code alterations. *Neuroepigenomics Aging Dis.* **2017**, *7*, 127–143.
40. Yun, M.; Wu, J.; Workman, J.L.; Li, B. Readers of histone modifications. *Cell Res.* **2011**, *21*, 564–578. [CrossRef] [PubMed]
41. Fenley, A.T.; Anandakrishnan, R.; Kidane, Y.H.; Onufriev, A.V. Modulation of nucleosomal DNA accessibility via charge-altering post-translational modifications in histone core. *Epigenetics Chromatin* **2018**, *11*, 11. [CrossRef] [PubMed]
42. Tropberger, P.; Pott, S.; Keller, C.; Kamieniarz-Gdula, K.; Caron, M.; Richter, F.; Li, G.; Mittler, G.; Liu, E.T.; Bu, M.; et al. Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell* **2013**, *152*, 859–872. [CrossRef] [PubMed]
43. Rajagopalan, M.; Balasubramanian, S.; Ioshikhes, I.; Ramaswamy, A. Structural dynamics of nucleosome mediated by acetylations at H3K56 and H3K115,122. *Eur. Biophys. J.* **2017**, *46*, 471–484. [CrossRef] [PubMed]
44. Yuan, J.; Pu, M.; Zhang, Z.; Lou, Z. Histone H3-K56 acetylation is important for genomic stability in mammals. *Cell Cycle* **2009**, *8*, 1747–1753. [CrossRef] [PubMed]

45. Yu, Y.; Song, C.; Zhang, Q.; Dimaggio, P.A.; Benjamin, A.; York, A.; Carey, M.F.; Grunstein, M. Histone H3 lysine 56 methylation regulates DNA replication through its interaction with PCNA. *Mol. Cell* **2012**, *46*, 7–17. [CrossRef] [PubMed]
46. Robin, P.; Fritsch, L.; Philipot, O.; Svinarchuk, F.; Centre, A.; De, N.; Scientifique, R.; Fre, C.; Lwoff, I.A.; Moquet, G.; et al. Post-translational modifications of histones H3 and H4 associated with the histone methyltransferases Suv39h1 and G9a. *Genome Biol.* **2007**, *8*, R270. [CrossRef] [PubMed]
47. DeLano, W.L. *The PyMOL Molecular Graphics System*; Version 2.0; Schrödinger LLC: New York, NY, USA, 2002.
48. Davey, C.A.; Sargent, D.F.; Luger, K.; Maeder, A.W.; Richmond, T.J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **2002**, *319*, 1097–1113. [CrossRef]
49. North, J.A.; Šimon, M.; Ferdinand, M.B.; Shoffner, M.A.; Picking, J.W.; Howard, C.J.; Mooney, A.M.; Van Noort, J.; Poirier, M.G.; Ottesen, J.J. Histone H3 phosphorylation near the nucleosome dyad alters chromatin structure. *Nucleic Acids Res.* **2014**, *42*, 4922–4933. [CrossRef] [PubMed]
50. Ruthenburg, A.J.; Wang, W.; Graybosch, D.M.; Li, H.; Allis, C.D.; Patel, D.J.; Verdine, G.L. Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat. Struct. Mol. Biol.* **2006**, *13*, 704–712. [CrossRef] [PubMed]
51. Wysocka, J.; Swigut, T.; Milne, T.A.; Dou, Y.; Zhang, X.; Burlingame, A.L.; Roeder, R.G.; Brivanlou, A.H.; Allis, C.D. WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **2005**, *121*, 859–872. [CrossRef] [PubMed]
52. Qin, H.T.; Li, H.Q.; Liu, F. Selective histone deacetylase small molecule inhibitors: Recent progress and perspectives. *Expert Opin. Pat.* **2017**, *27*, 621–636. [CrossRef] [PubMed]
53. Shakespear, M.R.; Halili, M.A.; Irvine, K.M.; Fairlie, D.P.; Sweet, M.J. Histone deacetylases as regulators of inflammation and immunity. *Trends Immunol.* **2011**, *32*, 335–343. [CrossRef] [PubMed]
54. Kalin, J.H.; Wu, M.; Gomez, A.V.; Song, Y.; Das, J.; Hayward, D.; Adejola, N.; Wu, M.; Panova, I.; Chung, H.J.; et al. Targeting the CoREST complex with dual histone deacetylase and demethylase inhibitors. *Nat. Commun.* **2018**, *9*, 53. [CrossRef] [PubMed]
55. Kamps, J.J.A.G.; Huang, J.; Poater, J.; Xu, C.; Pieters, B.J.G.E.; Dong, A.; Min, J.; Sherman, W.; Beuming, T.; Bickelhaupt, M.F.; et al. Chemical basis for the recognition of trimethyllysine by epigenetic reader proteins. *Nat. Commun.* **2015**, *6*, 8911. [CrossRef] [PubMed]
56. Fraga, M.F.; Ballestar, E.; Villar-Garea, A.; Boix-Chornet, M.; Espada, J.; Schotta, G.; Bonaldi, T.; Haydon, C.; Ropero, S.; Petrie, K.; et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat. Genet.* **2005**, *37*, 391–400. [CrossRef] [PubMed]
57. Bibb, J.A.; Chen, J.; Taylor, J.R.; Svenningsson, P.; Nishi, A.; Snyder, G.L.; Yan, Z.; Sagawa, Z.K.; Ouimet, C.C.; Nairn, A.C.; et al. Effects of chronic exposure to cocaine are regulated by the neuronal protein Cdk5. *Nature* **2001**, *410*, 376–380. [CrossRef] [PubMed]
58. Heller, E.A.; Hamilton, P.J.; Burek, D.D.; Lombroso, S.I.; Peña, C.J.; Neve, R.L.; Nestler, E.J. Targeted epigenetic remodeling of the cdk5 gene in nucleus accumbens regulates cocaine-and stress-evoked behavior. *J. Neurosci.* **2016**, *36*, 4690–4697. [CrossRef] [PubMed]
59. Nelson, P.; Kiriakidou, M.; Sharma, A.; Maniataki, E.; Mourelatos, Z. The microRNA world: Small is mighty. *Trends Biochem. Sci.* **2003**, *28*, 534–540. [CrossRef] [PubMed]
60. Fabbri, M.; Garzon, R.; Cimmino, A.; Liu, Z.; Zanesi, N.; Callegari, E.; Liu, S.; Alder, H.; Costinean, S.; Fernandez-Cymering, C.; et al. MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15805–15810. [CrossRef] [PubMed]
61. Zhang, X.; Wang, W.; Zhu, W.; Dong, J.; Cheng, Y.; Yin, Z.; Shen, F. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *Int. J. Mol. Sci.* **2019**, *20*, 5573. [CrossRef] [PubMed]
62. Tsai, W.W.; Wang, Z.; Yiu, T.T.; Akdemir, K.C.; Xia, W.; Winter, S.; Tsai, C.Y.; Shi, X.; Schwarzer, D.; Plunkett, W.; et al. TRIM24 links a non-canonical histone signature to breast cancer. *Nature* **2010**, *468*, 927–932. [CrossRef] [PubMed]
63. Glass, C.K.; Rosenfeld, M.G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **2000**, *14*, 121–141. [PubMed]
64. Björnström, L.; Sjöberg, M. Mechanisms of estrogen receptor signaling: Convergence of genomic and nongenomic actions on target genes. *Mol. Endocrinol.* **2005**, *19*, 833–842. [CrossRef] [PubMed]

65. Kovács, T.; Szabó-Melegh, E.; Ábrahám, M.I. Estradiol-induced epigenetically mediated mechanisms and regulation of gene expression. *Int. J. Mol. Sci.* **2020**, *21*, 3177. [CrossRef] [PubMed]
66. Kim, S.; Kaang, B.K. Epigenetic regulation and chromatin remodeling in learning and memory. *Exp. Mol. Med.* **2017**, *49*, 281–288. [CrossRef] [PubMed]
67. Wang, G.G.; Allis, C.D.; Chi, P. Chromatin remodeling and cancer, part I: Covalent histone modifications. *Trends Mol. Med.* **2007**, *13*, 363–372. [CrossRef] [PubMed]
68. Budden, D.M.; Hurley, D.G.; Cursons, J.; Markham, J.F.; Davis, M.J.; Crampin, E.J. Predicting expression: The complementary power of histone modification and transcription factor binding data. *Epigenetics Chromatin* **2014**, *7*, 36. [CrossRef] [PubMed]
69. Everett, L.; Hansen, M.; Hannenhalli, S. Regulating the regulators: Modulators of transcription factor activity. *Methods Mol. Biol.* **2010**, *674*, 297–312. [PubMed]
70. Bannister, A.J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **2011**, *21*, 381–395. [CrossRef] [PubMed]
71. Dormann, H.L.; Boo, S.T.; Allis, C.D.; Funabiki, H.; Fischle, W. Dynamic regulation of effector protein binding to histone modifications: The biology of HP1 switching. *Cell Cycle* **2006**, *5*, 2842–2851. [CrossRef] [PubMed]
72. Liu, B.; KH Yip, R.; Zhou, Z. Chromatin remodeling, DNA damage repair and aging. *Curr. Genom.* **2012**, *13*, 533–547. [CrossRef] [PubMed]
73. Rajhans, R.; Nair, S.; Holden, A.H.; Kumar, R.; Tekmal, R.R.; Vadlamudi, R.K. Oncogenic potential of the nuclear receptor coregulator proline-, glutamic acid–, leucine-rich protein 1/modulator of the nongenomic actions of the estrogen receptor. *Cancer Res.* **2007**, *67*, 5505–5512. [CrossRef] [PubMed]
74. Choi, B.Y.; Ko, J.K.; Shin, J. The transcriptional corepressor, PELP1, recruits HDAC2 and masks histones using two separate domains. *J. Biol. Chem.* **2004**, *279*, 50930–50941. [CrossRef] [PubMed]
75. Adams, G.E.; Chandru, A.; Cowley, S.M. Co-repressor, co-activator and general transcription factor: The many faces of the Sin3 histone deacetylase (HDAC) complex. *Biochem. J.* **2018**, *475*, 3921–3932. [CrossRef] [PubMed]
76. Im, A.P.; Sehgal, A.R.; Carroll, M.P.; Smith, B.D.; Tefferi, A.; Johnson, D.E.; Boyiadzis, M. DNMT3A and IDH mutations in acute myeloid leukemia and other myeloid malignancies: Associations with prognosis and potential treatment strategies. *Leukemia* **2014**, *28*, 1774–1783. [CrossRef] [PubMed]
77. Plch, J.; Hrabeta, J.; Eckschlager, T. KDM5 demethylases and their role in cancer cell chemoresistance. *Int. J. Cancer* **2019**, *144*, 221–231. [CrossRef] [PubMed]
78. Cheng, Y.; He, C.; Wang, M.; Ma, X.; Mo, F.; Yang, S.; Han, J.; Wei, X. Targeting epigenetic regulators for cancer therapy: Mechanisms and advances in clinical trials. *Signal Transduct. Target.* **2019**, *4*, 62–101. [CrossRef] [PubMed]
79. Koedoot, E.; Fokkelman, M.; Rogkoti, V.M.; Smid, M.; van de Sandt, I.; de Bont, H.; Pont, C.; Klip, J.E.; Wink, S.; Timmermans, M.A.; et al. Uncovering the signaling landscape controlling breast cancer cell migration identifies novel metastasis driver genes. *Nat. Commun.* **2019**, *10*, 2983. [CrossRef] [PubMed]
80. Li, H.; Ilin, S.; Wang, W.; Duncan, E.M.; Wysocka, J.; Allis, C.D.; Patel, D.J. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **2006**, *442*, 91–95. [CrossRef] [PubMed]
81. Iwase, S.; Xiang, B.; Ghosh, S.; Ren, T.; Lewis, P.W.; Cochrane, J.C.; Allis, C.D.; Picketts, D.J.; Patel, D.J.; Li, H.; et al. ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. *Nat. Struct. Mol. Biol.* **2011**, *18*, 769–776. [CrossRef] [PubMed]
82. Ooi, S.K.T.; Qiu, C.; Bernstein, E.; Li, K.; Jia, D.; Yang, Z.; Erdjument-Bromage, H.; Tempst, P.; Lin, S.P.; Allis, C.D.; et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **2007**, *448*, 714–717. [CrossRef] [PubMed]
83. Pinzi, L.; Rastelli, G. Molecular docking: Shifting paradigms in drug discovery. *Int. J. Mol. Sci.* **2019**, *20*, 4331. [CrossRef] [PubMed]
84. Bragg, W.H.; Bragg, W.L. The structure of the diamond. *Nature* **1913**, *91*, 557. [CrossRef]
85. Brink, C.; Hodgkin, D.; Lindsey, Y.; Pickworth, J.; Robertson, J.H.; White, J.G. Structure of vitamin B12: X-ray crystallographic evidence on the structure of vitamin B12. *Nature* **1954**, *174*, 1169–1171. [CrossRef] [PubMed]
86. Aloy, P.; Russell, R.B. Structural systems biology: Modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 188–197. [CrossRef] [PubMed]

87. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The protein data bank. *Acta Cryst. Sect. D Biol. Cryst.* **2002**, *58*, 899–907. [CrossRef] [PubMed]
88. Bah, A.; Vernon, R.M.; Siddiqui, Z.; Krzeminski, M.; Muhandiram, R.; Zhao, C.; Sonenberg, N.; Kay, L.E.; Forman-Kay, J.D. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* **2015**, *519*, 106–109. [CrossRef] [PubMed]
89. Jemth, P.; Karlsson, E.; Vögeli, B.; Guzovsky, B.; Andersson, E.; Hultqvist, G.; Dogan, J.; Güntert, P.; Riek, R.; Chi, C.N. Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins. *Sci. Adv.* **2018**, *4*, 4130–4144. [CrossRef] [PubMed]
90. Adrian, M.; Dubochet, J.; Lepault, J.; McDowall, A.W. Cryo-electron microscopy of viruses. *Nature* **1984**, *308*, 32–36. [CrossRef] [PubMed]
91. Richmond, T.J.; Finch, J.T.; Rushton, B.; Rhodes, D.; Klug, A. Structure of the nucleosome core particle at 7 resolution. *Nature* **1984**, *311*, 532–537. [CrossRef] [PubMed]
92. Wilson, M.D.; Benlekbir, S.; Fradet-Turcotte, A.; Sherker, A.; Julien, J.P.; McEwan, A.; Noordermeer, S.M.; Sicheri, F.; Rubinstein, J.L.; Durocher, D. The structural basis of modified nucleosome recognition by 53BP1. *Nature* **2016**, *536*, 100–103. [CrossRef] [PubMed]
93. Park, S.H.; Ayoub, A.; Lee, Y.T.; Xu, J.; Kim, H.; Zheng, W.; Zhang, B.; Sha, L.; An, S.; Zhang, Y.; et al. Cryo-EM structure of the human MLL1 core complex bound to the nucleosome. *Nat. Commun.* **2019**, *10*, 5540–5553. [CrossRef] [PubMed]
94. Wagner, F.R.; Dienemann, C.; Wang, H.; Stützer, A.; Tegunov, D.; Urlaub, H.; Cramer, P. Structure of SWI/SNF chromatin remodeller RSC bound to a nucleosome. *Nature* **2020**, *579*, 448–469. [CrossRef] [PubMed]
95. Chignola, F.; Gaetani, M.; Rebane, A.; Org, T.; Mollica, L.; Zucchelli, C.; Spitaleri, A.; Mannella, V.; Peterson, P.; Musco, G. The solution structure of the first PHD finger of autoimmune regulator in complex with non-modified histone H3 tail reveals the antagonistic role of H3R2 methylation. *Nucleic Acids Res.* **2009**, *37*, 2951–2961. [CrossRef] [PubMed]
96. Girish, T.S.; McGinty, R.K.; Tan, S. Multivalent interactions by the set8 histone methyltransferase with its nucleosome substrate. *J. Mol. Biol.* **2016**, *428*, 1531–1543. [CrossRef] [PubMed]
97. Henderson, R.; Baldwin, J.M.; Ceska, T.A.; Zemlin, F.; Beckmann, E.; Downing, K.H.; Baldwin, J.M.; Ceska, T.A.; Zemlin, F.; Beckmann, E.; et al. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **1990**, *213*, 899–929. [CrossRef]
98. Renaud, J.P.; Chung, C.W.; Danielson, U.H.; Egner, U.; Hennig, M.; Hubbard, R.E.; Nar, H. Biophysics in drug discovery: Impact, challenges and opportunities. *Nat. Rev. Drug Discov.* **2016**, *15*, 679–698. [CrossRef] [PubMed]
99. Bálint, M.; Horváth, I.; Mészáros, N.; Hetényi, C. Towards unraveling the histone code by fragment blind docking. *Int. J. Mol. Sci.* **2019**, *20*, 422. [CrossRef] [PubMed]
100. Bálint, M.; Jeszenői, N.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Systematic exploration of multiple drug binding sites. *J. Cheminform.* **2017**, *9*, 65–77. [CrossRef] [PubMed]
101. Zhou, K.; Gaullier, G.; Luger, K. Nucleosome structure and dynamics are coming of age. *Nat. Struct. Mol. Biol.* **2019**, *26*, 3–13. [CrossRef] [PubMed]
102. Li, G.; Widom, J. Nucleosomes facilitate their own invasion. *Nat. Struct. Mol. Biol.* **2004**, *11*, 763–769. [CrossRef] [PubMed]
103. Van Den Bedem, H.; Fraser, J.S. Integrative, dynamic structural biology at atomic resolution-It's about time. *Nat. Methods* **2015**, *12*, 307–318. [CrossRef] [PubMed]
104. Shimada, I.; Ueda, T.; Kofuku, Y.; Eddy, M.T.; Wüthrich, K. GPCR drug discovery: Integrating solution NMR data with crystal and cryo-EM structures. *Nat. Rev. Drug Discov.* **2018**, *18*, 59–82. [CrossRef] [PubMed]
105. Gruber, D.R.; Toner, J.J.; Miears, H.L.; Shernyukov, A.V.; Kiryutin, A.S.; Lomzov, A.A.; Endutkin, A.V.; Grin, I.R.; Petrova, D.V.; Kupryushkin, M.S.; et al. Oxidative damage to epigenetically methylated sites affects DNA stability, dynamics and enzymatic demethylation. *Nucleic Acids Res.* **2018**, *46*, 10827–10839. [CrossRef] [PubMed]
106. Xiang, S.Q.; le Paige, U.B.; Horn, V.; Houben, K.; Baldus, M.; van Ingen, H. Site-specific studies of nucleosome interactions by solid-state NMR spectroscopy. *Angew. Chem. Int. Ed.* **2018**, *57*, 4571–4575. [CrossRef] [PubMed]

107. Houliston, R.S.; Lemak, A.; Iqbal, A.; Ivanochko, D.; Duan, S.; Kaustov, L.; Ong, M.S.; Fan, L.; Senisterra, G.; Brown, P.J.; et al. Conformational dynamics of the TTD-PHD histone reader module of the UHRF1 epigenetic regulator reveals multiple histone-binding states, allosteric regulation, and druggability. *J. Biol. Chem.* **2017**, *292*, 20947–20959. [CrossRef] [PubMed]
108. Jenuwein, T.; Allis, C.D. Translating the histone code. *Science* **2001**, *293*, 1074–1080. [CrossRef] [PubMed]
109. Fischle, W.; Wang, Y.; Allis, C.D. Histone and chromatin cross-talk. *Curr. Opin. Cell Biol.* **2003**, *15*, 172–183. [CrossRef]
110. Patel, D.J.; Wang, Z. Readout of epigenetic modifications. *Annu. Rev. Biochem.* **2013**, *82*, 81–118. [CrossRef] [PubMed]
111. Brower-Toland, B.; Wacker, D.A.; Fulbright, R.M.; Lis, J.T.; Kraus, W.L.; Wang, M.D. Specific contributions of histone tails and their acetylation to the mechanical stability of nucleosomes. *J. Mol. Biol.* **2005**, *346*, 135–146. [CrossRef] [PubMed]
112. Blundell, T.L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002**, *1*, 45–54. [CrossRef] [PubMed]
113. Van Roey, K.; Uyar, B.; Weatheritt, R.J.; Dinkel, H.; Seiler, M.; Budd, A.; Gibson, T.J.; Davey, N.E. Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **2014**, *114*, 6733–6778. [CrossRef] [PubMed]
114. Davis, A.M.; Teague, S.J.; Kleywegt, G.J. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed.* **2003**, *42*, 2718–2736. [CrossRef] [PubMed]
115. Srivastava, A.; Nagai, T.; Srivastava, A.; Miyashita, O.; Tama, F. Role of computational methods in going beyond x-ray crystallography to explore protein structure and dynamics. *Int. J. Mol. Sci.* **2018**, *19*, 3401. [CrossRef] [PubMed]
116. Marchanka, A.; Simon, B.; Althoff-Ospelt, G.; Carlomagno, T. RNA structure determination by solid-state NMR spectroscopy. *Nat. Commun.* **2015**, *6*, 1–7. [CrossRef] [PubMed]
117. Wang, J.; Williams, B.; Chirasani, V.R.; Krokhotin, A.; Das, R.; Dokholyan, N.V. Limits in accuracy and a strategy of RNA structure prediction using experimental information. *Nucleic Acids Res.* **2019**, *47*, 5563–5572. [CrossRef] [PubMed]
118. Van Emmerik, C.L.; van Ingen, H. Unspinning chromatin: Revealing the dynamic nucleosome landscape by NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2019**, *110*, 1–19. [CrossRef] [PubMed]
119. Zhao, S.; Yang, M.; Zhou, W.; Zhang, B.; Cheng, Z.; Huang, J.; Zhang, M.; Wang, Z.; Wang, R.; Chen, Z.; et al. Kinetic and high-throughput profiling of epigenetic interactions by 3D-carbene chip-based surface plasmon resonance imaging technology. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 7245–7254. [CrossRef] [PubMed]
120. Rajakumara, E.; Wang, Z.; Ma, H.; Hu, L.; Chen, H.; Lin, Y.; Guo, R.; Wu, F.; Li, H.; Lan, F.; et al. PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression. *Mol. Cell* **2011**, *43*, 275–284. [CrossRef] [PubMed]
121. Flanagan, T.W.; Brown, D.T. Molecular dynamics of histone H1. *Biochim. Biophys. Acta Gene Regul. Mech.* **2016**, *1859*, 468–475. [CrossRef] [PubMed]
122. Mikasa, T.; Kugo, M.; Nishimura, S.; Taketani, S.; Ishijima, S.; Sagami, I. Thermodynamic characterization of the Ca 2+ -dependent interaction between SOUL and ALG-2. *Int. J. Mol. Sci.* **2018**, *19*, 3802. [CrossRef] [PubMed]
123. Weaver, T.M.; Morrison, E.A.; Musselman, C.A. Reading more than histones: The prevalence of nucleic acid binding among reader domains. *Molecules* **2018**, *23*, 2614. [CrossRef] [PubMed]
124. Peach, C.J.; Mignone, V.W.; Arruda, M.A.; Alcobia, D.C.; Hill, S.J.; Kilpatrick, L.E.; Woolard, J. Molecular pharmacology of VEGF-A isoforms: Binding and signalling at VEGFR2. *Int. J. Mol. Sci.* **2018**, *19*, 1264. [CrossRef] [PubMed]
125. Daniel, J.A.; Pray-Grant, M.G.; Grant, P.A. Effector proteins for methylated histones: An expanding family. *Cell Cycle* **2005**, *4*, 919–926. [CrossRef] [PubMed]
126. Jeszenői, N.; Horváth, I.; Bálint, M.; Van Der Spoel, D.; Hetényi, C. Mobility-based prediction of hydration structures of protein surfaces. *Bioinformatics* **2015**, *31*, 1959–1965. [CrossRef] [PubMed]
127. Jeszenői, N.; Bálint, M.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Exploration of interfacial hydration networks of target-ligand complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148–158. [CrossRef] [PubMed]

128. Schwartzentruber, J.; Korshunov, A.; Liu, X.Y.; Jones, D.T.W.; Pfaff, E.; Jacob, K.; Sturm, D.; Fontebasso, A.M.; Quang, D.A.K.; Tönjes, M.; et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **2012**, *482*, 226–231. [CrossRef] [PubMed]
129. Wang, Y.; Han, Y.; Fan, E.; Zhang, K. Analytical strategies used to identify the readers of histone modifications: A review. *Anal. Chim. Acta* **2015**, *891*, 32–42. [CrossRef] [PubMed]
130. Org, T.; Chignola, F.; Hetényi, C.; Gaetani, M.; Rebane, A.; Liiv, I.; Maran, U.; Mollica, L.; Bottomley, M.J.; Musco, G.; et al. The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. *EMBO Rep.* **2008**, *9*, 370–376. [CrossRef] [PubMed]
131. Kim, S.; Natesan, S.; Cornilescu, G.; Carlson, S.; Tonelli, M.; McClurg, U.L.; Binda, O.; Robson, C.N.; Markley, J.L.; Balaz, S.; et al. Mechanism of histone H3K4me3 recognition by the plant homeodomain of inhibitor of growth 3. *J. Biol. Chem.* **2016**, *291*, 18326–18341. [CrossRef] [PubMed]
132. Kostrhon, S.; Kontaxis, G.; Kaufmann, T.; Schirghuber, E.; Kubicek, S.; Konrat, R.; Slade, D. A histone-mimicking interdomain linker in a multidomain protein modulates multivalent histone binding. *J. Biol. Chem.* **2017**, *292*, 17643–17657. [CrossRef] [PubMed]
133. Bortoluzzi, A.; Amato, A.; Lucas, X.; Blank, M.; Ciulli, A. Structural basis of molecular recognition of helical histone H3 tail by PHD finger domains. *Biochem. J.* **2017**, *474*, 1633–1651. [CrossRef] [PubMed]
134. Liu, Y.; Qin, S.; Chen, T.Y.; Lei, M.; Dhar, S.S.; Ho, J.C.; Dong, A.; Loppnau, P.; Li, Y.; Lee, M.G.; et al. Structural insights into trans-histone regulation of H3K4 methylation by unique histone H4 binding of MLL3/4. *Nat. Commun.* **2019**, *10*, 36–47. [CrossRef] [PubMed]
135. Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nat. Rev. Drug Discov.* **2015**, *14*, 95–110. [CrossRef] [PubMed]
136. Horváth, I.; Jeszenői, N.; Bálint, M.; Paragi, G.; Hetényi, C. A Fragmenting protocol with explicit hydration for calculation of binding enthalpies of target-ligand complexes at a quantum mechanical level. *Int. J. Mol. Sci.* **2019**, *20*, 4384. [CrossRef] [PubMed]
137. Freire, E. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov. Today* **2008**, *13*, 869–874. [CrossRef] [PubMed]
138. Ohtaka, H.; Freire, E. Adaptive inhibitors of the HIV-1 protease. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 193–208. [CrossRef] [PubMed]
139. Carbonell, T.; Freire, E. Binding thermodynamics of statins to HMG-CoA reductase. *Biochemistry* **2005**, *44*, 11741–11748. [CrossRef] [PubMed]
140. Perniola, R.; Musco, G. The biophysical and biochemical properties of the autoimmune regulator (AIRE) protein. *Biochim. Biophys. Acta Mol. Basis Dis.* **2014**, *1842*, 326–337. [CrossRef] [PubMed]
141. Corbeski, I.; Dolinar, K.; Wienk, H.; Boelens, R.; Van Ingen, H. DNA repair factor APLF acts as a H2A-H2B histone chaperone through binding its DNA interaction surface. *Nucleic Acids Res.* **2018**, *46*, 7138–7152. [CrossRef] [PubMed]
142. Copeland, R.A. The drug-target residence time model: A 10-year retrospective. *Nat. Rev. Drug. Disc.* **2016**, *15*, 87–95. [CrossRef] [PubMed]
143. Zhao, S.; Zhang, B.; Yang, M.; Zhu, J.; Li, H. Systematic profiling of histone readers in *Arabidopsis thaliana*. *Cell Rep.* **2018**, *22*, 1090–1102. [CrossRef] [PubMed]
144. Hendershot, J.M.; O'Brien, P.J. Transient kinetic methods for mechanistic characterization of dna binding and nucleotide flipping. *Methods Enzym.* **2017**, *592*, 377–415.
145. Cui, Y.; Cho, I.H.; Chowdhury, B.; Irudayaraj, J. Real-time dynamics of methyl-CpG-binding domain protein 3 and its role in DNA demethylation by fluorescence correlation spectroscopy. *Epigenetics* **2013**, *8*, 1089–1100. [CrossRef] [PubMed]
146. Gunther, J.R.; Du, Y.; Rhoden, E.; Lewis, I.; Revennaugh, B.; Moore, T.W.; Kim, S.H.; Dingledine, R.; Fu, H.; Katzenellenbogen, J.A. A set of time-resolved fluorescence resonance energy transfer assays for the discovery of inhibitors of estrogen receptor-coactivator binding. *J. Biomol. Screen.* **2009**, *14*, 181–193. [CrossRef] [PubMed]
147. Hardwidge, P.R.; Parkhurst, K.M.; Parkhurst, L.J.; Maher, L.J. Reflections on apparent DNA bending by charge variants of bZIP proteins. *Biopolymers* **2003**, *69*, 110–117. [CrossRef] [PubMed]
148. Luo, Y.; North, J.A.; Poirier, M.G. Single molecule fluorescence methodologies for investigating transcription factor binding kinetics to nucleosomes and DNA. *Methods* **2014**, *70*, 108–118. [CrossRef] [PubMed]

149. Dias, D.M.; Ciulli, A. NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes. *Prog. Biophys. Mol. Biol.* **2014**, *116*, 101–112. [CrossRef] [PubMed]
150. Cheng, H.C. The power issue: Determination of KB or Ki from IC50-A closer look at the Cheng-Prusoff equation, the schild plot and related power equations. *J. Pharm. Toxicol. Methods* **2001**, *46*, 61–71. [CrossRef]
151. Chan-Penebre, E.; Kuplast, K.G.; Majer, C.R.; Boriack-Sjodin, P.A.; Wigle, T.J.; Johnston, L.D.; Rioux, N.; Munchhof, M.J.; Jin, L.; Jacques, S.L.; et al. A selective inhibitor of PRMT5 with in vivo and in vitro potency in MCL models. *Nat. Chem. Biol.* **2015**, *11*, 432–437. [CrossRef] [PubMed]
152. Meyners, C.; Meyer-Almes, F.J. Impact of binding mechanism on selective inhibition of histone deacetylase isoforms. *Chem. Biol. Drug Des.* **2017**, *90*, 1215–1225. [CrossRef] [PubMed]
153. Ma, F.; Jiang, S.; Zhang, C. yang Recent advances in histone modification and histone modifying enzyme assays. *Expert Rev. Mol. Diagn.* **2019**, *19*, 27–36. [CrossRef] [PubMed]
154. Wen, Q.; Gu, Y.; Tang, L.J.; Yu, R.Q.; Jiang, J.H. Peptide-templated gold nanocluster beacon as a sensitive, label-free sensor for protein post-translational modification enzymes. *Anal. Chem.* **2013**, *85*, 11681–11685. [CrossRef] [PubMed]
155. Yufang, H.; Siyu, C.; Yitao, H.; Hongjun, C.; Qin, W.Z.; Nie, Y.; Huang, S.Y. Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors. *Chem. Commun.* **2015**, *51*, 17611–17614.
156. Wang, Y.X.D.; Liu, J.T.; Jiang, H.J.J. Surface enhanced raman scattering based sensitive detection of histone demethylase activity using formaldehyde-selective reactive probe. *Chem. Commun.* **2013**, *49*, 8489. [CrossRef] [PubMed]
157. Guo, X.; Wang, L.; Li, J.; Ding, Z.; Xiao, J.; Yin, X.; He, S.; Shi, P.; Dong, L.; Li, G.; et al. Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature* **2015**, *517*, 640–644. [CrossRef] [PubMed]
158. Müller, M.M.; Fierz, B.; Bittova, L.; Liszczak, G.; Muir, T.W. A two-state activation mechanism controls the histone methyltransferase Suv39h1. *Nat. Chem. Biol.* **2016**, *12*, 188–193. [CrossRef] [PubMed]
159. Bauden, M.; Tassidis, H.; Ansari, D. In vitro cytotoxicity evaluation of HDAC inhibitor apicidin in pancreatic carcinoma cells subsequent time and dose dependent treatment. *Toxicol. Lett.* **2015**, *236*, 8–15. [CrossRef] [PubMed]
160. Liu, W.; Cui, Y.; Ren, W.; Irudayaraj, J. Epigenetic biomarker screening by FLIM-FRET for combination therapy in ER+ breast cancer. *Clin. Epigenetics* **2019**, *11*, 16. [CrossRef] [PubMed]
161. Wegener, D.; Deubzer, H.E.; Oehme, I.; Milde, T.; Hildmann, C.; Schwienhorst, A.; Witt, O. HKI 46F08, a novel potent histone deacetylase inhibitor, exhibits antitumoral activity against embryonic childhood cancer cells. *Anticancer Drugs* **2008**, *19*, 849–857. [CrossRef] [PubMed]
162. Wang, Y.M.; Gu, M.L.; Meng, F.S.; Jiao, W.R.; Zhou, X.X.; Yao, H.P.; Ji, F. Histone acetyltransferase p300/CBP inhibitor C646 blocks the survival and invasion pathways of gastric cancer cell lines. *Int. J. Oncol.* **2017**, *51*, 1860–1868. [CrossRef] [PubMed]
163. Gu, M.L.; Wang, Y.M.; Zhou, X.X.; Yao, H.P.; Zheng, S.; Xiang, Z.; Ji, F. An inhibitor of the acetyltransferases CBP/p300 exerts antineoplastic effects on gastrointestinal stromal tumor cells. *Oncol. Rep.* **2016**, *36*, 2763–2770. [CrossRef] [PubMed]
164. Ou, Y.; Wilson, R.E.; Weber, S.G. Methods of measuring enzyme activity ex vivo and in vivo. *Annu. Rev. Anal. Chem.* **2018**, *11*, 509–533. [CrossRef] [PubMed]
165. Louie, A.Y.; Hüber, M.M.; Ahrens, E.T.; Rothbächer, U.; Moats, R.; Jacobs, R.E.; Fraser, S.E.; Meade, T.J. In vivo visualization of gene expression using magnetic resonance imaging. *Nat. Biotechnol.* **2000**, *18*, 321–325. [CrossRef] [PubMed]
166. Södersten, E.; Toskas, K.; Rraklli, V.; Tiklova, K.; Björklund, Å.K.; Ringnér, M.; Perlmann, T.; Holmberg, J. A comprehensive map coupling histone modifications with gene regulation in adult dopaminergic and serotonergic neurons. *Nat. Commun.* **2018**, *9*, 1226–1242. [CrossRef] [PubMed]
167. Wang, Z.; Wu, J.; Humphries, B.; Kondoe, K.; Jiang, Y.; Shi, X.; Yang, C. Upregulation of histone-lysine methyltransferases plays a causal role in hexavalent chromium-induced cancer stem cell-like property and cell transformation. *Toxicol. Appl. Pharm.* **2018**, *342*, 22–30. [CrossRef] [PubMed]

168. Pascual, M.; Boix, J.; Felipo, V.; Guerri, C. Repeated alcohol administration during adolescence causes changes in the mesolimbic dopaminergic and glutamatergic systems and promotes alcohol intake in the adult rat. *J. Neurochem.* **2009**, *108*, 920–931. [CrossRef] [PubMed]
169. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
170. Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441. [CrossRef] [PubMed]
171. Ryde, U.; Söderhjelm, P. Ligand-binding affinity estimates supported by quantum-mechanical methods. *Chem. Rev.* **2016**, *116*, 5520–5566. [CrossRef] [PubMed]
172. Gaillard, T.; Panel, N.; Simonson, T. Protein side chain conformation predictions with an MMGBSA energy function. *Proteins Struct. Funct. Bioinforma.* **2016**, *84*, 803–819. [CrossRef] [PubMed]
173. Avgy-David, H.H.; Senderowitz, H. Toward focusing conformational ensembles on bioactive conformations: A molecular mechanics/quantum mechanics study. *J. Chem. Inf. Model.* **2015**, *55*, 2154–2167. [CrossRef] [PubMed]
174. Schindler, C.E.M.; de Vries, S.J.; Zacharias, M. iATTRACT: Simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 248–258. [CrossRef] [PubMed]
175. 175. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, 506–515. [CrossRef] [PubMed]
176. Lu, W.; Zhang, R.; Jiang, H.; Zhang, H.; Luo, C. Computer-aided drug design in epigenetics. *Front. Chem.* **2018**, *6*, 57. [CrossRef] [PubMed]
177. Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.* **2018**, *46*, 356–362. [CrossRef] [PubMed]
178. Di Luccio, E. Inhibition of nuclear receptor binding SET domain 2/multiple myeloma SET domain by LEM-06 implication for epigenetic cancer therapies. *J. Cancer Prev.* **2015**, *20*, 113–120. [CrossRef] [PubMed]
179. Sekhavat, A.; Sun, J.M.; Davie, J.R. Competitive inhibition of histone deacetylase activity by trichostatin A and butyrate. *Biochem. Cell Biol.* **2007**, *85*, 751–758. [CrossRef] [PubMed]
180. Tomar, J.S.; Peddinti, R.K. *A. baumannii* histone acetyl transferase Hpa2: Optimization of homology modeling, analysis of protein-protein interaction and virtual screening. *J. Biomol. Struct. Dyn.* **2017**, *35*, 1115–1126. [CrossRef] [PubMed]
181. Uba, A.I.; Yelekçi, K. Homology modeling of human histone deacetylase 10 and design of potential selective inhibitors. *J. Biomol. Struct. Dyn.* **2019**, *37*, 3627–3636.
182. Zhao, M.L.; Wang, W.; Nie, H.; Cao, S.S.; Du, L.F. In silico structure prediction and inhibition mechanism studies of AtHDA14 as revealed by homology modeling, docking, molecular dynamics simulation. *Comput. Biol. Chem.* **2018**, *75*, 120–130. [CrossRef] [PubMed]
183. Sixto-López, Y.; Bello, M.; Rodríguez-Fonseca, R.A.; Rosales-Hernández, M.C.; Martínez-Archundia, M.; Gómez-Vidal, J.A.; Correa-Basurto, J. Searching the conformational complexity and binding properties of HDAC6 through docking and molecular dynamic simulations. *J. Biomol. Struct. Dyn.* **2017**, *35*, 2794–2814. [CrossRef] [PubMed]
184. Park, I.; Hwang, Y.J.; Kim, T.H.; Viswanath, A.N.I.; Londhe, A.M.; Jung, S.Y.; Sim, K.M.; Min, S.J.; Lee, J.E.; Seong, J.; et al. In silico probing and biological evaluation of SETDB1/ESET-targeted novel compounds that reduce tri-methylated histone H3K9 (H3K9me3) level. *J. Comput. Aided. Mol. Des.* **2017**, *31*, 877–889. [CrossRef] [PubMed]
185. Scholte, L.L.S.; Mourão, M.M.; Pais, F.S.M.; Melesina, J.; Robaa, D.; Volpini, A.C.; Sippl, W.; Pierce, R.J.; Oliveira, G.; Nahum, L.A. Evolutionary relationships among protein lysine deacetylases of parasites causing neglected diseases. *Infect. Genet. Evol.* **2017**, *53*, 175–188. [CrossRef] [PubMed]
186. Park, H.; Kim, S.; Kim, Y.E.; Lim, S.J. A structure-based virtual screening approach toward the discovery of histone deacetylase inhibitors: Identification of promising zinc-chelating groups. *Chem. Med. Chem.* **2010**, *5*, 591–597. [CrossRef] [PubMed]
187. Iwamori, N.; Tominaga, K.; Sato, T.; Riehle, K.; Iwamori, T.; Ohkawa, Y.; Coarfa, C.; Ono, E.; Matzuk, M.M. MRG15 is required for pre-mRNA splicing and spermatogenesis. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 5408–5415. [CrossRef] [PubMed]

188. Wu, X.; Fang, Z.; Yang, B.; Zhong, L.; Yang, Q.; Zhang, C.; Huang, S.; Xiang, R.; Suzuki, T.; Li, L.L.; et al. Discovery of KDM5A inhibitors: Homology modeling, virtual screening and structure-activity relationship analysis. *Bioorganic Med. Chem. Lett.* **2016**, *26*, 2284–2288. [CrossRef] [PubMed]
189. Kannan, A.; Camilloni, C.; Sahakyan, A.B.; Cavalli, A.; Vendruscolo, M. A conformational ensemble derived using NMR methyl chemical shifts reveals a mechanical clamping transition that gates the binding of the HU protein to dna. *J. Am. Chem. Soc.* **2014**, *136*, 2204–2207. [CrossRef] [PubMed]
190. Tang, H.; Wang, X.S.; Huang, X.P.; Roth, B.L.; Butler, K.V.; Kozikowski, A.P.; Jung, M.; Tropsha, A. Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J. Chem. Inf. Model.* **2009**, *49*, 461–476. [CrossRef] [PubMed]
191. Alves-Avelar, L.A.; Held, J.; Engel, J.A.; Sureechatchaiyan, P.; Hansen, F.K.; Hamacher, A.; Kassack, M.U.; Mordmüller, B.; Andrews, K.T.; Kurz, T. Design and synthesis of novel anti-plasmodial histone deacetylase inhibitors containing an alkoxyamide connecting unit. *Arch. Pharm.* **2017**, *350*, 1600347. [CrossRef]
192. Melesina, J.; Robaa, D.; Pierce, R.J.; Romier, C.; Sippl, W. Homology modeling of parasite histone deacetylases to guide the structure-based design of selective inhibitors. *J. Mol. Graph. Model.* **2015**, *62*, 342–361. [CrossRef]
193. Fiser, A.; Gian-Do, R.K.; Ŝali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753–1773. [CrossRef] [PubMed]
194. Hillringhaus, L.; Yue, W.W.; Rose, N.R.; Ng, S.S.; Gileadi, C.; Loenarz, C.; Bello, S.H.; Bray, J.E.; Schofield, C.J.; Oppermann, U. Structural and evolutionary basis for the dual substrate selectivity of human KDM4 histone demethylase family. *J. Biol. Chem.* **2011**, *286*, 41616–41625. [CrossRef] [PubMed]
195. Bottomley, M.J.; Stier, G.; Pennacchini, D.; Legube, G.; Simon, B.; Akhtar, A.; Sattler, M.; Musco, G. NMR structure of the first PHD finger of autoimmune regulator protein (AIRE1): Insights into autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) disease. *J. Biol. Chem.* **2005**, *280*, 11505–11512. [CrossRef] [PubMed]
196. Madsent, D.; Kleywegt, G.J. Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.* **2002**, *35*, 137–139. [CrossRef]
197. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447. [CrossRef] [PubMed]
198. Ozbaykal, G.; Atilgan, A.R.; Atilgan, C. In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 2077–2090. [CrossRef] [PubMed]
199. He, M.; Li, W.; Zheng, Q.; Zhang, H. A molecular dynamics investigation into the mechanisms of alectinib resistance of three ALK mutants. *J. Cell. Biochem.* **2018**, *119*, 5332–5342. [CrossRef]
200. Li, M.; Simonetti, F.L.; Goncearenco, A.; Panchenko, A.R. MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic Acids Res.* **2016**, *44*, 494–501. [CrossRef] [PubMed]
201. Rentzsch, R.; Renard, B.Y. Docking small peptides remains a great challenge: An assessment using AutoDock vina. *Brief. Bioinform.* **2015**, *16*, 1045–1056. [CrossRef] [PubMed]
202. Hauser, A.S.; Windshügel, B. LEADS-PEP: A benchmark data set for assessment of peptide docking performance. *J. Chem. Inf. Model.* **2016**, *56*, 188–200. [CrossRef] [PubMed]
203. Antunes, D.A.; Devaurs, D.; Kavraki, L.E. Understanding the challenges of protein flexibility in drug design. *Expert Opin. Drug Discov.* **2015**, *10*, 1301–1313. [CrossRef] [PubMed]
204. Antunes, D.A.; Moll, M.; Devaurs, D.; Jackson, K.R.; Lizée, G.; Kavraki, L.E. DINC 2.0: A new protein-peptide docking webserver using an incremental approach. *Cancer Res.* **2017**, *77*, e55–e57. [CrossRef] [PubMed]
205. Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein-peptide docking: Opportunities and challenges. *Drug Discov. Today* **2018**, *23*, 1530–1537. [CrossRef] [PubMed]
206. Pevzner, Y.; Frugier, E.; Schalk, V.; Caflisch, A.; Woodcock, H.L. Fragment-based docking: Development of the CHARMMing web user interface as a platform for computer-aided drug design. *J. Chem. Inf. Model.* **2014**, *54*, 2612–2620. [CrossRef] [PubMed]
207. London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. Rosetta FlexPepDock web server-high resolution modeling of peptide-protein interactions. *Nucleic Acids Res.* **2011**, *39*, 249–253. [CrossRef] [PubMed]
208. Dominguez, C.; Boelens, R.; Bonvin, A.M.J.J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737. [CrossRef] [PubMed]

209. Lee, H.; Heo, L.; Lee, M.S.; Seok, C. GalaxyPepDock: A protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res.* **2015**, *43*, 431–435. [CrossRef] [PubMed]
210. Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **2015**, *43*, 419–424. [CrossRef] [PubMed]
211. Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P. PEP-FOLD3: Faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res.* **2016**, *44*, 44, 449–454. [CrossRef] [PubMed]
212. Agostini, F.; Zanzoni, A.; Klus, P.; Marchese, D.; Cirillo, D.; Tartaglia, G.G. CatRAPID omics: A web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* **2013**, *29*, 2928–2930. [CrossRef] [PubMed]
213. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H.J. PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Res.* **2005**, *33*, 363–367. [CrossRef] [PubMed]
214. Weng, G.; Wang, E.; Wang, Z.; Liu, H.; Zhu, F.; Li, D.; Hou, T. HawkDock: A web server to predict and analyze the protein-protein complex based on computational docking and MM/GBSA. *Nucleic Acids Res.* **2019**, *47*, 322–330. [CrossRef] [PubMed]
215. Labbe, C.M.; Pencheva, T.; Jereva, D.; Desvillechabrol, D.; Becot, J.; Villoutreix, B.O.; Pajeva, I.; Miteva, M.A. AMMOS2: A web server for protein-ligand-water complexes refinement via molecular mechanics. *Nucleic Acids Res.* **2017**, *45*, 350–355. [CrossRef] [PubMed]
216. Sen, S.; Sanyal, S.; Srivastava, D.K.; Dasgupta, D.; Roy, S.; Das, C. Transcription factor 19 interacts with histone 3 lysine 4 trimethylation and controls gluconeogenesis via the nucleosome-remodeling-deacetylase complex. *J. Biol. Chem.* **2017**, *292*, 20362–20378. [CrossRef] [PubMed]
217. Shao, Z.; Xu, P.; Xu, W.; Li, L.; Liu, S.; Zhang, R.; Liu, Y.C.; Zhang, C.; Chen, S.; Luo, C. Discovery of novel DNA methyltransferase 3A inhibitors via structure-based virtual screening and biological assays. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 342–346. [CrossRef] [PubMed]
218. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. [CrossRef] [PubMed]
219. Allen, W.J.; Balius, T.E.; Mukherjee, S.; Brozell, S.R.; Moustakas, D.T.; Lang, P.T.; Case, D.A.; Kuntz, I.D.; Rizzo, R.C. DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* **2015**, *36*, 1132–1156. [CrossRef] [PubMed]
220. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef] [PubMed]
221. Chen, S.; Wang, Y.; Zhou, W.; Li, S.; Peng, J.; Shi, Z.; Hu, J.; Liu, Y.C.; Ding, H.; Lin, Y.; et al. Identifying novel selective non-nucleoside DNA methyltransferase 1 inhibitors through docking-based virtual screening. *J. Med. Chem.* **2014**, *57*, 9028–9041. [CrossRef] [PubMed]
222. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.E.; Berendsen, H.J.C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718. [CrossRef] [PubMed]
223. Stocker, U.; Van Gunsteren, W.F. Molecular dynamics simulation of hen egg white lysozyme: A test of the GROMOS96 force field against nuclear magnetic resonance data. *Proteins Struct. Funct. Genet.* **2000**, *40*, 145–153. [CrossRef]
224. Huang, D.; Rossini, E.; Steiner, S.; Caflisch, A. Structured water molecules in the binding site of bromodomains can be displaced by cosolvent. *Chem. Med. Chem.* **2014**, *9*, 573–579. [CrossRef] [PubMed]
225. Ulucan, O.; Keskin, O.; Erman, B.; Gursoy, A. A comparative molecular dynamics study of methylation state specificity of JMJD2A. *PLoS ONE* **2011**, *6*, 24664–24673. [CrossRef] [PubMed]
226. De Ruiter, A.; Oostenbrink, C. Advances in the calculation of binding free energies. *Curr. Opin. Struct. Biol.* **2020**, *61*, 207–212. [CrossRef] [PubMed]
227. Kukol, A. *Molecular Modeling of Proteins, 2nd ed*; Humana Press: New York, NY, USA, 2014; pp. 1–474.
228. Bianchi, C.; Zangi, R. Molecular dynamics study of the recognition of dimethylated CpG sites by MBD1 protein. *J. Chem. Inf. Model.* **2015**, *55*, 636–644. [CrossRef] [PubMed]
229. Gao, C.; Herold, J.M.; Kireev, D. Assessment of free energy predictors for ligand binding to a methyllysine histone code reader. *J. Comput. Chem.* **2012**, *33*, 659–665. [CrossRef] [PubMed]

230. Ikebe, J.; Sakuraba, S.; Kono, H. H3 histone tail conformation within the nucleosome and the impact of K14 acetylation studied using enhanced sampling simulation. *PLoS Comput. Biol.* **2016**, *12*, e1004788. [CrossRef] [PubMed]
231. Heinzelmann, G.; Henriksen, N.M.; Gilson, M.K. Attach-pull-release calculations of ligand binding and conformational changes on the first BRD4 bromodomain. *J. Chem. Theory Comput.* **2017**, *13*, 3260–3275. [CrossRef] [PubMed]
232. Berendsen, H.J.C.; Grigera, R.J.; Straatsma, P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271. [CrossRef]
233. Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260. [CrossRef] [PubMed]
234. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
235. Horn, H.W.; Swope, W.C.; Pitera, J.W.; Madura, J.D.; Dick, T.J.; Hura, G.L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678. [CrossRef] [PubMed]
236. Choubey, S.K.; Jeyaraman, J. A mechanistic approach to explore novel HDAC1 inhibitor using pharmacophore modeling, 3D-QSAR analysis, molecular docking, density functional and molecular dynamics simulation study. *J. Mol. Graph. Model.* **2016**, *70*, 54–69. [CrossRef] [PubMed]
237. Choubey, S.K.; Jeyakanthan, J. Molecular dynamics and quantum chemistry-based approaches to identify isoform selective HDAC2 inhibitor-a novel target to prevent Alzheimer's disease. *J. Recept. Signal Transduct.* **2018**, *38*, 266–278. [CrossRef] [PubMed]
238. Karthi, N.; Karthiga, A.; Kalaiyarasu, T.; Stalin, A.; Manju, V.; Singh, S.K.; Cyril, R.; Lee, S.M. Exploration of cell cycle regulation and modulation of the DNA methylation mechanism of pelargonidin: Insights from the molecular modeling approach. *Comput. Biol. Chem.* **2017**, *70*, 175–185. [CrossRef] [PubMed]
239. Mitra, S.; Dash, R. Structural dynamics and quantum mechanical aspects of shikonin derivatives as CREBBP bromodomain inhibitors. *J. Mol. Graph. Mod.* **2018**, *83*, 42–52. [CrossRef] [PubMed]
240. Raj, U.; Kumar, H.; Varadwaj, P.K. Molecular docking and dynamics simulation study of flavonoids as BET bromodomain inhibitors. *J. Biomol. Struct. Dyn.* **2017**, *35*, 2351–2362. [CrossRef] [PubMed]
241. Ran, T.; Zhang, Z.; Liu, K.; Lu, Y.; Li, H.; Xu, J.; Xiong, X.; Zhang, Y.; Xu, A.; Lu, S.; et al. Insight into the key interactions of bromodomain inhibitors based on molecular docking, interaction fingerprinting, molecular dynamics and binding free energy calculation. *Mol. Biosyst.* **2015**, *11*, 1295–1304. [CrossRef] [PubMed]
242. Sivanandam, M.; Manjula, S.; Kumaradhas, P. Investigation of activation mechanism and conformational stability of N-(4-chloro-3-trifluoromethyl-phenyl)-2-ethoxybenzamide and N-(4-chloro-3-trifluoromethyl-phenyl)-2-ethoxy-6-pentadecyl-benzamide in the: Active site of p300 histone acetyl transferase. *J. Biomol. Struct. Dyn.* **2019**, *37*, 4006–4018. [CrossRef] [PubMed]
243. Suryanarayanan, V.; Singh, S.K. Unravelling novel congeners from acetyllysine mimicking ligand targeting a lysine acetyltransferase PCAF bromodomain. *J. Biomol. Struct. Dyn.* **2018**, *36*, 4303–4319. [CrossRef] [PubMed]
244. Yuan, Y.; Hu, Z.; Bao, M.; Sun, R.; Long, X.; Long, L.; Li, J.; Wu, C.; Bao, J. Screening of novel histone deacetylase 7 inhibitors through molecular docking followed by a combination of molecular dynamics simulations and ligand-based approach. *J. Biomol. Struct. Dyn.* **2018**, *37*, 4092–4103. [CrossRef] [PubMed]
245. Mallik, B.S.; Pai, A.; Shenoy, R.R.; Jayashree, B.S. Novel flavonol analogues as potential inhibitors of JMJD3 histone demethylase—A study based on molecular modelling. *J. Mol. Graph. Model.* **2017**, *72*, 81–87. [CrossRef] [PubMed]
246. Chen, L.; Zheng, Q.C.; Zhang, H.X. Insights into the effects of mutations on Cren7-DNA binding using molecular dynamics simulations and free energy calculations. *Phys. Chem. Chem. Phys.* **2015**, *17*, 5704–5711. [CrossRef] [PubMed]
247. Grauffel, C.; Stote, R.H.; Dejaegere, A. Molecular dynamics for computational proteomics of methylated histone H3. *Biochim. Biophys. Acta Gen. Subj.* **2015**, *1850*, 1026–1040. [CrossRef] [PubMed]
248. Hassanzadeh, M.; Bagherzadeh, K.; Amanlou, M. A comparative study based on docking and molecular dynamics simulations over HDAC-tubulin dual inhibitors. *J. Mol. Graph. Model.* **2016**, *70*, 170–180. [CrossRef] [PubMed]

249. Singh, A.N.; Sharma, N. Epigenetic modulators as potential multi-targeted drugs against hedgehog pathway for treatment of cancer. *Protein J.* **2019**, *38*, 537–550. [CrossRef] [PubMed]
250. Tambunan, U.S.F.; Wulandari, E.K. Identification of a better Homo sapiens Class II HDAC inhibitor through binding energy calculations and descriptor analysis. *Bmc Bioinform.* **2010**, *11*, 16–26. [CrossRef] [PubMed]
251. Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461. [CrossRef] [PubMed]
252. Aldeghi, M.; Bodkin, M.J.; Knapp, S.; Biggin, P.C. Statistical analysis on the performance of molecular mechanics poisson-boltzmann surface area versus absolute binding free energy calculations: Bromodomains as a case study. *J. Chem. Inf. Model.* **2017**, *57*, 2203–2221. [CrossRef] [PubMed]
253. Shi, J.; Vakoc, C.R. The mechanisms behind the therapeutic activity of BET bromodomain inhibition. *Mol. Cell* **2014**, *54*, 728–736. [CrossRef] [PubMed]
254. Filippakopoulos, P.; Knapp, S. Targeting bromodomains: Epigenetic readers of lysine acetylation. *Nat. Rev. Drug Discov.* **2014**, *13*, 337–356. [CrossRef] [PubMed]
255. Huang, S.; Song, C.; Wang, X.; Zhang, G.; Wang, Y.; Jiang, X.; Sun, Q.; Huang, L.; Xiang, R.; Hu, Y.; et al. Discovery of new SIRT2 Inhibitors by utilizing a consensus docking/scoring strategy and structure-activity relationship analysis. *J. Chem. Inf. Model.* **2017**, *57*, 669–679. [CrossRef] [PubMed]
256. Ballante, F.; Marshall, G.R. An automated strategy for binding-pose selection and docking assessment in structure-based drug design. *J. Chem. Inf. Model.* **2016**, *56*, 54–72. [CrossRef] [PubMed]
257. Wang, Y.; Yang, J.; Hong, T.; Chen, X.; Cui, L. SIRT2: Controversy and multiple roles in disease and physiology. *Ageing Res. Rev.* **2019**, *55*, 100961–100976. [CrossRef] [PubMed]
258. Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. Semiempirical calculations of binding enthalpy for protein-ligand complexes. *Int. J. Quantum Chem.* **2004**, *97*, 747–763. [CrossRef]
259. Nikitina, E.; Sulimov, V.; Grigoriev, F.; Kondakova, O.; Luschenka, S. Mixed implicit/explicit solvation modelsin quantum mechanical calculations of binding enthalpy for protein-ligand complexes. *Int. J. Quantum Chem.* **2006**, *106*, 1943–1963. [CrossRef]
260. Sermer, D.; Pasqualucci, L.; Wendel, H.G.; Melnick, A.; Younes, A. Emerging epigenetic-modulating therapies in lymphoma. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 494–507. [CrossRef] [PubMed]
261. Makita, S.; Tobinai, K. Targeting EZH2 with tazemetostat. *Lancet Oncol.* **2018**, *19*, 586–587. [CrossRef]
262. Italiano, A.; Soria, J.C.; Toulmonde, M.; Michot, J.M.; Lucchesi, C.; Varga, A.; Coindre, J.M.; Blakemore, S.J.; Clawson, A.; Suttle, B.; et al. Tazemetostat, an EZH2 inhibitor, in relapsed or refractory B-cell non-Hodgkin lymphoma and advanced solid tumours: A first-in-human, open-label, phase 1 study. *Lancet Oncol.* **2018**, *19*, 649–659. [CrossRef]
263. Mohammad, F.; Weissmann, S.; Leblanc, B.; Pandey, D.P.; Højfeldt, J.W.; Comet, I.; Zheng, C.; Johansen, J.V.; Rapin, N.; Porse, B.T.; et al. EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. *Nat. Med.* **2017**, *23*, 483–492. [CrossRef] [PubMed]
264. Kim, H.K.; Roberts, C.W.M. Targeting EZH2 in cancer. *Nat. Med.* **2016**, *22*, 128–134. [CrossRef] [PubMed]
265. Kim, Y.H.; Bagot, M.; Pinter-Brown, L.; Rook, A.H.; Porcu, P.; Horwitz, S.M.; Whittaker, S.; Tokura, Y.; Vermeer, M.; Zinzani, P.L.; et al. Mogamulizumab versus vorinostat in previously treated cutaneous T-cell lymphoma (MAVORIC): An international, open-label, randomised, controlled phase 3 trial. *Lancet Oncol.* **2018**, *19*, 1192–1204. [CrossRef]
266. Zagni, C.; Floresta, G.; Monciino, G.; Rescifina, A. The Search for potent, small-molecule HDACIs in cancer treatment: A decade after vorinostat. *Med. Res. Rev.* **2017**, *37*, 1373–1428. [CrossRef] [PubMed]
267. Grant, S.; Easley, C.; Kirkpatrick, P. Vorinostat. *Nat. Rev. Drug Discov.* **2007**, *6*, 21–22. [CrossRef] [PubMed]
268. Banerjee, S.N.; Moore, D.W.; Broker, T.R.; Chow, L.T. Vorinostat, a pan-HDAC inhibitor, abrogates productive HPV-18 DNA amplification. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11138–11147. [CrossRef] [PubMed]
269. Wang, L.; de Oliveira, L.R.; Huijberts, S.; Bosdriesz, E.; Pencheva, N.; Brunen, D.; Bosma, A.; Song, J.Y.; Zevenhoven, J.; Los-de Vries, G.T.; et al. An acquired vulnerability of drug-resistant melanoma with therapeutic potential. *Cell* **2018**, *173*, 1413–1425. [CrossRef] [PubMed]

D9

EMBO reports

EMBO open

scientific report

The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression

Tõnis Org[1], Francesca Chignola[2], Csaba Hetényi[3], Massimiliano Gaetani[2], Ana Rebane[1], Ingrid Liiv[1], Uko Maran[3], Luca Mollica[2], Matthew J. Bottomley[4], Giovanna Musco[2+] & Pärt Peterson[1++]

[1]Molecular Pathology, University of Tartu, Tartu, Estonia, [2]Biomolecular NMR Laboratory, Dulbecco Telethon Institute c/o S. Raffaele Scientific Institute, Milan, Italy, [3]Institute of Chemical Physics, University of Tartu, Tartu, Estonia, and [4]Istituto di Ricerche di Biologia Molecolare, Pomezia (Rome), Italy

Mutations in the gene autoimmune regulator (*AIRE*) cause autoimmune polyendocrinopathy candidiasis ectodermal dystrophy. AIRE is expressed in thymic medullary epithelial cells, where it promotes the expression of tissue-restricted antigens. By the combined use of biochemical and biophysical methods, we show that AIRE selectively interacts with histone H3 through its first plant homeodomain (PHD) finger (AIRE–PHD1) and preferentially binds to non-methylated H3K4 (H3K4me0). Accordingly, *in vivo* AIRE binds to and activates promoters containing low levels of H3K4me3 in human embryonic kidney 293 cells. We conclude that AIRE–PHD1 is an important member of a newly identified class of PHD fingers that specifically recognize H3K4me0, thus providing a new link between the status of histone modifications and the regulation of tissue-restricted antigen expression in thymus.

Keywords: AIRE; negative selection; NMR; protein structure

EMBO reports (2008) **9**, 370–376. doi:10.1038/embor.2008.11

INTRODUCTION

Autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED) is a monogenic autosomal recessive syndrome characterized by a breakdown of self-tolerance, leading to autoimmune reactions in several organs and providing a useful model for molecular studies of autoimmunity (Mathis & Benoist, 2007). The disease is caused by mutations in autoimmune regulator (AIRE; Fig 1A), a transcriptional activator (Nagamine *et al*, 1997). AIRE promotes the thymic expression of many tissue-restricted antigens, enabling the negative selection of developing T cells and thus precluding self-reactivity (Anderson *et al*, 2002; Liston *et al*, 2003); however, the mechanisms are so far unknown. AIRE controls genes in genomic clusters, indicating a role in epigenetic regulation (Derbinski *et al*, 2005). Indeed, AIRE contains two plant homeodomain (PHD) fingers—small zinc-binding domains often found in chromatin-associated proteins (Aasland *et al*, 1995; Bienz, 2006). The PHD finger has emerged as a module that transduces histone-lysine methylation events. In particular, BPTF, ING2 and RAG2 PHD fingers recognize histone H3 trimethylated at lysine (K) 4 (H3K4me3; Li *et al*, 2006; Pena *et al*, 2006; Shi *et al*, 2006; Wysocka *et al*, 2006; Matthews *et al*, 2007), whereas the SMCX PHD finger binds to H3K9me3 (Iwase *et al*, 2007).

Here, we show that AIRE binds to histone H3 through its first PHD finger (AIRE–PHD1). In contrast with BPTF, ING2 and RAG2, AIRE–PHD1 preferentially binds to histone H3 non-methylated at lysine 4 (H3K4me0). Our results, in agreement with recent studies of the DNMT3L and BHC80 PHD fingers (Lan *et al*, 2007; Ooi *et al*, 2007), show a new role for the PHD finger as an H3K4me0 reader.

RESULTS AND DISCUSSION

AIRE–PHD1 binds to histone H3

To investigate the role of AIRE in chromatin-regulating complexes, we examined whether AIRE interacts with histones. Indeed, when incubated with whole histones, glutathione-*S*-transferase (GST)-AIRE (full-length) interacted with a histone that was identified as

[1]Molecular Pathology, University of Tartu, Tartu 50411, Estonia
[2]Biomolecular NMR Laboratory, Dulbecco Telethon Institute c/o S. Raffaele Scientific Institute, Milan 20132, Italy
[3]Institute of Chemical Physics, University of Tartu, Tartu 51010, Estonia
[4]Istituto di Ricerche di Biologia Molecolare, via Pontina km 30.600, Pomezia (Rome) 00040, Italy
[+]Corresponding author. Tel: +39 0226434824; Fax: +39 0226434153; E-mail: musco.giovanna@hsr.it
[++]Corresponding author. Tel: +372 7374 202; Fax: +372 7374 207; E-mail: part.peterson@ut.ee

Received 16 August 2007; revised 7 December 2007; accepted 9 January 2008; published online 22 February 2008

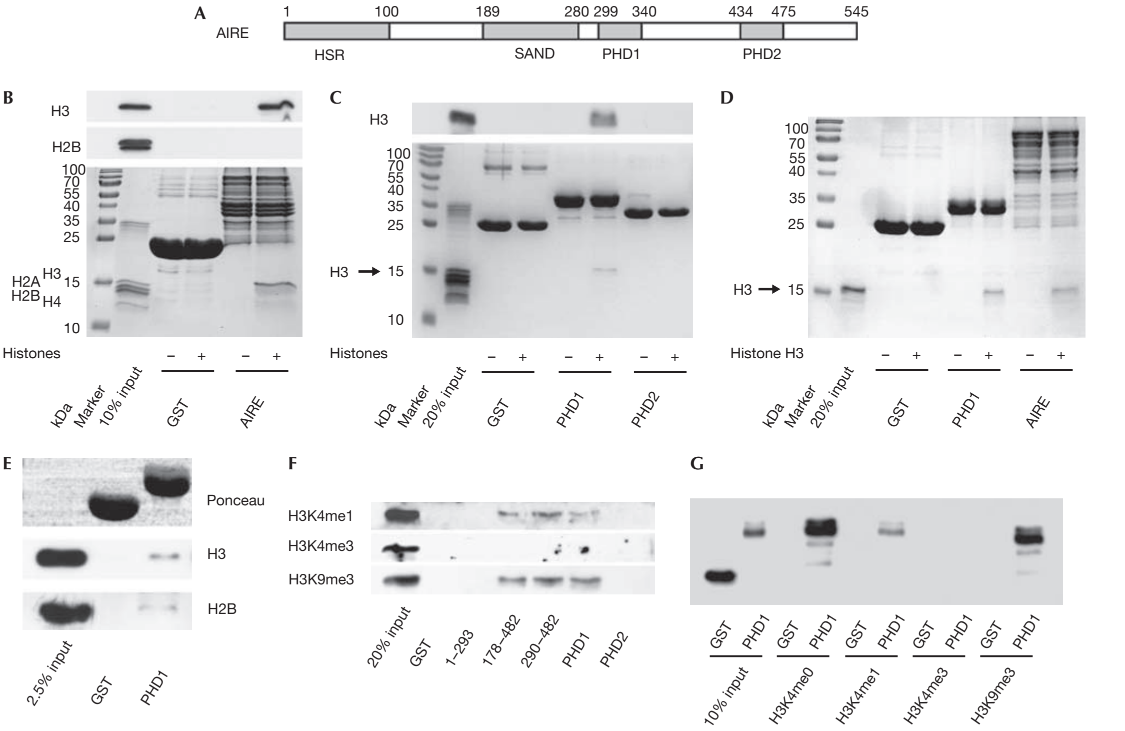


Fig 1 | AIRE interacts with H3K4me0 by means of AIRE–PHD1. (**A**) A schematic representation of the AIRE protein. Grey boxes represent HSR (homogenously staining region), PHD (plant homeodomain) and SAND (Sp100, AIRE-1, NucP41/P75 and *Drosophila* DEAF-1). (**B**) GST-AIRE interaction with whole histones visualized by Coomassie staining (bottom), and detected by western blot using anti-H3 and anti-H2B (top and middle). (**C**) A similar experiment to (**B**) but using GST-PHD proteins, detected by western blot using anti-H3 (top) or Coomassie staining (bottom). (**D**) GST-PHD1 and GST-AIRE, but not GST alone, interact with purified recombinant histone H3, visualized by Coomassie staining. (**E**) GST-PHD1, but not GST alone, interacts with native mononucleosomes detected by western blot against anti-H3 (middle) and anti-H2B (bottom). Equal input of GST proteins is shown with Ponceau red staining (top). (**F**) Interaction between GST-AIRE fusion proteins and whole histones, detected by anti-H3K4me1, anti-H3K4me3 and anti-H3K9me3. (**G**) Interaction between GST-AIRE–PHD1 fusion proteins and amino-terminal histone H3 peptides (H3K4me0, H3K4me1, H3K4me3 and H3K9me3), all detected by anti-GST. AIRE, autoimmune regulator; GST, glutathione-*S*-transferase.

H3 by western blotting (Fig 1B). By using various GST fusions, we found that AIRE–PHD1 is necessary and sufficient to interact with histone H3 (Fig 1C; supplementary Fig S1A,B online). Furthermore, the interaction is direct, as both full-length AIRE and AIRE–PHD1 bound to recombinant purified H3 (Fig 1D) but not to H2B (supplementary Fig S1C online). AIRE–PHD1 has a zinc-dependent fold and, accordingly, H3 binding is greatly reduced by EDTA or mutation of zinc-chelating cysteines, including the pathological mutation C311Y (Bjorses *et al*, 2000; supplementary Fig S1D–F online). AIRE–PHD1 also interacted with a small fraction of native mononucleosomes, as assessed by western blot against H3 and H2B (Fig 1E), and by analysing bound DNA (supplementary Fig S1G online). Thus, AIRE interacts, by means of its first PHD finger, specifically with histone H3 in both isolated and nucleosomal contexts.

AIRE–PHD1 preferentially binds to H3K4me0

Western blot analysis of H3/AIRE–PHD1 complex formation by using antibodies for H3K4me1, H3K4me3 and H3K9me3 indicated that H3K4 trimethylation hinders interaction (Fig 1F), whereas H3K9 trimethylation does not. Binding experiments with amino-terminal histone H3 unmodified (H3K4me0) or modified (H3K4me1, H3K4me3 and H3K9me3) 20-mer peptides showed that these N-terminal residues of histone H3 are sufficient for binding to AIRE–PHD1 (Fig 1G). Although both H3K4me0 and H3K9me3 peptides bound to AIRE–PHD1 with similar affinities, binding decreased with increasing H3K4 methylation (Fig 1G), indicating that AIRE–PHD1 preferentially binds to H3K4me0.

To confirm the specificity of AIRE–PHD1 for H3K4me0, we compared the binding of histone H3 N-terminal peptides—H3K4me0, H3K4me1, H3K4me2 and H3K4me3—to AIRE–PHD1 by using two dimensional ^{1}H-^{15}N nuclear magnetic resonance (NMR). A discrete set of chemical shift changes was observed on addition of all four histone H3 peptides to AIRE–PHD1 (supplementary Fig S2A,B online). However, the intensity of the changes was inversely related to the methylation level of the H3 peptide: the H3K4me0 peptide induced the largest changes (maximum average chemical shift change $\Delta\delta^{av}_{max}$= 0.9 p.p.m.; Fig 2).

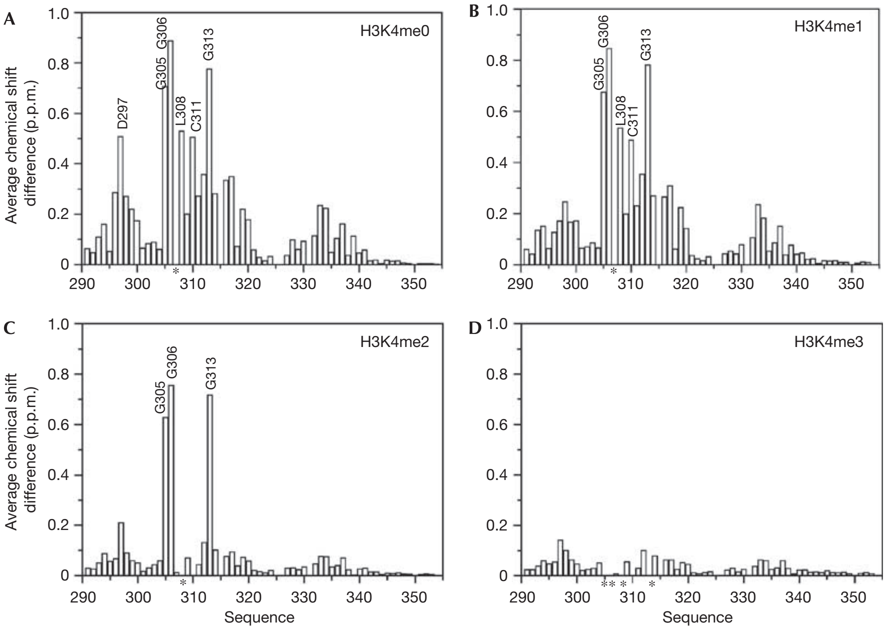


Fig 2 | Distribution of the backbone amide chemical shift changes within AIRE–PHD1 on binding to H3 amino-terminal peptides. Histograms showing the average backbone chemical shift differences observed in the ^{15}N-labelled AIRE–PHD1 (0.2 mM) on addition of a twofold excess of (**A**) H3K4me0, (**B**) H3K4me1, (**C**) H3K4me2 and (**D**) H3K4me3; the average chemical shift differences decrease with the methylation level of H3K4. Asterisks indicate residues for which the backbone amide signals disappear during titration owing to line broadening. AIRE, autoimmune regulator; H3K4me0, histone H3 non-methylated at lysine 4; H3K4me1 (H3K4me2, H3K4me3), histone H3 monomethylated (respectively, dimethylated and trimethylated) at lysine 4; PHD, plant homeodomain.

The addition of H3K4me0 and H3K4me1 peptides resulted in chemical shift changes in the slow- to intermediate-exchange regime (supplementary Fig S2A online), indicating low micromolar binding affinities. By contrast, the NMR data on addition of H3K4me2 and H3K4me3 peptides were in the fast-exchange regime, indicating millimolar binding affinities (supplementary Fig S2B online).

The greater binding affinity of AIRE–PHD1 for H3K4me0 peptides was confirmed by both tryptophan fluorescence spectroscopy and isothermal titration calorimetry (ITC), yielding dissociation constants of ~4 μM, ~20 μM and >0.5 mM for H3K4me0, H3K4me1 and H3K4me2, respectively (supplementary Fig S2C online; Table 1). Notably, H3K4me3 did not show any significant interaction with AIRE–PHD1 in either binding assay.

In agreement with the GST fusion pull-down experiments, fluorescence spectroscopy showed no binding of H3K4me0 to AIRE–PHD1 containing the APECED-causing C311Y mutation (Bjorses *et al*, 2000). Nevertheless, a second pathological mutant, V301M (Soderbergh *et al*, 2000), was still able to bind to H3K4me0, indicating that this mutation is not located in the H3 interaction site (Table 1).

The mapping of the H3/AIRE interaction site uniquely to AIRE–PHD1 was further confirmed by NMR titrations of histone H3 peptides into AIRE–PHD2, which bound neither methylated nor H3K4me0 peptides (data not shown).

Model of AIRE–PHD1 and histone H3 interactions

We generated a model of AIRE–PHD1 complexed with the H3K4me0 peptide on the basis of the crystal structure of the BPTF–PHD finger bound to H3K4me3 and performed molecular dynamics calculations for 10 ns. During the simulations, the peptide interacted stably with the first β-strand of AIRE–PHD1, creating a third antiparallel β-strand (Fig 3). The additional β-strand allowed the formation of four hydrogen (H) bonds from the backbone of H3 residues R2, K4 and T6 to the AIRE–PHD1 residues C310, L308 and G306, respectively (Fig 3B). Accordingly, the amides of C310, L308 and G306 showed high protection factors in NMR deuterium exchange experiments, confirming their involvement in H-bonds (Fig 3B). The N terminus of the peptide was anchored through intermolecular H-bonds with the backbone carbonyl oxygen atoms of residues P331–G333 (Fig 3B). Furthermore, hydrophobic interactions between the methyl group of A1 and the pyrrolidine ring of P331, and between the methylene groups of K4 and L308 further contributed to the stabilization of the complex. The formation of salt bridges between the side chains of R2 and D312, and between K4 and D297 seemed to be crucial for binding specificity, as indicated experimentally by the large NMR chemical shift changes for G313 (near to D312) and D297 (Fig 2). Indeed, fluorescence spectroscopy and ITC assays showed that the alanine mutations R2A in the H3 peptide and D312A in AIRE–PHD1 markedly reduced the

Table 1 | Values of the dissociation constants between H3 peptides and AIRE–PHD1 wild type (WT) and mutants measured by fluorescence spectroscopy and isothermal titration calorimetry

AIRE–PHD1	Peptide	K_D (μM), fluorescence spectroscopy	K_D (μM), isothermal titration calorimetry
WT	H3K4me0	4.7 ± 0.8	6.5 ± 0.2
WT	H3K4me1	21.4 ± 5.9	55.6 ± 1.2
WT	H3K4me2	> 500	714 ± 90
WT	H3K4me3	ND	NM
V301M	H3K4me0	6.8 ± 0.4	NM
C311Y	H3K4me0	ND	NM
D297A	H3K4me0	173.0 ± 18.6	146.8 ± 6.1
D312A	H3K4me0	ND	ND
WT	H3K4me0-R2A	ND	NM
D297A	H3K4me3	ND	NM

ND, not detectable, denotes binding too weak to be reliably quantified; NM, not measured.

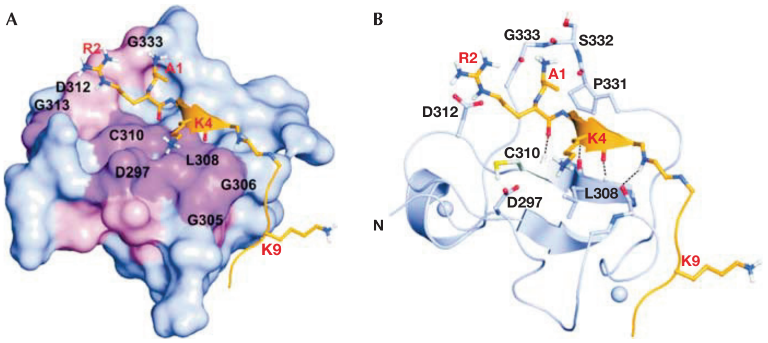


Fig 3 | Model of AIRE–PHD1 in complex with H3K4me0. (**A**) Surface representation of AIRE–PHD1 in complex with H3K4me0. Residues with the highest chemical shifts are shown in magenta ($\Delta\delta > 0.4$ p.p.m.) and pink ($0.2 < \Delta\delta < 0.4$ p.p.m.). (**B**) Ribbon representation of a representative structure of the complex of AIRE1–PHD1 (blue) with H3K4me0 (orange). Inter-backbone hydrogen bonds and Zn^{2+} ions are represented by dotted lines and spheres, respectively. AIRE, autoimmune regulator; H3K4me0, histone H3 non-methylated at lysine 4; PHD, plant homeodomain.

binding affinity (Table 1; Fig 4C) without affecting the protein fold (supplementary Fig S3 online). Similarly, pull-down experiments with whole histones and the H3K4me0 peptide, together with fluorescence spectroscopy and ITC measurements performed on AIRE–PHD1-D297A showed reduced binding (Table 1; Fig 4). Furthermore, no binding was observed in fluorescence spectroscopy and ITC experiments when H3K4me3 was titrated into AIRE–PHD1-D297A (Table 1). Importantly, simulations of AIRE–PHD1 with H3K4me1 or H3K4me3 were not compatible with complex formation, showing displacement of K4 owing to steric clashes with D297, with the consequent breakage of the additional β-strand (supplementary Fig S4 online).

Nature of the binding interface

The model of AIRE–PHD1 complexed with H3K4me0 was in perfect agreement with the experimental chemical shift perturbation data, as the peptide-binding region coincided with the binding surface identified by NMR spectroscopy (Fig 3A). In fact, the H3K4me0 peptide induced chemical shift changes in AIRE–PHD1 residues that map only on one side of the protein surface, involving residues in the N terminus of the PHD finger, the first β-strand, and the loop connecting the first and the second β-strands (D297, G305, G306, L308, C310, D312 and G313; Fig 2; supplementary Fig S5 online). A similar pattern of chemical shift changes indicated the same binding site for H3K4me1. However, H3K4me1 induced smaller changes for residues E296–A300, indicating that binding to this region is reduced by K4 methylation (Fig 2B).

H3K4me2 and H3K4me3 also induced similar patterns of chemical shift changes, indicating a similar interaction site with AIRE–PHD1. However, the changes were markedly reduced in size, in keeping with a weak interaction (Fig 2C,D). Remarkably, residues G305, G306 and G313 showed strong shifts when bound to H3K4me2 and disappeared completely from the NMR spectrum owing to line-shape broadening on binding to H3K4me3, indicating an involvement of this region in peptide binding.

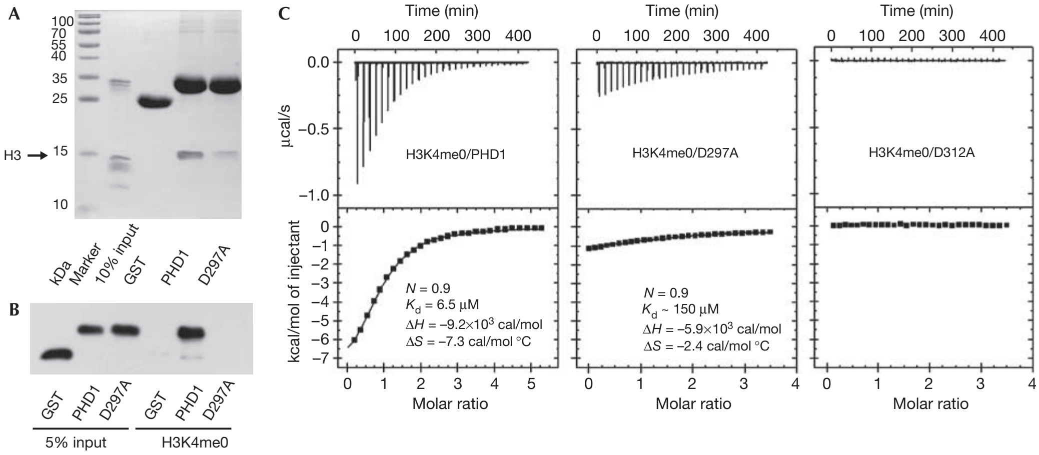


Fig 4 | Mutations of D297 and D312 abolish AIRE–PHD1 binding to histone H3. (**A**) Pull-down assay of AIRE–PHD1 (PHD1) and AIRE–PHD1-D297A (D297A) mutant proteins with histones. (**B**) Interaction between AIRE–PHD1 (PHD1), AIRE–PHD1-D297A (D297A) mutant proteins and H3K4me0 peptide detected by anti-GST. (**C**) ITC data for binding of H3K4me0 peptide to AIRE–PHD1 (PHD1), AIRE–PHD1-D297A (D297A) and AIRE–PHD1-D312A (D312A). The upper panels show the sequential heat pulses for peptide–protein binding, and the lower panels show the integrated data, corrected for heat of dilution and fit to a single-site-binding model using a nonlinear least-squares method (line). *N*, K_d, Δ*H* and Δ*S* represent measured stoichiometric ratio, dissociation binding constant, differential enthalpy and differential entropy, respectively. AIRE, autoimmune regulator; GST, glutathione-S-transferase; H3K4me0, histone H3 non-methylated at lysine 4; ITC, isothermal titration calorimetry; PHD, plant homeodomain.

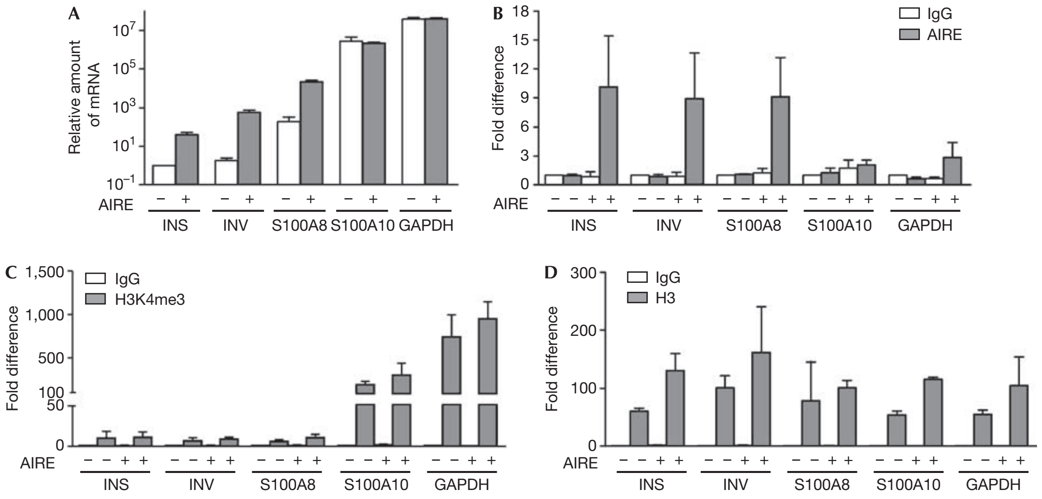


Fig 5 | AIRE binds to chromatin as a transcriptional activator. (**A**) Relative expression levels of insulin (INS), involucrin (INV), S100A8, S100A10 and GAPDH genes in HEK-AIRE (+) and HEK-control (−) cell lines are shown in logarithmic scale and in comparison with the insulin messenger RNA level in HEK-control cell line (=1). DNA ChIP with (**B**) anti-AIRE, (**C**) anti-H3K4me3 and (**D**) anti-H3. The fold differences are normalized to input fractions and shown in comparison with the background level (ChIP with IgG from HEK-control cells (=1)) of each primer set. The data are the averages of two or more independent experiments. AIRE, autoimmune regulator; ChIP, chromatin immunoprecipitation; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; HEK, human embryonic kidney; H3K4me3, histone H3 trimethylated at lysine 4.

Structural comparison with other PHD fingers

Our data suggest a regulatory mechanism mediated by AIRE–PHD1 that differs from that of ING2 and BTPF, the PHD fingers of which bind to H3K4me3 and discriminate against H3K4me0. A structural based sequence alignment (supplementary Fig S6 online) suggests that AIRE–PHD1 is representative of a newly identified subclass of PHD fingers (Lan *et al*, 2007). AIRE–PHD1 differs structurally from the ING2 and BPTF PHD fingers owing to the lack of conserved aromatic residues used to coordinate the trimethyl ammonium ion of H3K4me3 by π-cation interactions. Instead, the crucial elements of the methylated lysine-binding aromatic cage seen in ING2 and BPTF (supplementary Fig S6 online) are substituted by negatively charged (D297) and small hydrophobic (A317) residues in AIRE–PHD1. Our data show that D297 is involved in the interaction of AIRE with H3K4me0, providing an alternative to the recognition of histone H3 by aromatic caging. Notably, D297 is conserved in other PHD finger proteins, for example, Sp110 and Sp140, which might constitute a subset of H3K4me0 readers (supplementary Fig S6 online). Recently, the PHD finger of BHC80 and the cysteine-rich domain of DNMT3L were shown to recognize H3K4me0 by an analogous mechanism, in which the H3 peptide binds to the surface of the domain, forming an additional β-strand that is anchored by the side chain and N-amine group of H3A1. Importantly, these proteins also have an acidic residue comparable to D297, which forms a salt bridge with K4. Although there are many similarities between these two structures and the AIRE–PHD1/H3K4me0 complex presented here, the AIRE–PHD1 finger differs in the additional recognition of the H3R2 side chain, which makes an important contribution to the high affinity of this interaction, as shown by our peptide mutagenesis experiments.

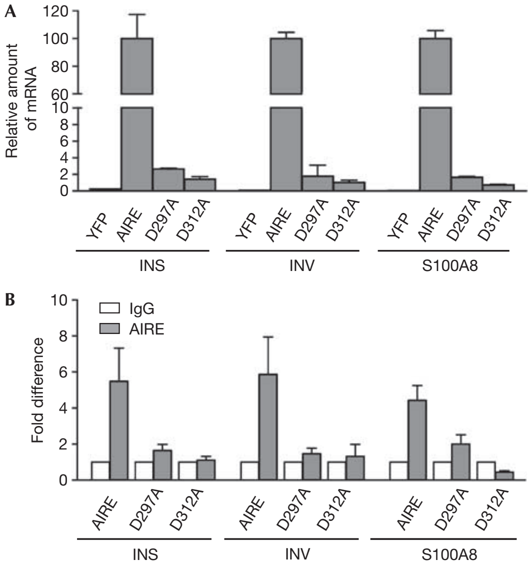


Fig 6 | Influence of the AIRE–PHD1 mutations on transcriptional activation and chromatin binding. (**A**) Relative expression levels of insulin (INS), involucrin (INV) and S100A8 genes in HEK-AIRE (AIRE), HEK-AIRE-D297A (D297A) and HEK-AIRE-D312A (D312A) mutant and HEK-control (YFP) cell lines are shown in comparison with the messenger RNA levels in HEK-AIRE cell lines (= 100%). (**B**) DNA ChIP analysis with anti-AIRE and IgG was performed from the stably transfected cells, as indicated. The fold differences are normalized to input fractions and shown in comparison with the background level (= 1) with each primer set in each condition. The data are the averages of two independent experiments. AIRE, autoimmune regulator; ChIP, chromatin immunoprecipitation; HEK, human embryonic kidney; PHD, plant homeodomain.

AIRE interacts with chromatin

We have shown previously that transiently transfected AIRE enhances target gene expression in human embryonic kidney (HEK)293 cells (Pitkanen *et al*, 2005). So far, no cell line has been described with endogenous AIRE expression; therefore, we transfected HEK293 cells with an AIRE-encoding or control plasmid and generated stable cell lines called HEK-AIRE and HEK-control. We first tested HEK-AIRE compared with HEK-control cell lines for expression levels of tissue-restricted antigens that are downregulated in AIRE-deficient mouse thymic medullary epithelial cells (Derbinski *et al*, 2005). Indeed, the HEK-AIRE cell line showed enhanced expression of such antigens, including insulin, the principal autoantigen in type I diabetes (Babaya *et al*, 2005), involucrin and S100A8 (Fig 5A). The last two genes are AIRE target genes clustered on human chromosome 1q21 (Marenholz *et al*, 2001). Conversely, the expression levels of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and that of another S100 family protein, S100A10, were unaffected by AIRE (Fig 5A). Next, we studied *in vivo* histone binding by protein chromatin immunoprecipitation (ChIP) assays and observed that AIRE is found in complexes with a small fraction of histone H3 but not with H3K4me3. By contrast, binding of ING2, used as a positive control, was detected for both H3 and H3K4me3 (supplementary Fig S7 online). By using DNA ChIP analysis, we found that AIRE interacts with the insulin, involucrin and S100A8 promoter regions, but much less with the S100A10 and GAPDH promoters (Fig 5B). In agreement with the low expression levels observed, the insulin, involucrin and S100A8 promoters almost completely lacked H3K4me3, whereas the highly expressed S100A10 and GAPDH promoters were enriched with H3K4me3 (Fig 5C). The overall levels of histone H3 were comparable on all promoters studied (Fig 5D). To analyse the influence of AIRE–PHD1 mutations that impaired the interaction with H3 *in vitro*, we generated stable cell lines expressing AIRE with PHD1 mutations D297A and D312A. Importantly, the activation of the AIRE target genes (Fig 6A), as well as binding to their promoters (Fig 6B), was clearly reduced by both mutations. Although AIRE specificity towards chromatin might be influenced by other protein and DNA interactions (Ruan *et al*, 2007), the data presented here indicate that AIRE preferentially binds to and activates the promoters containing low levels of H3K4me3. On the basis of these results, we propose a speculative model for the regulation of tissue-restricted antigen expression in thymic epithelial cells (supplementary Fig S8 online). Normally, tissue-restricted antigens are silenced in immature thymic epithelial cells as they lack the active chromatin mark H3K4me3 on their promoters. During differentiation into mature thymic medullary epithelial cells, activation of AIRE expression (Kyewski & Klein, 2006) enables the read-out of non-methylated H3K4 as a signal to activate tissue-restricted antigen genes. AIRE binding to the non-methylated H3K4 on tissue-restricted antigen promoters results in recruitment of other transcriptional regulators, for example, CBP (Pitkanen *et al*, 2005) and activation of transcription.

Our results provide new information on the role of AIRE in sensing epigenetic chromatin modifications through direct binding of AIRE–PHD1 to histone H3 N-terminal residues. Collectively, our data show that AIRE belongs to a new subset of PHD finger-containing proteins that preferentially recognize H3K4me0. Future studies should therefore explore the epigenetic role of AIRE in thymic expression of tissue-restricted antigens to advance further our understanding of this important regulator of autoimmunity.

METHODS

Plasmid construction and *in vitro* binding assays. The construction of plasmids, information on antibodies and peptides used, as well as protein expression and binding assays are described in the supplementary information online.

NMR binding, fluorescence titration assays and isothermal titration calorimetry thermodynamic analysis. Details on NMR titrations, fluorescence spectroscopy and thermodynamic measurements are described in the supplementary information online.

Assembly of the complex structures and molecular dynamics calculations. The PHD finger structures from the human NURF BPTF PHD finger-H3K4me3 complex (2fuu) and AIRE1–PHD1 (1xwh) were superimposed by using the Lsqman program (Cα atom RMSD: 2.1 Å). Molecular dynamics simulations and analysis were performed using the GROMACS 3.1.3 package with GROMOS force field. The details of the protocol are available in the supplementary information online.

Cell lines, expression analysis and chromatin immunoprecipitation. The establishment of HEK-AIRE and HEK-control cell lines is described in the supplementary information online. DNA ChIP was performed essentially according to Upstate Chromatin Immunoprecipitation Assay protocol. Quantitative PCR analysis and primer sequences are provided in the supplementary information online.

Supplementary information is available at *EMBO reports* online (http://www.emboreports.org).

ACKNOWLEDGEMENTS
T.O., A.R., I.L. and P.P. were supported by grants from the Wellcome Trust and European Union Framework programme 6 (Thymaide and Euraps), and from the Estonian Science Foundation (6663 and 6490). We acknowledge A. Häling for technical help. G.M. acknowledges support from Fondazione Telethon, Fondazione Cariplo and Compagnia S. Paolo, and thanks M. Bianchi and D. Gabellini for useful discussions. C.H. and U.M. thank the Foundation Innove (www.innove.ee) project no. 1.0101-0310 for financial support.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

REFERENCES

Aasland R, Gibson TJ, Stewart AF (1995) The PHD finger: implications for chromatin-mediated transcriptional regulation. *Trends Biochem Sci* **20:** 56–59

Anderson MS *et al* (2002) Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298:** 1395–1401

Babaya N, Nakayama M, Eisenbarth GS (2005) The stages of type 1A diabetes. *Ann NY Acad Sci* **1051:** 194–204

Bienz M (2006) The PHD finger, a nuclear protein-interaction domain. *Trends Biochem Sci* **31:** 35–40

Bjorses P, Halonen M, Palvimo JJ, Kolmer M, Aaltonen J, Ellonen P, Perheentupa J, Ulmanen I, Peltonen L (2000) Mutations in the *AIRE* gene: effects on subcellular location and transactivation function of the autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy protein. *Am J Hum Genet* **66:** 378–392

Derbinski J, Gabler J, Brors B, Tierling S, Jonnakuty S, Hergenhahn M, Peltonen L, Walter J, Kyewski B (2005) Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J Exp Med* **202:** 33–45

Iwase S, Lan F, Bayliss P, de la Torre-Ubieta L, Huarte M, Heng Qi H, Whetstine JR, Bonni A, Roberts TM, Shi Y (2007) The X-linked mental retardation gene *SMCX/JARID1C* defines a family of histone H3 lysine 4 demethylases. *Cell* **128:** 1077–1088

Kyewski B, Klein L (2006) A central role for central tolerance. *Annu Rev Immunol* **24:** 571–606

Lan F, Collins RE, De Cegli R, Alpatov R, Horton JR, Shi X, Gozani O, Cheng X, Shi Y (2007) Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **448:** 718–722

Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442:** 91–95

Liston A, Lesage S, Wilson J, Peltonen L, Goodnow CC (2003) Aire regulates negative selection of organ-specific T cells. *Nat Immunol* **4:** 350–354

Marenholz I, Zirra M, Fischer DF, Backendorf C, Ziegler A, Mischke D (2001) Identification of human epidermal differentiation complex (EDC)-encoded genes by subtractive hybridization of entire YACs to a gridded keratinocyte cDNA library. *Genome Res* **11:** 341–355

Mathis D, Benoist C (2007) A decade of AIRE. *Nat Rev Immunol* **7:** 645–650

Matthews AG *et al* (2007) RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature* **450:** 1106–1110

Nagamine K *et al* (1997) Positional cloning of the APECED gene. *Nat Genet* **17:** 393–398

Ooi SK *et al* (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* **448:** 714–717

Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV, Gozani O, Zhao R, Kutateladze TG (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442:** 100–103

Pitkanen J *et al* (2005) Cooperative activation of transcription by autoimmune regulator AIRE and CBP. *Biochem Biophys Res Commun* **333:** 944–953

Ruan QG *et al* (2007) The autoimmune regulator directly controls the expression of genes critical for thymic epithelial function. *J Immunol* **178:** 7173–7180

Shi X *et al* (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442:** 96–99

Soderbergh A, Rorsman F, Halonen M, Ekwall O, Bjorses P, Kampe O, Husebye ES (2000) Autoantibodies against aromatic L-amino acid decarboxylase identifies a subgroup of patients with Addison's disease. *J Clin Endocrinol Metab* **85:** 460–463

Wysocka J *et al* (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442:** 86–90

hetenyi.csaba_83_23

D10

International Journal of
Molecular Sciences

MDPI

Article

The Structural Effects of Phosphorylation of Protein Arginine Methyltransferase 5 on Its Binding to Histone H4

Rita Börzsei [1,2], Bayartsetseg Bayarsaikhan [1], Balázs Zoltán Zsidó [1,2], Beáta Lontay [3] and Csaba Hetényi [1,2,*]

[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, 7624 Pécs, Hungary
[2] János Szentágothai Research Centre & Centre for Neuroscience, University of Pécs, 7624 Pécs, Hungary
[3] Department of Medical Chemistry, Faculty of Medicine, University of Debrecen, 4032 Debrecen, Hungary
[*] Correspondence: hetenyi.csaba@pte.hu

Abstract: The protein arginine methyltransferase 5 (PRMT5) enzyme is responsible for arginine methylation on various proteins, including histone H4. PRMT5 is a promising drug target, playing a role in the pathomechanism of several diseases, especially in the progression of certain types of cancer. It was recently proved that the phosphorylation of PRMT5 on T80 residue increases its methyltransferase activity; furthermore, elevated levels of the enzyme were measured in the case of human hepatocellular carcinoma and other types of tumours. In this study, we constructed the complexes of the unmodified human PRMT5-methylosome protein 50 (MEP50) structure and its T80-phosphorylated variant in complex with the full-length histone H4 peptide. The full-length histone H4 was built in situ into the human PRMT5-MEP50 enzyme using experimental H4 fragments. Extensive molecular dynamic simulations and structure and energy analyses were performed for the complexed and apo protein partners, as well. Our results provided an atomic level explanation for two important experimental findings: (1) the increased methyltransferase activity of the phosphorylated PRMT5 when compared to the unmodified type; (2) the PRMT5 methylates only the free form of histone H4 not bound in the nucleosome. The atomic level complex structure H4-PRMT5-MEP50 will help the design of new inhibitors and in uncovering further structure–function relationships of PRMT enzymes.

Keywords: ligand; epigenetics; post-translational modification; signal transduction

check for updates

Citation: Börzsei, R.; Bayarsaikhan, B.; Zsidó, B.Z.; Lontay, B.; Hetényi, C. The Structural Effects of Phosphorylation of Protein Arginine Methyltransferase 5 on Its Binding to Histone H4. *Int. J. Mol. Sci.* **2022**, *23*, 11316. https://doi.org/10.3390/ijms231911316

Academic Editor: Christina Piperi

Received: 23 August 2022
Accepted: 21 September 2022
Published: 26 September 2022

1. Introduction

Post-translational modification (PTM) is a fundamental mechanism occurring on proteins of different roles in epigenetic regulation [1–4]. Histone H4 is a building block of the nucleosome, the smallest unit of the chromosome [5]. It also contributes to the epigenetic "histone code" system [6], a combination of PTMs mostly on the N-terminal tail of H4 and other histones. PTMs often involve the covalent attachment of atomic groups to proteins catalysed by different enzymes, also called writers [7].

Protein arginine methyltransferases (PRMT) are writers that add methyl groups to the N-terminal arginine residues of H4 (or other substrates) [8]. It was recently recognised that arginine methylation via PRMTs is associated with several diseases, especially cancer progression [8,9]. Consequently, PRMTs are promising novel drug targets in tumour therapy, as is indicated by the numerous PRMT inhibitors that appeared in preclinical and clinical development [8].

PRMT5 is a member of the PRMT family, catalysing the arginine monomethylation and monomethylation of several non-histone and histone proteins, including histone H2A [10–12], H3 [13,14], and H4 [8,11,15,16]. Its activity is linked to mRNA splicing, DNA repair mechanisms, drug resistance, and the regulation of immune cell function [8]. An increased activity and overexpression of PRMT5 was identified in several cancers, making it a promising drug target [8]. PRMT5 is localised in both the cytoplasm and nucleus, in complex with methylosome protein 50 (MEP50), creating an association with numerous partner proteins and several histone and non-histone ligands [15]. However, it was experimentally proved that, likewise to other PRMTs, PRMT5 cannot catalyse the arginine methylation of histones if bound in the nucleosome [11,16].

The phosphorylation and dephosphorylation of tyrosine [17,18] and threonine [19] residues of PRMT5 have an effect on its enzyme activity, and therefore, on the pathomechanism of tumour formation. For example, in the case of hepatocellular carcinoma, the phosphorylation/dephosphorylation of PRMT5 on T80 modulates its methyltransferase activity, and the dephosphorylating myosine phosphatase has a tumour suppressor role [9]. Due to the role of PRMT5 in tumourigenesis, the regulation of its enzymatic activity is the major point of interest. It can be regulated at the molecular level, primarily by the formation of the methylosome complex, containing PRMT5 and its various partners, such as MEP50 [17]. However, the major regulatory action on PRMT5 was described by Rho A activating kinase (ROK) and myosin phosphatase (MP), which also counteract on the T80 phosphorylation site of PRMT5, regulating its methyltransferase activity, both in vitro and in vivo. MP modulates the symmetrical dimethylation of histone core proteins in the cell nucleus via the dephosphorylation of PRMT5 at its activating phosphorylation site, causing changes in gene expression. In tumour cells, the inhibitory phosphorylation of MP is increased, leading to higher phosphorylation levels of PRMT5 at T80 by ROK [9].

The experimental atomic resolution structure of human PRMT5 in complex with MEP50, a methyl-donor ligand and an eight-amino-acid-long histone H4 fragment, was revealed 10 years ago [17]. Some years earlier, structures of non-human PRMT5 were published, as well [20,21]. However, complex structures with the full-length histone H4 have not been published yet.

The aim of this study was to construct the human PRMT5-MEP50 structure in complex with the full-length histone H4 peptide in order to provide a structural explanation for the increased methyltransferase activity of the T80-phosphorylated enzyme (PRMT5$_P$), as well as for the inactivity of PRMT5 on nucleosome-bound histone H4.

2. Results and Discussion

2.1. Unmodified PRMT5 in Complex with the Full-Length H4 Protein

The explanation of the difference in enzymatic activity (activation energy) of the two PRMT5 variants necessitates the atomic resolution structures of the H4-PRMT5 complexes. However, PRMT5(-MEP50) in complex with the full-length histone H4 has not been measured yet (see Table S1 for available PRMT5 structures [17,20–34]). Although the number of experimental human PRMT5 complexes increased recently due to its importance in cancer therapy, only one structure (PDB code: 4gqb, [17]) contains an eight-amino-acid-long N-terminal peptide fragment of histone H4 bound to the catalytic domain of PRMT5. The experimental determination of a full-length histone structure may be challenging, partly due to the high flexibility of the N-terminal tail [4]. However, the catalytic domain is positioned far from T80 in space, and therefore, the PRMT5-bound structure of N-terminal tail of histone H4 alone did not provide a sufficient basis for an explanation of the effects of T80 phosphorylation. Thus, the building of the full-length H4 in complex with PRMT5 was necessary to provide an explanation of the effects of T80 phosphorylation. Building the unmodified complex H4-PRMT5(-MEP50) was challenging, as there is no information in the literature about how the full-length H4 fits to PRMT5. The structure of the full-length histone H4 (Figure 1A) is available, e.g., in a nucleosome-bound form (PDB code: 1kx5, [35]). However, the simple superimposition of this full-length nucleosomal H4 to the N-terminal H4 fragment (4gqb) did not result in a collision-free H4-PRMT5 complex. Thus, an in situ, fragment-based construction of the full-length histone H4 structure was performed, using available histone H4 fragment structures, starting from the 8th amino acid of H4 in 4gqb (Methods, Figure 1B). The superimposed H4 fragments were covalently attached, and the H4-PRMT5(-MEP50) complex was energy-minimized and submitted to a 720 ns-long molecular dynamic (MD) simulation.

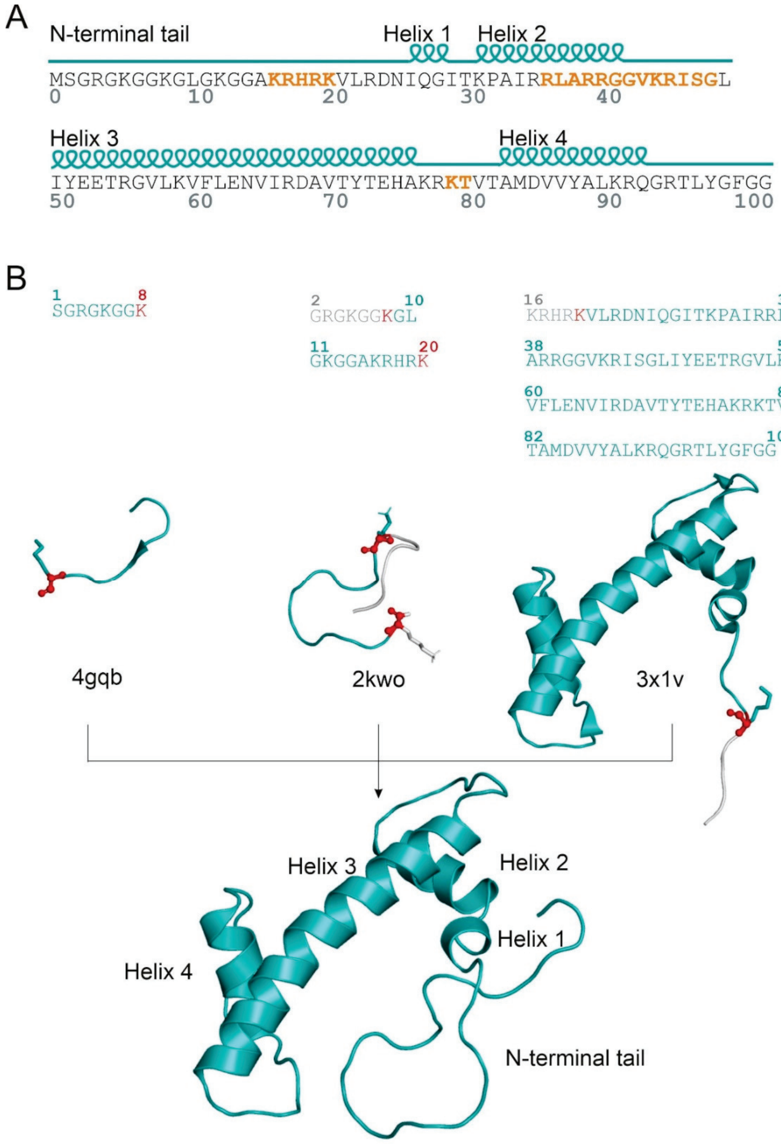


Figure 1. (**A**) Sequence and secondary structure of histone H4 (Uniprot code: P62805). DNA binding regions are highlighted with orange; (**B**) the process of in situ, fragment-based construction of the full-length histone H4 (down, cartoon, teal) by the usage of peptide fragments of different lengths obtained from experimental structures (PDB codes: 4gqb, 2kwo, and 3x1v). Residues (sticks, teal) and backbone atoms (N, Cα, C, O, and spheres) used for alignment are highlighted with red, while the overlapped regions (grey) were cut.

The schematic energy profile of the human PRMT5-catalysed methylation of H4 is shown in Figure 2A. In the case of the T80 phosphorylation of the enzyme (PRMT5$_{\mathrm{P}}$), it can be expected (Introduction) that the activation energy barrier will decrease (Figure 2A) when compared to the unmodified enzyme (PRMT5).

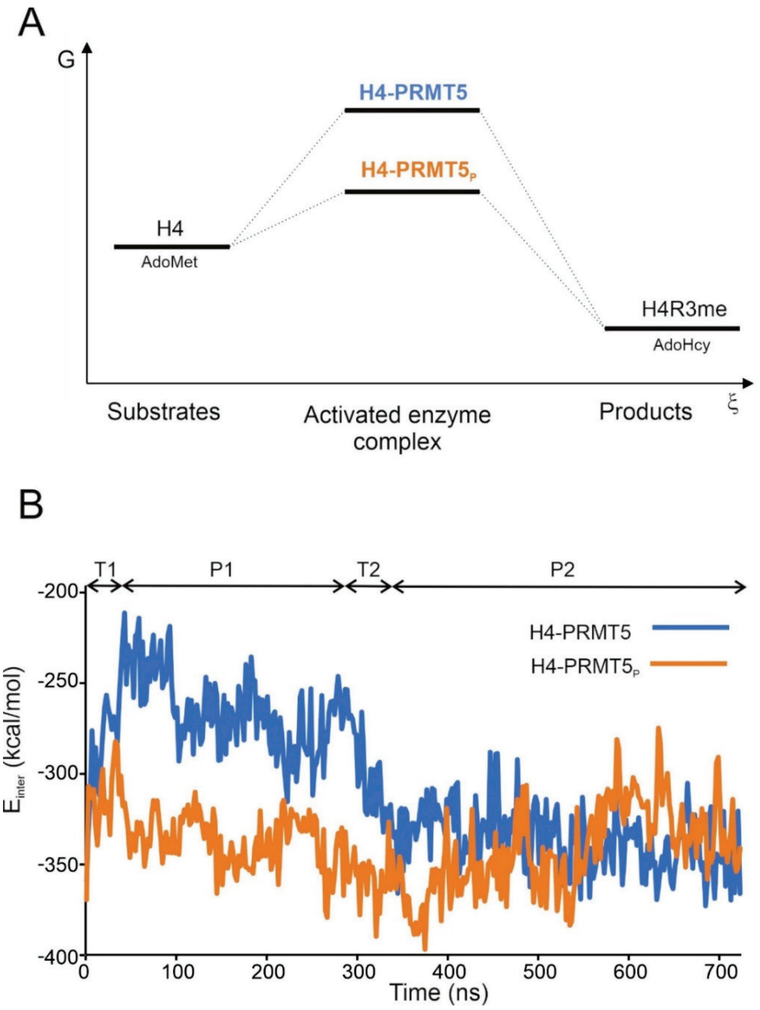


Figure 2. (**A**) Schematic free energy (G) vs. reaction coordinate (ξ) profile of human PRMT5-catalysed methylation of histone H4. Substrate histone H4 peptide and the methyl donor S-adenosyl-L-methionine (AdoMet) are marked. The free energy of the activated enzyme complexes is relatively high. However, the phosphorylated enzyme complex (H4-PRMT5$_P$, orange) has to cross a lower energy barrier than the unmodified complex (H4-PRMT5, blue). Products R3-methylated histone H4 (H4R3me) and S-adenosyl-L-homocysteine (AdoHcy) are also shown; (**B**) intermolecular interaction energy (E$_{inter}$) changed during the 720 ns-long MD simulation for the unmodified (H4-PRMT5, blue) and the T80-phosphorylated (H4-PRMT5$_P$, orange) enzyme complexes. The plot includes two transient (T1, T2) and two plateau (P1, P2) regions. T1 is an equilibration section for the conformational optimalization, while T2 refers to the energy drop between the activated (P1) and the pre-product (P2) states of the enzyme–substrate complexes.

The interaction energy (E$_{inter}$, Methods) between H4 and PRMT5(-MEP50) was calculated for all snapshots of the MD simulation (Figure 2B). Two transient (T1, T2) and two horizontal (plateau) (P1, P2) regions can be distinguished (Figure 2B). E$_{inter}$ showed a relatively quick change up to 120 kcal/mol in the transient regions T1 and T2, and it fluctuated with a maximal amplitude of 50 kcal/mol in the P1 and P2 regions for a longer time period of at least 250 ns (Table S2). T1 can be assigned as a technical equilibrating region of conformational optimization of the complex, and therefore, it was omitted from the evaluations. At the same time, T2 may correspond to the energy drop connecting the activated and the pre-product states of the enzyme–substrate complex (H4-PRMT5, Figure 2A). The term "pre-product state" is used for a bound H4 conformation appropriately prepared for subsequent methylation, but not yet methylated.

Representative structures of H4-PRMT5 were selected (Method) from both P1 and P2 plateaus and analysed. The interactions between the catalytic domain of PRMT5 and histone H4 N-terminal residues remained stable throughout the MD simulation. At the same time, a considerable change of the H4 structure was observed between the activated (P1) and pre-product (P2) states, also reflected by the aforementioned drop in the total E_{inter} (Figure 2B). In the pre-product structure of H4-PRMT5 (plateau P2), histone H4 interacted with six major regions of PRMT5, including the catalytic sites (e.g., residues E435 and E444), H146, R201, Y304-Q309, D317-Q322, and E483-D491, reflected by the favourable (large negative) E_{inter} contributions of the above regions, listed as bar charts calculated for a representative structure of the P2 complex, and marked with coloured spheres in Figure 3 (Methods). Although Helix 3 of H4 spans over PRMT5:T80, significant interaction could not be measured at T80 (Figure 3). Instead, the C-terminal part of Helix 3 (R67 and D68) showed a remarkable interaction (Figure 3) with the neighbouring PRMT5 residues (marked with blue and magenta in Figure 3). In contrast with the hypothesis of a previous study [16], MEP50 does not play a direct role in histone binding (S4).

2.2. The H4-PRMT5-MEP Complex vs. Apo Protein Structures

To examine the conformational changes of histone H4 and PRMT5-MEP50 during complex formation (Section 1), their structures were extracted from the complex and submitted to MD simulations of 1000 and 580 ns, respectively. The root mean square fluctuation (RMSF) of each residue was calculated, and regions with an RMSF higher than 3 Å were collected (Figure 4). Interestingly, two of the four PRMT5 regions (residues 145–148 and 490–491) with an RMSF higher than 3 Å took place in histone H4 binding (Figure 4A). At the same time, residues in these regions had the highest E_{inter} in the pre-product H4-PRMT5 complex structure (Figure 3). Unlike PRMT5, MEP50 showed conformational rigidity (RMSF < 3 Å, Figure 4B), indicating the lack of flexible regions necessary for H4 binding.

In the case of histone H4, the highest RMSF occurred at the linear N-terminal tail (residues 1–33, Figure 4C), including residue R3, methylated by PRMT5. While this region is obviously highly flexible, it is crucial in PRMT5 binding, also indicated by the per-residue E_{inter} contribution (Figure 3). Similarly, the region of residues 39–53 also showed high fluctuation and were involved in the binding of PRMT5 phosphorylated on T80 ($PRMT5_P$, see Section 3 for details), but did not show significant interaction with PRMT5 (Figure 3) The third, small region with RMSF > 3 Å was focused on K59, important (Figure 3) in stabilizing the PRMT5 complex structure (Figure 4C).

The calculation of the root mean squared deviation (RMSD) was also performed to estimate the time scale of conformational changes of H4 binding to PRMT5. Considerable changes of the bound structures were measured in terms of Cα RMSD values (Figure S1) of 7.7 (P1) and 9.1 (P2) Å when compared to the representative apo form (last frame). This considerable change in the H4 structure can be attributed to Helix 3, which had a linear conformation in the nucleosome (Figure 5) that is very similar to the representative apo conformation (Figure 5). At the same time, in the pre-product state of the H4-PRMT5 complex (P2), Helix 3 broke in the middle and adopted a V-shaped conformation (Figure 5). This conformational change was further analysed, and the Cα RMSD of Helix H3 was calculated for the unbound H4 MD trajectory, using the pre-product H4 V-shaped conformation (P2) as a reference structure (Figure S2). Helix 3 of unbound H4 showed flexibility centred at H4:G56, resulting in V-shaped conformations (Figure S2) similar to the complexed H4 (Figure 5). The unbound H4 MD simulation showed that a ca. 400 of 1000 ns (Figures S1 and S2) time frame is necessary for the conformational change of Helix 3 from the V-shaped to the linear form of the representative apo H4 structure (Figure 5).

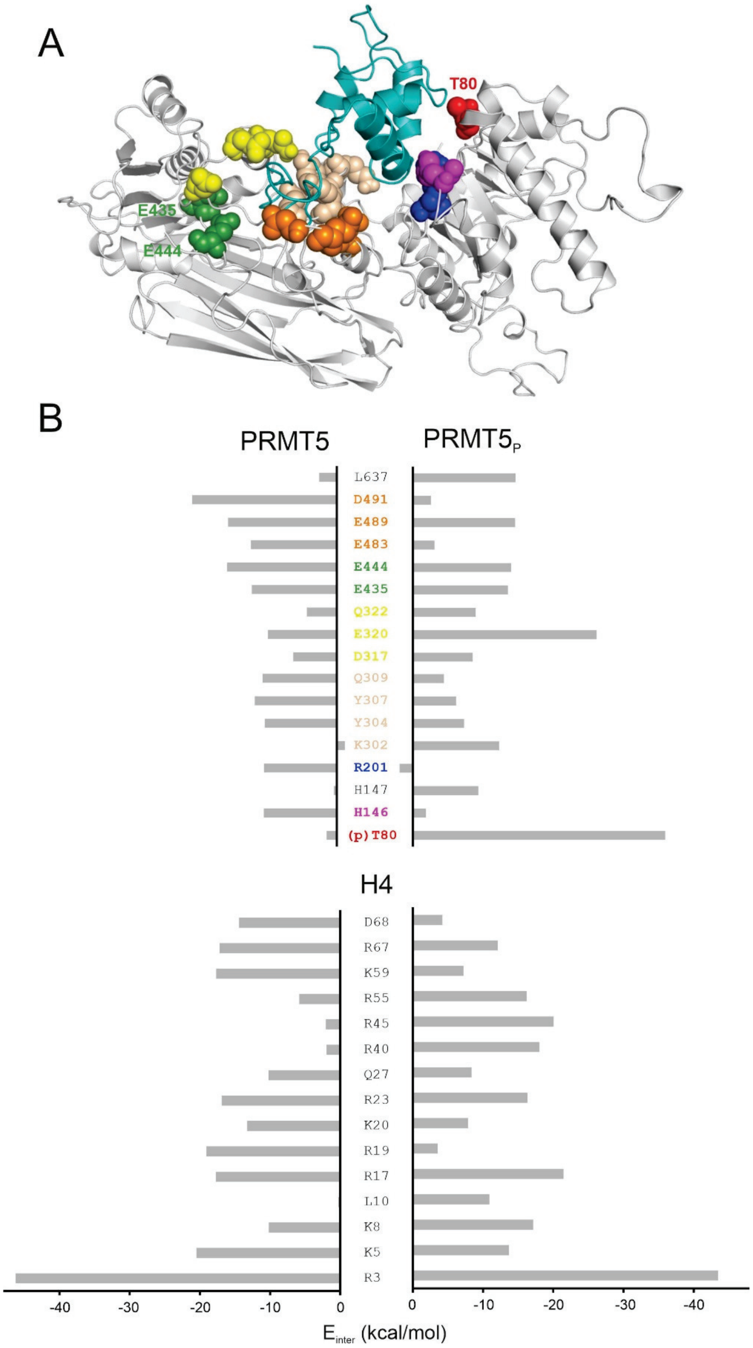


Figure 3. (**A**) The representative structure of the unmodified H4-PRMT5 complex in the pre-product

state (P2 in Figure 1B). Note that MEP50 is not shown in this figure due to space restrictions. For the structure of the full H4-PRMT5-MEP50 complex, please refer to Figure S4. Enzyme residues with the lowest E_{inter} are highlighted with colours and spheres, represented in both the structure and the energy bar chart; (**B**) E_{inter} values are calculated for the enzyme and the histone H4 residues in the unmodified PRMT5 and the phosphorylated (PRMT5$_P$) complexes in the pre-product (P2) state.

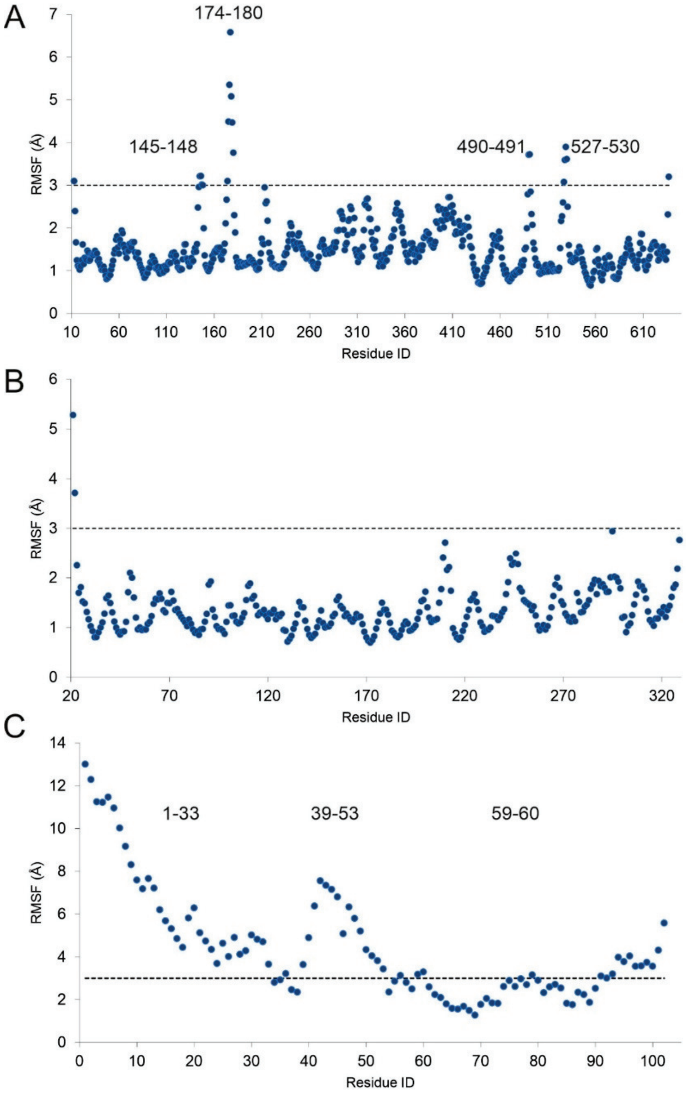
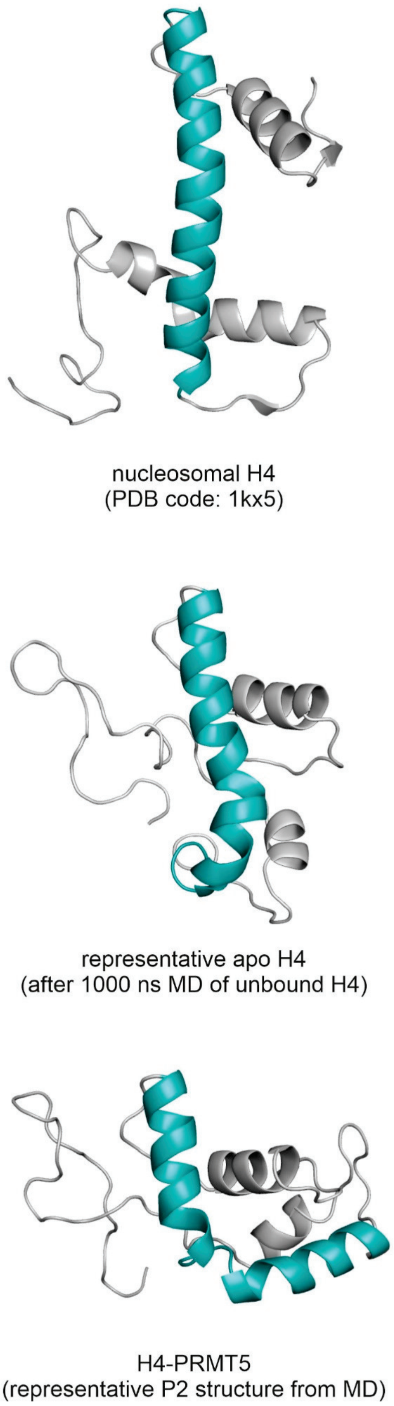


Figure 4. Root mean square fluctuation (RMSF) of each residue in the apo enzyme PRMT5 (**A**), MEP50 (**B**), and the unbound H4 (**C**) during 580, 580, and 1000 ns-long MD simulations, respectively. Regions with higher than 3 Å fluctuation are also marked at the top of the corresponding peaks.



nucleosomal H4
(PDB code: 1kx5)

representative apo H4
(after 1000 ns MD of unbound H4)

H4-PRMT5
(representative P2 structure from MD)

Figure 5. Histone H4 conformations (cartoon, grey) in the nucleosomal, unbound (apo) and PRMT5-complexed forms. Helix 3 (highlighted in teal) adopts a V-shaped conformation in the PRMT5-complexed structure, while it is mostly linear in the nucleosomal and apo forms.

2.3. Structural Effects of Phosphorylation on T80

The structural explanation of the increased methyltransferase activity of PRMT5$_P$ necessitates the building of the atomic resolution structure of the H4-PRMT5$_P$-MEP50 complex for a comparison with the unmodified PRMT5 version that is described in Section 1. The H4-PRMT5$_P$-MEP50 complex was constructed by adding a phosphate group to residue T80 of the energy-minimized H4-PRMT5 complex structure (Methods). The complex was energy-minimized and subjected to a 720 ns-long MD simulation, and E_{inter} was calculated between H4 and PRMT5$_P$-MEP50 along the whole trajectory (Figure 2B).

On average, H4-PRMT5$_P$ showed a lower E_{inter} when compared to the unmodified H4-PRMT5, regardless of the method used to select the representative P2 structure (Table S2). In region P2, the E_{inter} values of both complexes were comparably low (Figure 2B, Table S2), indicating that both systems reached an energetically favorable (pre-product) state (Figure 2A). The overall lower E_{inter} of H4-PRMT5$_P$ for the full trajectory is due to the lack of plateau P1 of a relatively high E_{inter} at H4-PRMT5 (Figure 2B, Table S2). Thus, in H4-PRMT5$_P$, the phosphorylation of T80 resulted in a favourable E_{inter}, lowering the activation energy barrier. The lower activation energy also means an increased methyltransferase activity of PRMT5$_P$, which was verified experimentally [9].

The abovementioned lowering of E_{inter} comes from the change of interaction network between the phosphorylated T80 (pT80) residue of PRMT5 and histone H4, as reflected by the per-residue E_{inter} analysis (Figure 3). The highest change of E_{inter} occurred on pT80, E320, and K302 of PRMT5$_P$ (Figure 3). Residue pT80 formed stable H-bridge interactions with R40 and R45 of histone H4 (Figures 6A and S3). The stabilization of these interactions for several hundreds of nanoseconds is reflected by the corresponding distance plots prepared for the full MD simulation (Figure S4). In the case of unmodified H4-PRMT5, the complex was formed by the C-terminal (R67 and D68) residues of Helix 3 of H4 interacting with a PRMT5 region different from T80 (blue and magenta in Figure 3). These interactions resulted in a V-shaped conformation (Figure 5) of Helix 3 in the H4-PRMT5 complex, while a linear Helix 3 was observed in H4-PRMT5$_P$ (Figure S5), similar to the apo form of H4 (Figure 5). This distortion of Helix 3 is an important factor of its unfavourable average E_{inter} (Figure 3, Table S2) in the case of the unmodified H4-PRMT5 complex. At the same time, binding of Helix 3 in an unchanged, linear form, that is, a "binding competent conformation" [36] contributed to the stronger interaction in H4-PRMT5$_P$.

Likewise, to the H4-PRMT5$_P$ complex, histone H4 residue R45 is also involved in the interaction of the phosphate groups of nucleosomal DNA. The H4-DNA interactions were listed using an experimental nucleosome structure (PDB code: 1kx5, [35], Methods, Table S3) and are shown in Figure 6b, with a close-up on the interacting residues marked as sticks. Among the interacting H4 residues (Figure 1A), R45 is one of the most important binding partners of the DNA phosphate groups (the phosphate groups at dT+7 and dG+8 are involved in the interactions with R45, Table S3, Figure 6b). Such electrostatic interactions are of primary importance in the stabilization of the nucleosome [37]. The abovementioned arginine–phosphate interactions are well-documented for nucleic acid partners [38], due to the ideal geometry, charge distribution, and flexibility of the arginine side-chain [39].

As our model shows that the phosphate group of pT80 residue of PRMT5$_P$ interacts with histone H4 residue R45 (Figure 6A), the above interactions (Figure 6B) of R45 with nucleosomal DNA cannot be formed in the presence of PRMT5$_P$. This structural conclusion of the present study is in line with the experimental fact that the nucleosome-bound histone H4 is not a substrate of PRMT5-MEP50 [11,16].

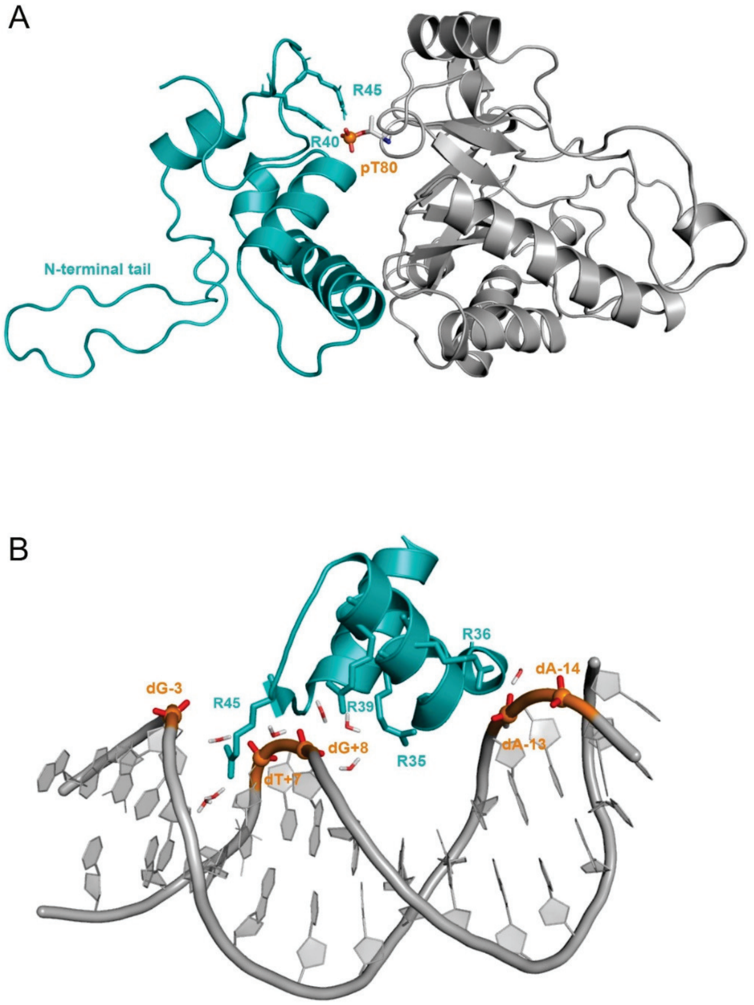


Figure 6. (**A**) Histone H4 (teal, cartoon) bound to PRMT5 (cartoon, grey) phosphorylated on T80 (pT80). Main anchoring residues, R40, R45, and pT80 are represented with sticks. MEP50 and certain parts of PRMT5 and H4 are not shown; (**B**) histone H4 (teal, cartoon) bound to DNA (grey, cartoon) in the experimental nucleosome structure (PDB code: 1kx5). An abridged structure of H4 is shown. Anchoring points are highlighted with sticks. Phosphates (orange) are represented with spheres in (**A**,**B**). In the 1kx5 structure, histone H4 and DNA are depicted as chain ID(s) F and I, and J, respectively. Note that histone H4 is embedded into the nucleosome, except for its N-terminal tail hanging out. The DNA binding domain of H4 is composed by K16-K20, based on the UniProt database (numbering according to PDB), which was identified in the nucleosome complex (1kx5), as well. However, residues before this region (Table S3) can also create interactions with DNA, due to the flexibility of the N-terminal tail. Furthermore, other H4 regions, such as R36-G48 and K79-T80, were also found to interact with DNA (Figure 1A, Table S3).

3. Materials and Methods

3.1. In Situ Fragment-Based Construction of H4 in Complex with PRMT5

For building the human H4-PRMT5 complex, the crystal structure of the human PRMT5-MEP50 complex, bound to a histone H4 fragment (1–8 residues), was used as the starting structure (4gqb). The missing amino acids of PRMT5-MEP50 were built by

SWISSMODEL online server [40], and terminals were capped. To build the full-length histone H4, peptide fragments from at least the 8th residue were needed. Therefore, all H4 structures linked in the UniProt database [41] under the human H4 UniProt entry (P62805) were checked in the PDB databank [42] and filtered based on the first resolved H4 residue and the length of the experimentally revealed H4 fragment. Structures of the human transcriptional protein (2kwo, [43]) and the nucleosome core particle (3x1v, [44]) bound to histone H4 were chosen, containing H4 residues 2–20 and 16–102, respectively (Figure 2B). After several attempts of superimposing of the abovementioned fragments to the resolved H4-PRMT5 complex, only the backbone alignment (with C, N, O, and Cα atoms) of histone H4 residues, such as K8 and K20, resulted a complex without collision (Figure 2B). The overlapping residues were cut (H4 residues 2–8 and 16–20 from the structures of 2kwo and 3x1v, respectively), and the superimposed H4 fragments were covalently attached in Maestro [45]; the H4-PRMT5(-MEP50) complex was equilibrated by the two-step energy minimization procedure (detailed later).

The root mean square deviation (RMSD) for Cα atoms was calculated during the MD simulation to check the equilibration of the structures. The RMSD of MEP50 fluctuated between 1.5 and 2.0 A (Figure S1), showing a low conformational flexibility; therefore, this structure was used as reference structure in all superposition. The Cα RMSD of PRMT5 and MEP50 was separately calculated after the least square fitting of snapshots to the first MEP50 structure. The RMSD of the unbound H4 structure was also calculated after the least square fitting of snapshots to the assembled H4 structure obtained from the energy-minimized H4-PRMT5 complex during the MD simulation. RMSD was calculated by Equation (1):

$$\mathrm{RMSD}(t_1, t_2) = \left[\frac{1}{M}\sum\nolimits_{i=1}^{N} m_i \|r_i(t_1) - r_i(t_2)\|^2\right]^{\frac{1}{2}}; M = \sum\nolimits_{i=1}^{N} m_i \tag{1}$$

where m_i is the atomic mass, and $r_i(t)$ is the position of atom i at time t.

The root mean square fluctuation (RMSF) was also calculated for each residue in the case of the apo PRMT5-MEP50 and unbound H4 structures, after the snapshots were fitted to the same structures like in the RMSD calculation. GROMACS [46] was used for all RMSD (command: *gmx rms*) and RMSF (command: *gmx rmsf*) calculations. The RMSF was calculated by Equation (2):

$$\mathrm{RMSF}_i = \left[\frac{1}{T}\sum_{t_j=1}^{T} \|r_i(t_j) - r_i^{\mathrm{ref}}\|^2\right]^{\frac{1}{2}} \tag{2}$$

where r_i is the position of the particle i, T is the time of the MD simulation, and ref denotes the reference position of the particle i.

To build the phosphorylated H4-PRMT5$_P$ complex, the phosphate group was covalently attached to T80:PRMT5 by PyMol [47]. The parameters of the phosphorylated threonine were obtained from a previous study [48].

3.2. Energy Minimization

Complexes and peptides were submitted to a two-step (steepest descent and conjugate gradient) energy minimization procedure before the MD simulation by GROMACS [46]. Molecules were placed in the centre of a cubic box, with a distance of 10 Å between the box and the solute atoms. The simulation box was filled with TIP3P [49] explicit water molecules and counter ions to neutralize the total charge of the system. The convergence threshold of steepest descent and conjugant gradient step of minimization was set to 100 and 10 kJ mol^{-1} nm^{-2}, respectively.

3.3. Molecular Dynamic Simulation

The complex, the unbound histone H4, and the apo PRMT5-MEP50 were submitted separately to a 720 ns-, a 1000 ns-, and a 580 ns-long MD simulation, respectively. In all cases, a TIP3P [49] explicit water model with an AMBER99SB-ILDN force field [50] was applied using the GROMACS program package [46], following the two-step energy minimization procedure (described above). Histone H4 and the enzymes could move freely; position restraints were not applied. For temperature-coupling, the velocity rescale and the Parrinello–Rahman algorithm were used. The solute and solvent were coupled separately, with a reference temperature of 310.15 K and a coupling time constant of 0.1 ps. The pressure was coupled by the Parrinello–Rahman algorithm and a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1}, and reference pressure of 1 bar. Particle mesh Ewald summation was used for long-range electrostatics. Van der Waals and Coulomb interactions had a cut-off at 11 Å. Periodic boundary conditions were treated after the finish of the calculations. After each trajectory, the periodic boundary effects were handled, the system was centred in the box, and the target molecules in subsequent frames were fit on the top of the first frame. The final trajectory, including all atomic coordinates of all frames, were converted to portable xdr-based xtc binary files.

3.4. Interaction Energy Calculations

The sum of Lennard-Jones (LJ) and Coulomb (Cb) intermolecular interaction energies were calculated [51] (3). The Coulomb term was globally calculated with a distance-dependent dielectric function [52] (4) and Amber partial charges [50,53], with per-residues during the simulations, and was represented as intermolecular interaction energy (E_{inter}).

$$\mathbf{E_{inter}} = \mathbf{E_{LJ}} + \mathbf{E_{Coulomb}} = \sum_{\mathbf{i,j}}^{\mathbf{N_E\,N_S}} \left(\frac{\mathbf{A_{ij}}}{\mathbf{r_{ij}^{12}}} - \frac{\mathbf{B_{ij}}}{\mathbf{r_{ij}^{6}}} + \frac{\mathbf{q_i q_j}}{\mathbf{4\pi\varepsilon_0\varepsilon_r r_{ij}}} \right) \tag{3}$$

$$\mathbf{A_{ij}} = \boldsymbol{\varepsilon_{ij}}\mathbf{R_{ij}^{12}}$$

$$\mathbf{B_{ij}} = \mathbf{2}\boldsymbol{\varepsilon_{ij}}\mathbf{R_{ij}^{6}}$$

$$\mathbf{R_{ij}} = \mathbf{R_i} + \mathbf{R_j}$$

$$\boldsymbol{\varepsilon_{ij}} = \sqrt{\boldsymbol{\varepsilon_i \varepsilon_j}}$$

where ε_{ij} is the potential well depth at equilibrium between the ith (substrate) and jth (enzyme) atoms; ε_0 is the permittivity of vacuum; ε_r = 1, relative permittivity; R_{ij} is the inter-nuclear distance at equilibrium between ith (substrate) and jth (enzyme) atoms; q is the partial charge of an atom; r_{ij} is the actual distance between the ith (substrate) and jth (enzyme) atoms; N_E is the number of enzyme atoms; N_S is the number of substrate atoms.

$$\boldsymbol{\varepsilon_r} = \mathbf{A} + \frac{\mathbf{B}}{\mathbf{1 + ke^{-\lambda Br}}} \tag{4}$$

where B = $\varepsilon_0 - A$, ε_0 is the dielectric constant of water at 25 °C, and A, λ, and k are constants.

The top ten residues with the lowest E_{inter} values at both the enzyme and the substrate sites for the unmodified and phosphorylated H4-PRMT5 complexes were chosen and merged to prepare the E_{inter} bar chart (Figure 3).

3.5. Selection of Representative Structures by Structural Clustering and Interaction Energy Differences

Representative structures were selected using a structure-based clustering from the following four sets of structures: (i) unmodified H4-PRMT5 complex structures from the P1 and (ii) P2 plateaus; (iii) phosphorylated H4-PRMT5 complex structures from the P1 and (iv) P2 plateaus. The clustering procedure contained the following steps: The average atomic coordinates were calculated for all four set of structures using a bash script, which prints the x, y, and z atomic coordinates of all structures in the set into separate text files. The atomic coordinates were structured into a pdb file format and used as average structures. Finally, the RMSD values between the average structure and each complex were calculated by an in-house program, rmsd, and the structure with the lowest RMSD value was selected as the representative structure.

An E_{inter}-based selection of representative structures was also performed in the pre-product state (P2), as E_{inter} should have a similar value in the unmodified and phosphorylated structures. Accordingly, representatives for the P2 section were determined by calculating the E_{inter}-differences of the unmodified H4-PRMT5 and H4-PRMT5$_P$ complexes of the last twenty-five frames. Structures with the lowest E_{inter}-difference were chosen as unmodified and phosphorylated H4-PRMT5 representatives for the P2 section (Figure 2B).

3.6. Determination of DNA Binding Domain of H4 in Nucleosome

The structure of the nucleosome core particle (1kx5, [35]) contained two full-length histone H4s (chain ID: B and F). Histone H4 residues within 3.5 Å distance from the DNA chains were collected for both of the H4 peptides, using an in-house program. Amino acids taking place in DNA binding in the case of both peptides were determined as DNA binding domains of H4 in the nucleosome (Table S3).

4. Conclusions

A three-dimensional structure of the unmodified and phosphorylated human PRMT5-MEP50, in complex with the full-length histone H4 protein, was modeled. Molecular dynamic simulations and subsequent analyses provided an atomic level explanation for two important experimental findings: (1) the increased methyltransferase activity of the phosphorylated PRMT5 when compared to the unmodified type [9]; (2) the PRMT5 methylates only the free form of histone H4 not bound to the nucleosome [11,16,20]. We expect that our findings will foster the design of new inhibitors and help in uncovering further structure–function relationships of PRMT enzymes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms231911316/s1.

Author Contributions: Conceptualization, C.H. and R.B.; methodology, R.B., C.H. and B.B.; investigation, R.B., B.B. and B.Z.Z.; resources, C.H.; data curation, R.B.; writing—original draft preparation, R.B., C.H. and B.L.; writing—review and editing, R.B., C.H., B.B., B.Z.Z. and B.L.; visualization, R.B.; supervision, C.H.; project administration, C.H.; funding acquisition, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Hungarian National Research, Development, and Innovation Office (K123836). Project no. TKP2021-EGA-16 was implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the EGA-16 funding scheme. Project no. TKP2021-EGA-13 was implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the EGA-13 funding scheme and the Lorand Eotvos Research Network. Project no. RRF-2.3.1-21-2022-00015 was implemented with the support provided by the European Union. This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The work was supported by the ÚNKP-21-5 and ÚNKP-21-3-II New National Excellence Program of the Ministry for Innovation and Technology and by the PTE ÁOK-KA 2021/KA-2021-39. The project was supported by the European Union, co-financed by the European Social Fund, project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research were achieved using the DECI resource Archer2 based in the UK at the National Supercomputing Service with support from the PRACE aisbl.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

PTM	Post-translational modification
PRMT	Protein arginine methyltransferase
PRMT5	Protein arginine methyltransferase 5
H4	Histone H4 peptide
PRMT5$_P$	Human PRMT5 enzyme phosphorylated on T80 residue
pT80	Phosphorylated T80 residue
MEP50	Methylosome protein 50
G	Free energy
E_{inter}	The sum of Coulomb and Lennard-Jones intermolecular interaction energy
MD	Molecular dynamics
PDB	Protein Data Bank
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation

References

1. Griffin, G.K.; Wu, J.; Iracheta-Vellve, A.; Patti, J.C.; Hsu, J.; Davis, T.; Dele-Oni, D.; Du, P.P.; Halawi, A.G.; Ishizuka, J.J.; et al. Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature* **2021**, *595*, 309–314. [CrossRef] [PubMed]
2. Izzo, L.T.; Wellen, K.E. Histone lactylation links metabolism and gene regulation. *Nature* **2019**, *574*, 492–493. [CrossRef] [PubMed]
3. Wojcik, F.; Dann, G.P.; Beh, L.Y.; Debelouchina, G.T.; Hofmann, R.; Muir, T.W. Functional crosstalk between histone H2B ubiquitylation and H2A modifications and variants. *Nat. Commun.* **2018**, *9*, 1394. [CrossRef]
4. Zsidó, B.Z.; Hetényi, C. Molecular Structure, Binding Affinity, and Biological Activity in the Epigenome. *Int. J. Mol. Sci.* **2020**, *21*, 4134. [CrossRef] [PubMed]
5. Hwang, W.L.; Deindl, S.; Harada, B.T.; Zhuang, X. Histone H4 tail mediates allosteric regulation of nucleosome remodelling by linker DNA. *Nature* **2014**, *512*, 213–217. [CrossRef]
6. Strahl, B.D.; Allis, C.D. The language of covalent histone modifications. *Nature* **2000**, *403*, 41–45. [CrossRef] [PubMed]
7. Heinz, S.; Romanoski, C.E.; Benner, C.; Allison, K.A.; Kaikkonen, M.U.; Orozco, L.D.; Glass, C.K. Impact of natural genetic variation on enhancer selection and function. *Nature* **2013**, *503*, 487–492. [CrossRef]
8. Jarrold, J.; Davies, C.C. PRMTs and Arginine Methylation: Cancer's Best-Kept Secret? *Trends Mol. Med.* **2019**, *25*, 993–1009. [CrossRef]
9. Sipos, A.; Iván, J.; Bécsi, B.; Darula, Z.; Tamás, I.; Horváth, D.; Medzihradszky, K.F.; Erdődi, F.; Lontay, B. Myosin phosphatase and RhoA-activated kinase modulate arginine methylation by the regulation of protein arginine methyltransferase 5 in hepatocellular carcinoma cells. *Sci. Rep.* **2017**, *7*, 40590. [CrossRef]
10. Ancelin, K.; Lange, U.C.; Hajkova, P.; Schneider, R.; Bannister, A.J.; Kouzarides, T.; Surani, M.A. Blimp1 associates with Prmt5 and directs histone arginine methylation in mouse germ cells. *Nat. Cell Biol.* **2006**, *8*, 623–630. [CrossRef]
11. Fulton, M.D.; Cao, M.; Ho, M.-C.; Zhao, X.; Zheng, Y.G. The macromolecular complexes of histones affect protein arginine methyltransferase activities. *J. Biol. Chem.* **2021**, *297*, 101123. [CrossRef] [PubMed]
12. Tee, W.-W.; Pardo, M.; Theunissen, T.W.; Yu, L.; Choudhary, J.S.; Hajkova, P.; Surani, M.A. Prmt5 is essential for early mouse development and acts in the cytoplasm to maintain ES cell pluripotency. *Genes Dev.* **2010**, *24*, 2772–2777. [CrossRef] [PubMed]
13. Dong, F.; Li, Q.; Yang, C.; Huo, D.; Wang, X.; Ai, C.; Kong, Y.; Sun, X.; Wang, W.; Zhou, Y.; et al. PRMT2 links histone H3R8 asymmetric dimethylation to oncogenic activation and tumorigenesis of glioblastoma. *Nat. Commun.* **2018**, *9*, 4552. [CrossRef] [PubMed]
14. Pal, S.; Vishwanath, S.N.; Erdjument-Bromage, H.; Tempst, P.; Sif, S. Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes. *Mol. Cell. Biol.* **2004**, *24*, 9630–9645. [CrossRef]
15. Antonysamy, S. The Structure and Function of the PRMT5:MEP50 Complex. In *Macromolecular Protein Complexes*; Harris, J.R., Marles-Wright, J., Eds.; Subcellular Biochemistry; Springer International Publishing: Cham, Switzerland, 2017; Volume 83, pp. 185–194, ISBN 978-3-319-46501-2.

16. Burgos, E.S.; Wilczek, C.; Onikubo, T.; Bonanno, J.B.; Jansong, J.; Reimer, U.; Shechter, D. Histone H2A and H4 N-terminal Tails Are Positioned by the MEP50 WD Repeat Protein for Efficient Methylation by the PRMT5 Arginine Methyltransferase *. *J. Biol. Chem.* **2015**, *290*, 9674–9689. [CrossRef]
17. Antonysamy, S.; Bonday, Z.; Campbell, R.M.; Doyle, B.; Druzina, Z.; Gheyi, T.; Han, B.; Jungheim, L.N.; Qian, Y.; Rauch, C.; et al. Crystal structure of the human PRMT5:MEP50 complex. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17960–17965. [CrossRef]
18. Liu, F.; Zhao, X.; Perna, F.; Wang, L.; Koppikar, P.; Abdel-Wahab, O.; Harr, M.W.; Levine, R.L.; Xu, H.; Tefferi, A.; et al. JAK2V617F-mediated phosphorylation of PRMT5 down-regulates its methyltransferase activity and promotes myeloproliferation. *Cancer Cell* **2011**, *19*, 283–294. [CrossRef]
19. Lattouf, H.; Kassem, L.; Jacquemetton, J.; Choucair, A.; Poulard, C.; Trédan, O.; Corbo, L.; Diab-Assaf, M.; Hussein, N.; Treilleux, I.; et al. LKB1 regulates PRMT5 activity in breast cancer. *Int. J. Cancer* **2019**, *144*, 595–606. [CrossRef]
20. Ho, M.-C.; Wilczek, C.; Bonanno, J.B.; Xing, L.; Seznec, J.; Matsui, T.; Carter, L.G.; Onikubo, T.; Kumar, P.R.; Chan, M.K.; et al. Structure of the Arginine Methyltransferase PRMT5-MEP50 Reveals a Mechanism for Substrate Specificity. *PLoS ONE* **2013**, *8*, e57008. [CrossRef]
21. Sun, L.; Wang, M.; Lv, Z.; Yang, N.; Liu, Y.; Bao, S.; Gong, W.; Xu, R.-M. Structural insights into protein arginine symmetric dimethylation by PRMT5. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20538–20543. [CrossRef]
22. Kawamura, S.; Palte, R.L.; Kim, H.-Y.; Saurí, J.; Sondey, C.; Mansueto, M.S.; Altman, M.D.; Machacek, M.R. Design and Synthesis of Unprecedented 9- and 10-Membered Cyclonucleosides with PRMT5 Inhibitory Activity. *Bioorg. Med. Chem.* **2022**, *66*, 116820. [CrossRef] [PubMed]
23. Smith, C.R.; Aranda, R.; Bobinski, T.P.; Briere, D.M.; Burns, A.C.; Christensen, J.G.; Clarine, J.; Engstrom, L.D.; Gunn, R.J.; Ivetac, A.; et al. Fragment-Based Discovery of MRTX1719, a Synthetic Lethal Inhibitor of the PRMT5 MTA Complex for the Treatment of MTAP-Deleted Cancers. *J. Med. Chem.* **2022**, *65*, 1749–1766. [CrossRef]
24. Jensen-Pergakes, K.; Tatlock, J.; Maegley, K.A.; McAlpine, I.J.; McTigue, M.; Xie, T.; Dillon, C.P.; Wang, Y.; Yamazaki, S.; Spiegel, N.; et al. SAM-Competitive PRMT5 Inhibitor PF-06939999 Demonstrates Antitumor Activity in Splicing Dysregulated NSCLC with Decreased Liability of Drug Resistance. *Mol. Cancer Ther.* **2022**, *21*, 3–15. [CrossRef] [PubMed]
25. Quiroz, R.V.; Reutershan, M.H.; Schneider, S.E.; Sloman, D.; Lacey, B.M.; Swalm, B.M.; Yeung, C.S.; Gibeau, C.; Spellman, D.S.; Rankic, D.A.; et al. The Discovery of Two Novel Classes of 5,5-Bicyclic Nucleoside-Derived PRMT5 Inhibitors for the Treatment of Cancer. *J. Med. Chem.* **2021**, *64*, 3911–3939. [CrossRef]
26. Candito, D.A.; Ye, Y.; Quiroz, R.V.; Reutershan, M.H.; Witter, D.; Gadamsetty, S.B.; Li, H.; Saurí, J.; Schneider, S.E.; Lam, Y.; et al. Development of a Flexible and Robust Synthesis of Tetrahydrofuro[3,4-b]Furan Nucleoside Analogues. *J. Org. Chem.* **2021**, *86*, 5142–5151. [CrossRef]
27. McKinney, D.C.; McMillan, B.J.; Ranaghan, M.J.; Moroco, J.A.; Brousseau, M.; Mullin-Bernstein, Z.; O'Keefe, M.; McCarren, P.; Mesleh, M.F.; Mulvaney, K.M.; et al. Discovery of a First-in-Class Inhibitor of the PRMT5–Substrate Adaptor Interaction. *J. Med. Chem.* **2021**, *64*, 11148–11168. [CrossRef]
28. Mulvaney, K.M.; Blomquist, C.; Acharya, N.; Li, R.; Ranaghan, M.J.; O'Keefe, M.; Rodriguez, D.J.; Young, M.J.; Kesar, D.; Pal, D.; et al. Molecular Basis for Substrate Recruitment to the PRMT5 Methylosome. *Mol. Cell* **2021**, *81*, 3481–2495.e7. [CrossRef]
29. Palte, R.L.; Schneider, S.E.; Altman, M.D.; Hayes, R.P.; Kawamura, S.; Lacey, B.M.; Mansueto, M.S.; Reutershan, M.; Siliphaivanh, P.; Sondey, C.; et al. Allosteric Modulation of Protein Arginine Methyltransferase 5 (PRMT5). *ACS Med. Chem. Lett.* **2020**, *11*, 1688–1693. [CrossRef] [PubMed]
30. Lin, H.; Wang, M.; Zhang, Y.W.; Tong, S.; Leal, R.A.; Shetty, R.; Vaddi, K.; Luengo, J.I. Discovery of Potent and Selective Covalent Protein Arginine Methyltransferase 5 (PRMT5) Inhibitors. *ACS Med. Chem. Lett.* **2019**, *10*, 1033–1038. [CrossRef]
31. Bonday, Z.Q.; Cortez, G.S.; Grogan, M.J.; Antonysamy, S.; Weichert, K.; Bocchinfuso, W.P.; Li, F.; Kennedy, S.; Li, B.; Mader, M.M.; et al. LLY-283, a Potent and Selective Inhibitor of Arginine Methyltransferase 5, PRMT5, with Antitumor Activity. *ACS Med. Chem. Lett.* **2018**, *9*, 612–617. [CrossRef] [PubMed]
32. Mavrakis, K.J.; McDonald, E.R.; Schlabach, M.R.; Billy, E.; Hoffman, G.R.; deWeck, A.; Ruddy, D.A.; Venkatesan, K.; Yu, J.; McAllister, G.; et al. Disordered Methionine Metabolism in MTAP/CDKN2A-Deleted Cancers Leads to Dependence on PRMT5. *Science* **2016**, *351*, 1208–1213. [CrossRef]
33. Duncan, K.W.; Rioux, N.; Boriack-Sjodin, P.A.; Munchhof, M.J.; Reiter, L.A.; Majer, C.R.; Jin, L.; Johnston, L.D.; Chan-Penebre, E.; Kuplast, K.G.; et al. Structure and Property Guided Design in the Identification of PRMT5 Tool Compound EPZ015666. *ACS Med. Chem. Lett.* **2016**, *7*, 162–166. [CrossRef] [PubMed]
34. Chan-Penebre, E.; Kuplast, K.G.; Majer, C.R.; Boriack-Sjodin, P.A.; Wigle, T.J.; Johnston, L.D.; Rioux, N.; Munchhof, M.J.; Jin, L.; Jacques, S.L.; et al. A Selective Inhibitor of PRMT5 with in Vivo and in Vitro Potency in MCL Models. *Nat. Chem. Biol.* **2015**, *11*, 432–437. [CrossRef] [PubMed]
35. Davey, C.A.; Sargent, D.F.; Luger, K.; Maeder, A.W.; Richmond, T.J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study. *J. Mol. Biol.* **2002**, *319*, 1097–1113. [CrossRef]
36. Abdelsattar, A.S.; Mansour, Y.; Aboul-ela, F. The Perturbed Free-Energy Landscape: Linking Ligand Binding to Biomolecular Folding. *ChemBioChem* **2021**, *22*, 1499–1516. [CrossRef]
37. Fu, I.; Geacintov, N.E.; Broyde, S. Molecular dynamics simulations reveal how H3K56 acetylation impacts nucleosome structure to promote DNA exposure for lesion sensing. *DNA Repair* **2021**, *107*, 103201. [CrossRef]

38. Luscombe, N.M.; Laskowski, R.A.; Thornton, J.M. Amino acid–base interactions: A three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874. [CrossRef]
39. Fuhrmann, J.; Clancy, K.W.; Thompson, P.R. Chemical Biology of Protein Arginine Modifications in Epigenetic Regulation. *Chem. Rev.* **2015**, *115*, 5413–5461. [CrossRef]
40. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. [CrossRef] [PubMed]
41. Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bursteinas, B. The UniProt Consortium UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
42. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
43. Zeng, L.; Zhang, Q.; Li, S.; Plotnikov, A.N.; Walsh, M.J.; Zhou, M.-M. Mechanism and Regulation of Acetylated Histone Binding by the Tandem PHD Finger of DPF3b. *Nature* **2010**, *466*, 258–262. [CrossRef] [PubMed]
44. Padavattan, S.; Shinagawa, T.; Hasegawa, K.; Kumasaka, T.; Ishii, S.; Kumarevel, T. Structural and functional analyses of nucleosome complexes with mouse histone variants TH2a and TH2b, involved in reprogramming. *Biochem. Biophys. Res. Commun.* **2015**, *464*, 929–935. [CrossRef] [PubMed]
45. *Schrödinger Release 2017–4: Maestro*; Schrödinger, LLC: New York, NY, USA, 2017.
46. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25. [CrossRef]
47. The PyMOL Molecular Graphics System. Version 1.8; Schrödinger, LLC: New York, NY, USA, 2015.
48. Homeyer, N.; Horn, A.H.C.; Lanig, H.; Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: Phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281–289. [CrossRef] [PubMed]
49. Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105*, 9954–9960. [CrossRef]
50. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958. [CrossRef]
51. Horváth, I.; Jeszenői, N.; Bálint, M.; Paragi, G.; Hetényi, C. A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level. *Int. J. Mol. Sci.* **2019**, *20*, E4384. [CrossRef]
52. Mehler, E.L.; Solmajer, T. Electrostatic effects in proteins: Comparison of dielectric and charge models. *Protein Eng. Des. Sel.* **1991**, *4*, 903–910. [CrossRef]
53. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations IV: Van der Waals parameterization. *J. Phys. Chem. B* **2012**, *116*, 7088–7101. [CrossRef]

D11

International Journal of
Molecular Sciences

MDPI

Article

Towards Unraveling the Histone Code by Fragment Blind Docking

Mónika Bálint [1], István Horváth [2], Nikolett Mészáros [3] and Csaba Hetényi [1,*]

[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary; monibalint18@gmail.com

[2] Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720 Szeged, Hungary; horvathi@gmx.de

[3] Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary; eperke93@gmail.com

* Correspondence: csabahete@yahoo.com

Received: 19 December 2018; Accepted: 16 January 2019; Published: 19 January 2019

Abstract: Histones serve as protein spools for winding the DNA in the nucleosome. High variability of their post-translational modifications result in a unique code system often responsible for the pathomechanisms of epigenetics-based diseases. Decoding is performed by reader proteins via complex formation with the N-terminal peptide tails of histones. Determination of structures of histone-reader complexes would be a key to unravel the histone code and the design of new drugs. However, the large number of possible histone complex variations imposes a true challenge for experimental structure determination techniques. Calculation of such complexes is difficult due to considerable size and flexibility of peptides and the shallow binding surfaces of the readers. Moreover, location of the binding sites is often unknown, which requires a blind docking search over the entire surface of the target protein. To accelerate the work in this field, a new approach is presented for prediction of the structure of histone H3 peptide tails docked to their targets. Using a fragmenting protocol and a systematic blind docking method, a collection of well-positioned fragments of the H3 peptide is produced. After linking the fragments, reconstitution of anchoring regions of the target-bound H3 peptide conformations was possible. As a first attempt of combination of blind and fragment docking approaches, our new method is named fragment blind docking (FBD).

Keywords: peptide; interaction; translation; methylation; target; ligand

1. Introduction

In the past decades, epigenetics has opened up new pathways of drug discovery [1]. Among epigenetic events, post-translational modifications (PTM) of histone proteins are of particular interest [2–4]. The epigenetic role of these PTMs can be explained by their altering effect on the chromatin structure influencing histone–DNA and histone–histone contacts in nucleosomes of chromatin fibers. Histones are small, conserved eukaryotic proteins, with a very flexible N-terminal tail (Figure 1A) composed of ca. thirty-six amino acids [5] at histone H3. The tail can be covalently modified at side-chains of amino acids K, R, T, and S. The resulted PTMs may have diverse chemical composition such as methylation, acetylation, phosphorylation, etc. If considering methylation as an example, three (mono-, di-, and tri-methylated) PTMs can be derived by replacing hydrogen atoms of the charged amino group of the side-chain of K. The N-terminal histone H3 peptide has seven locations of K (Figure 1A) and methylation can result in 4^7 PTM variations (four comes from the three PTMs plus the non-modified K). In this way, an enormous number of PTM variations can be derived if all above-mentioned amino acids and modifying groups are considered.

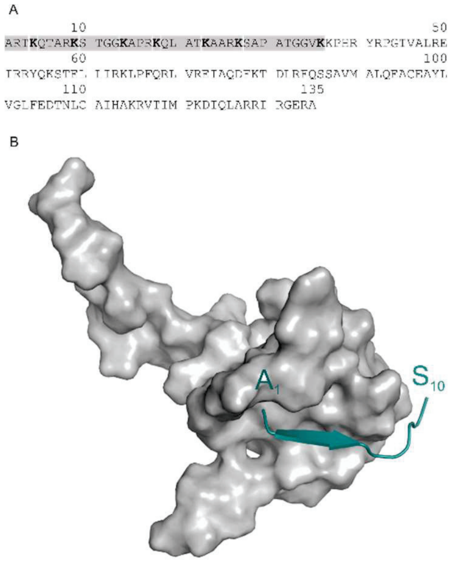

A
10
50
ARTKQTARKS TGGKAPRKQL ATKAARKSAP ATGGVKKPHR YRPGTVALRE
60
100
IRRYQKSTEL LIRKLPFQRL VREIAQDFKT DLRFQSSAVM ALQEACEAYL
110
135
VGLFEDTNLC AIHAKRVTIM PKDIQLARRI RGERA
B
A_1
S_{10}

Figure 1. (**A**) The sequence of H3 with the 36-amino-acid long N-terminal peptide marked in grey. (**B**) The complex of the H3 decapeptide tail (green) bound to AIRE-PHD (System 2ke1) test case.

Beyond the genetic code, a "histone code" was proposed [6] based on this large number of PTM variations. The histone code is fundamental in the epigenetics of chromatin-related pathomechanisms of various diseases [2,3] and can be "decoded and translated" by chromatin-associated reader proteins [5–7]. Atomic resolution structures of histone-reader (writer) complexes are keys of unraveling the histone code and drug design. The large number of PTM variations of histones yields a similarly large number of possible complexes. Large scale structure determination of such complexes is challenging even for high throughput crystallographic [8,9] techniques.

To answer the challenge, the use of theoretical approaches would be an alternative to experimental techniques. Computational docking [10–12] of the histone peptide to the reader (writer) target would be an obvious theoretical approach in this case. However, there are practical problems with the use of this approach. First of all, the whole N-terminal peptide tail is too large for fast docking approaches [13] due to its large torsional flexibility resulting in a complicated search problem [14]. Secondly, even the approximate location of binding sites of the histone peptide on the target surface is unknown in many cases. Approximate, pre-docking location of the binding sites is further encumbered by the shallow binding surfaces of reader proteins without deep pockets.

Docking of fragments instead of the whole histone peptide would tackle the first problem. Fragment-based approaches [15] have been used in past studies with success. In the case of peptide docking, fragmenting will result in a reduced degree of torsional freedom, and a relieved search. The second problem can be answered by the blind docking approach [16–18] which scans the entire surface of the target molecule without prior knowledge of the binding site. In a recent paper, a systematic blind docking method was released [19] for finding all possible binding modes of a ligand on a target.

In the present study, a combination of the fragment and systematic blind docking approaches is introduced and tested for complexes of histone peptides with their targets. The resulted methodology is named after the parent techniques as fragment blind docking (FBD).

2. Results and Discussion

A set of five histone H3 peptide–target complexes (Table 1) was used for elaboration and test of FBD. The complexes contain both methylated (3qla, 5tdw) and non-methylated (2ke1, 2pvc, 4lk9) histone H3 tails. Experimental complex structures deposited in the Protein Databank (PDB, [20]) mostly include deca-peptide-sized sections of the tail [21]. Seemingly, it is challenging to capture all 36 amino acids of the tail for current structure determination techniques such as X-ray or NMR.

This can be explained by the mobility, and by the lack of or weak interactions of the C-terminal end of the tail with the target protein which will be further analysed in the next paragraphs. The above experimental difficulties and need for of determination of the histone-target complexes motivated the elaboration of FBD. In the forthcoming Sections, the main steps (Figure 1) of FBD are introduced using the structure of H3 peptide in complex with autoimmune regulator protein plant homeodomain (System 2ke1) as an example (Figure 1B).

Table 1. Test systems.

PDB ID	Target	Ligand (Histone H3 Peptide) *
2ke1	autoimmune regulator protein plant homeodomain (AIRE-PHD)	ARTKQTARKS
2pvc	DNA (cytosine-5)-methyltransferase 3-like (DNMT3L)	ARTKQTA
3qla	transcriptional regulator ATRX-ADD domain (ATRX-DNMT3-DNMT3L)	ARTKQTARK(Me$_3$)S
4lk9	histone acetyltransferase KAT6A	ARTKQTARKSTGG
5tdw	Set domain containing protein 3	ARTK(Me$_3$)QTARKST

* One letter amino acid codes are used with PTMs marked in brackets after the modified amino acids.

2.1. Fragmenting

As the name of FBD indicates, the protocol (Figure 2) is based on the splitting of the original histone H3 peptide ligand (H3) into fragments. The fragments have smaller size, and fewer active torsions than the original peptide. Due to the decrease of their overall freedom, they are expected to impose less challenge on the docking search algorithm which is the rationale behind the fragmenting approaches. However, as fragment docking has been used in focused mode (see the Introduction for references), it is not clear how small fragments give the best possible results in a blind search over the entire surface of the target. To answer this question, fragments of H3 were designed according to a systematic scheme (Figure 3). The size of fragments ranged between di- and tetra-peptides. Notably, we did not use fragments larger than tetra-peptides as the above benefits of fragmenting would diminish beyond this size [12]. According to the starting position of fragmenting, two series of fragments were produced and named as N- and C-terminal fragments. To avoid single amino acid fragments, dipeptide fragmentation was carried out for the cases of tri- or tetrapeptide fragmenting, if the remaining C- or N terminal sequence were either a tetra- or pentapetides, respectively.

This fragmenting scheme is systematic and at the same time diversifies the sets of fragments used for docking. The original H3 peptide was cut at the amide bond between the carbon and the nitrogen atoms. In general, acetyl (Ac-) and imino-methyl (-NHMe) groups are used to cap/block the free N- and C-terminal cut ends of the peptides to mimic the backbone. In the case of the example system 2ke1, fragmenting resulted in thirteen different peptides (Figure 3). The structures of all fragments were prepared (Methods) and collected in a library for the blind docking cycles of the Wrapping step.

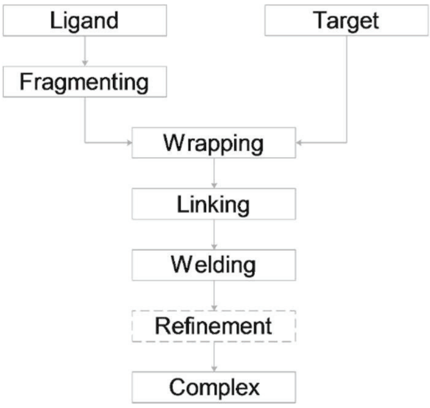


Figure 2. Fragment blind docking (FBD).

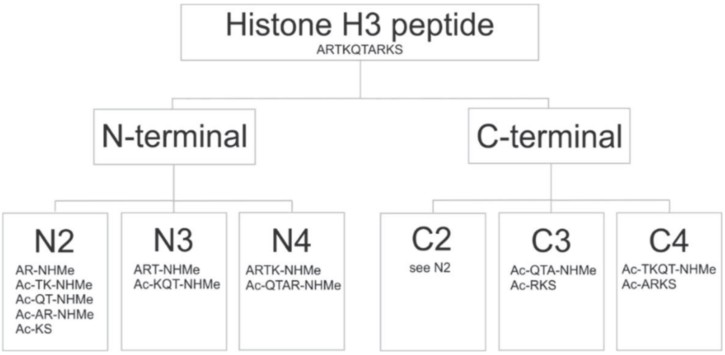


Figure 3. The fragmenting scheme.

2.2. *Wrapping*

The next step of FBD involves the Wrapper module of a new method [19] which covers the entire surface of the target molecule by a mono-layer of the copies of a peptide ligand using a series of blind docking cycles. Wrapping the target into ligand copies allows systematic mapping of all possible binding modes of a ligand [19]. In the present study, wrapping of target proteins of all test systems were performed by all fragments. In the case of 2ke1, blind docking of 13 unique fragments yielded more than nine thousand docked ligand conformations (see Methods) which was distilled into 529 binding modes (binding sites and conformations) used in the next, linking step. All binding modes were compared to the corresponding experimental conformation of the full H3 peptide using a standard procedure as implemented in the Wrapper module [19]. Briefly, a root mean squared deviation (RMSD, Equation 1) value was calculated for the binding modes obtained in a wrapping cycle, and the best RMSD values were collected for each peptide fragment (Supplementary Table S1). An abridged version of the results is shown in Table 2 with binding modes of an RMSD $\leq$ 4 Å. This agreement with the experimental conformation was found appropriate for the linking step (Figure 2) described in the next Section. Serial number of the wrapping cycle, and the energy rank (see Methods for ranking details) holding the binding mode are also indicated in Table 2 and Supplementary Table S1. The results of the fragmenting-wrapping tests of our systems (Table 1) allowed investigation of various factors influencing FBD, such as fragment size, chemistry of binding, secondary structure of the ligand and fragment ends as described in the following paragraphs.

Table 2. Docking results *.

System	Fragment Type	Fragment Sequence	#Cycle	#Rank	RMSD (Å)
2ke1	N2 = C2	AR-NHMe	1	4	1.9
	N2 = C2	Ac-TK-NHMe	1	1	2.8
	N3	Ac-KQT-NHMe	1	1	2.4
2pvc	N2	AR-NHMe	2	5	3.2
	N2	Ac-TK-NHMe	1	3	2.8
	N3	Ac-KQTA	1	1	3.0
	C2	Ac-KQ-NHMe	1	2	1.7
3qla	N2 = C2	Ac-QT-NHMe	1	12	2.0
	N2 = C2	Ac-K(Me$_3$)S	1	2	3.7
	N4	ARTK-NHMe	1	1	3.3
4lk9	N2	AR-NHMe	1	1	2.9
	N2	Ac-TK-NHMe	1	4	4.0
5tdw	N2	AR-NHMe	1	1	1.7

* Structures are shown in Supplementary Figures S1–S5. $^{\#}$ Serial number.

2.3. Fragment Size

It can be observed that in most of the cases, dipeptide fragments were selected for Table 2. Among the dipeptide fragments, docking of the AR-NHMe fragment was the most successful at Systems 2ke1, 4lk9, and 5tdw. Excellent RMSD values were obtained ranging from 1.7 to 2.0 Å. In these three cases the structure with the best RMSD was found in the first wrapping cycle, ranking in the top five cluster representative. The R_2 residue of this dipeptide seems to be an important anchoring residue of the H3 histone tail. For an in-depth analysis of target–ligand interactions, the number of intermolecular contacts (N_{inter}) and per-residue intermolecular interaction energy values (E_{inter}) were calculated for the energy-minimized complex structures of Table 1 as described in Methods. The results of the analysis (Figure 4, Supplementary Tables S2 and S3) underline that R_2 is indeed an important anchoring residue having the largest N_{inter} and best E_{inter} values. For system 3qla, a mis-docked AR-NHMe (Supplementary Table S1) was produced due to the lack of negatively charged residues and the targeted ATRX-DNMT3-DNMT3L (ATRX-ADD) domain side resulting in a single contact at R_2 (Figure 4, Supplementary Table S2). In case of 2pvc, the AR-NHMe fragment found the binding conformation at an RMSD of 3.2 Å in the second cycle of wrapping. This is due to the relatively small E_{inter} of R_2, if compared to that of K4 which is the main anchoring residue (Figure 4). The importance of K_4 is also reflected by its successful docking when K_4 was part of the fragmented sequence (N2, N3 and C2 in Table 2).

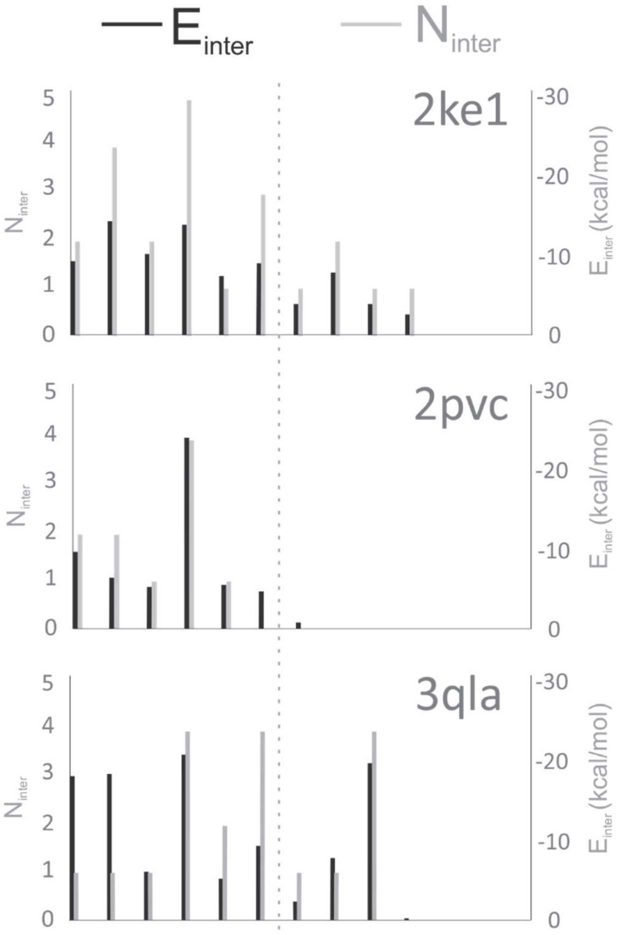


Figure 4. *Cont.*

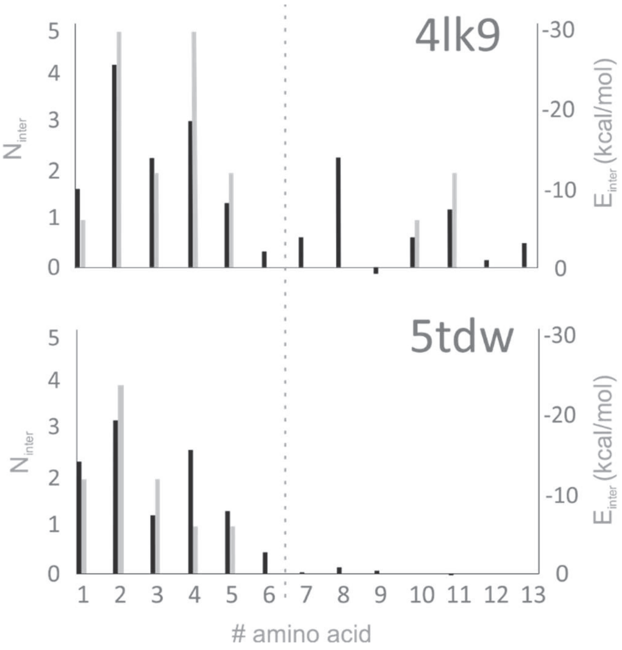


Figure 4. Per-residue values of target–ligand intermolecular interaction energy (E_{inter}) and number of intermolecular interactions (N_{inter}) calculated for the first 13 amino acids of the histone H3 ligand. The dotted line represents an approximate border of the N-terminal anchoring region where most of the target–ligand interactions act.

Previous docking studies claimed [12] that appropriately docked conformations can be obtained for small peptides with low number of active torsions whereas large peptides are true challenges for most of the docking procedures [12,22,23]. In agreement with these previous reports, low RMSD values were obtained mostly for dipeptide fragments of the N2 or C2 series (Table 2). Depending on the starting point of fragmenting (C- or N-terminus) different sets of peptides could be obtained. Plausibly, identical fragment sets can be obtained if, e.g., dipeptide fragments are produced from a H3 peptide of even number of residues (Figure 3). However, the N-terminal fragmenting was successful for any fragment lengths as the anchoring residues are close to the N-terminus of H3. In the case of C-terminal fragmenting, the first few fragments often bind weakly (Figure 4, System 4lk9) to the target protein, or in some cases they are completely immersed in the bulk solvent (Figure 4, System 5tdw). At the same time, the last fragments are usually larger peptides (tri- or tetra-peptides) which makes it more difficult to find the binding conformations known from the reference structure, as also explained above.

In some cases, tripeptide (2ke1), and tetra-peptide (2pvc) fragments were also docked at good RMSD values of 2–3 in the first ranks (Table 2). Therefore, tri- and tetra-peptides can also be useful in FBD for other systems. However, dipeptides performed overall better than larger ones for (partial) reconstruction of the original H3 ligand.

2.4. Chemistry of Binding

The investigated H3 peptide sequences are identical in all five Systems (with variable lengths) and unmodified in three Systems (Table 1). In the cases of 3qla and 5tdw, tri-methylation occurs at K4 and K9, respectively. Such PTMs are key elements of the histone code as it was discussed in the Introduction. The investigated target proteins were different in the five Systems. Identity of the H3 ligand sequences offers the possibility of observation and analysis of differences on the interacting target side. The K_4 binding site is negatively charged due to the presence of E (2ke1, 4lk9) and/or D (2ke1, 2pvc, 3qla) residues. K_4 found its position correctly, except for 5tdw, where the binding site is composed of T and F accommodating the tri-methylated K_4 side-chain in the reference crystallographic structure. At the same time, H3 fragment with the same PTM at K_9 docked correctly (Table 2) to its negatively charged pocket with E, Q and also hydrophobic (Y) residues.

In three (2ke1, 2pvc, and 4lk9) cases, the first two dipeptide fragments are the best docked, which is somewhat expected due to the anchoring role of both R2 and K4 residues. In Figure 4, it can be observed that anchoring (see dotted line in Figure 4) of the peptide sequence is achieved by mainly two residues, R_2 and K_4, which can explain the positive docking results obtained especially for the fragments containing these residues (Table 2). More than that, the interaction energies calculated for these residues are also the strongest in the H3 histone tail (see dotted line in Figure 4). In case of 3qla the anchoring was also achieved by the tri-methylated K_9 (Figure 4).

In cases of 2ke1, 2pvc, and 4lk9, consecutive fragments (AR-NHMe, Ac-TK-NHMe) have excellent RMSD $\leq$ 4 Å. Beside the good match with the reference structure, the distance between C_T and N_T atoms (Figure 5), in these consecutive segments fall below 0.75 Å, set as a criterion for selection of linkable ligand binding modes (see Section Linking). Hence, these ligand-binding modes are suitable for covalent re-linking.

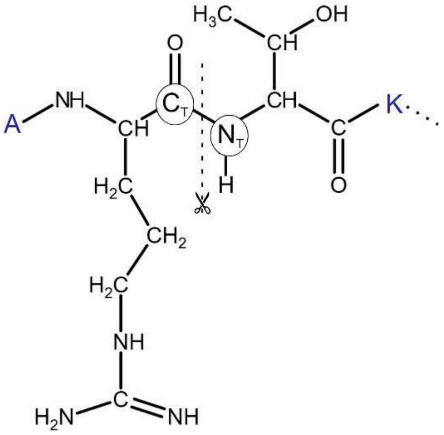


Figure 5. Chemical formula of the N-terminus (ARTK) of histone H3 with a detailed Lewis structure of R_2 and T_3. The peptide was cut into di-peptide fragments between the C_T and N_T atoms of the amide bond.

Notably, the binding modes with the above-mentioned excellent RMSD values (Systems 2ke1, 4lk9, and 5tdw) were produced in the first wrapping cycle, i.e., one hundred blind docking runs (see Methods) were enough to identify them. In the case of System 2pvc, the best binding mode of an RMSD of 3.2 Å was found in the second cycle of wrapping, requiring an additional cycle of one hundred runs. This exemplifies the systematic approach of Wrapper which surely finds a binding mode in higher cycles even if it was not identified in the first one.

2.5. Secondary Structure of the Ligand.

Target-bound H3 histone peptides of the tests systems adopted a variety of secondary structures, such as coil (2pvc, 5tdw), β-sheet (2ke1, 3qla) and α-helix (4lk9). In the case of 4lk9, due to the α-helical secondary structure only the first two dipeptide sequences were found successfully with 2.9 and 4.0 Å RMSD, respectively (N2 series, Table 2). In the second fragment, the Ac-TK-NHMe peptide, is part of the α-helix in the original 4lk9 structure, resulting in the increase of RMSD if compared with that of the first fragment. Similarly, further fragments from the α-helical region could not find the reference conformation below 4 Å RMSD. Fast docking approaches such as AutoDock 4.2 often have difficulties in reproducing helical secondary structures of peptides. This is probably due to missing explicit solvent model and inadequate consideration of intra-backbone H-bridges. Notably, the scoring function of fast docking methods are trained primarily [24] for optimization of intermolecular (target–ligand) interactions and not intramolecular ones. For this reason, in AutoDock 4.2 there is an option for restraining backbone torsions of the ligand [25]. However, as a real test, we aimed at fully flexible blind docking of fragments without using any prior knowledge of their bound conformation. Finding the C-terminal region of the H3 peptide tail was difficult in all of the test cases indifferent of the secondary structure, due to the shallow binding site and weak interaction with the target protein, as reflected by the calculated E_{inter} and N_{inter} values (Figure 4), as well. However, in the case of our 4lk9 test system, where the H3 peptide tail has an α-helix secondary structure when bound to the target protein, the intramolecular hydrogen bonds in the helix made the docking even more challenging than in other test cases.

2.6. Fragment Ends

The fragmenting method of FBD allowed to check the role of terminal end groups (Figure 3) of peptide ligands. The AR sequence appears at two different positions of the H3 peptide (Table 1). Thus, two dipeptide fragments were formed, one with a free, positively charged and another one with a capped N-terminal. Docking results showed that the positive charge is essential to find the reference binding position at an RMSD of 1.9 Å RMSD (due to a hydrogen bond with P331 (Supplementary Table S2) carbonyl oxygen of the autoimmune regulator (AIRE) protein. In the case of the capped (Ac-AR-NHMe) version, this interaction could not occur due to lack of the positive charge. This example hints that for appropriate modelling of terminal fragments, the original (charged) end should be retained. Capping should be used only at the cleaved amide bonds as it was indicated in Section Fragmenting.

2.7. Linking

Having all binding modes of all fragments produced in the previous steps, pairwise linking of fragments of the same length is performed automatically by an algorithm elaborated for FBD and described in the following paragraphs. The algorithm probes all possible pairwise combinations and produces the longest possible peptide from the connected amino acid pairs. The work flow of the algorithm (Figure 6) is described for the example of the N-terminal hexa-peptide ARTKQT of histone H3. Let us suppose that this hexa-peptide was cleaved into three dipeptide fragments AR, TK, and QT, as it was described in Section Fragmenting. For FBD, the dipeptides were blocked with Ac- and/or -NHMe groups at the cleavage sites. The linker algorithm takes the first dipeptide pair of AR-NHMe of n_1 binding modes and Ac-TK-NHMe of n_2 binding modes, and removes the blocking groups. In this way, two series of free radicals are obtained with terminal carbon (C_T) and nitrogen (N_T, Figure 5) atoms available for re-forming the amide bonds. However, not all fragment binding modes are in a correct position to allow the formation of the amide bond. In some cases, the distance between C_T and N_T (d_{CN}) is too large to allow re-formation of a covalent bond. To select the copies with appropriate distances, all (n1 × n2) d_{CN} values are calculated and saved in a matrix (Table 3). The d_{CN} values are generated for all dipeptide fragment pairs, and therefore, two d_{CN} matrices are produced in the case of our hexa-peptide example. There is a user-defined minimal distance tolerance $d_{CN,min}$ = 0.75 Å comparable to the half-length of a C_T–N_T bond in an amide group. A collision is identified between C_T and N_T if the actual $d_{CN} \leq d_{CN,min}$ and the corresponding element of the collision matrix (c_{CN}) is set to zero (otherwise one). A maximal distance tolerance ($d_{CN,max}$) of 6 Å is also defined to exclude fragment pairs too far from each-other, and a trimming matrix t_{CN} is generated based on this value. Notably, $d_{CN,max}$ should not be too large, it must have an meaningful physical value.

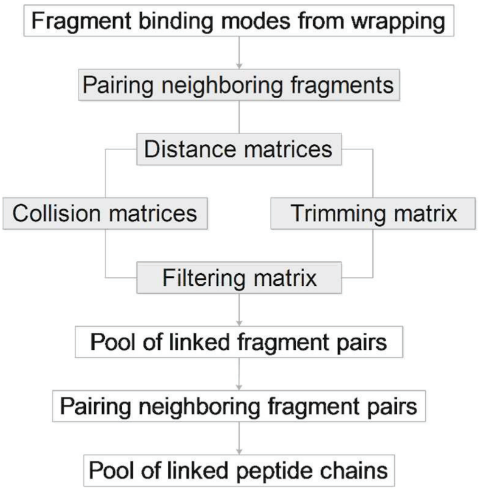

Fragment binding modes from wrapping
Pairing neighboring fragments
Distance matrices
Collision matrices
Trimming matrix
Filtering matrix
Pool of linked fragment pairs
Pairing neighboring fragment pairs
Pool of linked peptide chains

Figure 6. The linking algorithm (grey boxes refer to repeated tasks for all fragments).

Table 3. Matrices used during the linking process.

Symbol	Description
d_{all}	Smallest distance calculated between all heavy atoms
c_{all}	Collision matrix between any heavy atoms
d_{CN}	Distance calculated between C_T and N_T atoms
c_{CN}	Collision matrix from d_{CN}, c_{CN} = 1 if there is no collision, otherwise 0
t_{CN}	Trimming matrix from d_{CN}, t_{CN} = 0 if trimmed, otherwise 1
f_{CN}	Filtering matrix from the collision and trimming matrices, f_{CN} = 1 if c_{CN} = t_{CN} = 1, otherwise 0

To avoid overall collisions between any atoms of the fragment pairs, a distance matrix d_{all} with pair-wise distances (Table 3) between all heavy atoms is also generated. A collision matrix c_{all} is also calculated from d_{all} to identify the steric collisions between heavy atoms of the fragments. A collision is identified if the actual $d_{all} \leq d_{all,min}$, where $d_{all,min}$ is a user-defined minimal distance tolerance, and a $d_{all,min}$ = 0.75 Å, the same value as the above $d_{CN,min}$ works well for peptide ligands. The elements of c_{all} are set to zero by default, and one if there is no collision between a pair of atoms. Finally, a filtering matrix f_{CN} is produced which tells if a fragment pair can be considered for welding and refinement (see next Section). Values of f_{CN} is set to one if t_{CN} and c_{CN} are equal to one otherwise zero.

After sorting the elements of the distance matrix in an increasing order of d_{CN} the main loop selects the first (next), unchecked fragment pair with the actual smallest d_{CN} from the list and checks the corresponding element of the filtering matrix. If the f_{CN} = 0 then the next fragment pair will be checked, otherwise the structure of the actual fragment pair is saved to the candidate pool containing structures of possible fragment pairs in separate directories.

After producing a pool of candidate structures of the first fragment pair of AR and TK in our example, the same procedure is repeated for the next fragment pair of TK and QT if considering our example hexa-peptide ARTKQT. Having all fragment pairs (two pairs in our example) finished, the candidate pools are further processed to link the fragment pairs into triads. Accordingly, the first fragment pair AR-TK with the smallest d_{CN} is selected. If the same TK fragment copy occurs in one of the TK-QT pairs, the structure of the hexa-peptide ARTKQT is produced (Figure 3, AR_2-TK_3-QT_1 or AR_2-TK_3-QT_4, etc.). If not, then the next AR-TK pair will be checked for a common TK with the TK-QT pairs and the algorithm proceeds until all AR-TK pairs are checked. The linker produces structure pools at all levels of the above pairing process. The process works on arbitrary long peptide chains. That is, pools of fragment pairs, triads, and tetrads are produced depending on the length of the actual peptide. A statistics of the pools is written into a report file (Supplementary Table S5).

2.8. Welding and Refinement.

The pools of linked candidate structures are further processed to re-form (weld) the covalent bonds between atoms C_T and N_T between fragments AR and TK in our example (Figure 5). During welding, AR and TK are rotated along the C_α–C_T (angle Ψ) and N_T–C_α (angle Φ) bonds, respectively. Rotations are performed systematically with a step size of 1 degree. One rotation step at angle Ψ is followed by a series of steps of a complete turn-around Φ, up to 360 degrees. After each rotation step, d_{CN} is calculated and stored with the corresponding angles. From among the stored d_{CN} values, the smallest one is selected and the corresponding structure is resulted by welding. The same welding procedure is followed for the remaining fragments of ARTKQTARKS presented in the linked candidate pool.

Following the linking and welding processes, structural refinement of the paired fragments is also recommended using a common molecular mechanics energy-minimization, preferable in explicit water model (Methods). In case of System 2ke1, docking found the first two di-peptide fragments at RMSD values of 1.9 (AR-NHMe) and 2.8 (Ac-TK-NHMe) Å, respectively (Table 2). These two fragments can also be identified as AR04 and TK01 in Supplementary Table S5. These fragments obtained from docking were linked and welded as it was described above and also shown in Figure 7. After refinement of the welded fragment pairs (Figure 7), the d_{CN} of the amide bond changed from 1.5 to 1.3 Å. The optimized structure of ARTK, matches the X-ray structure at a 1.3 Å RMSD. This is a remarkable improvement, considering the above-mentioned RMSD of the di-peptide fragments alone.

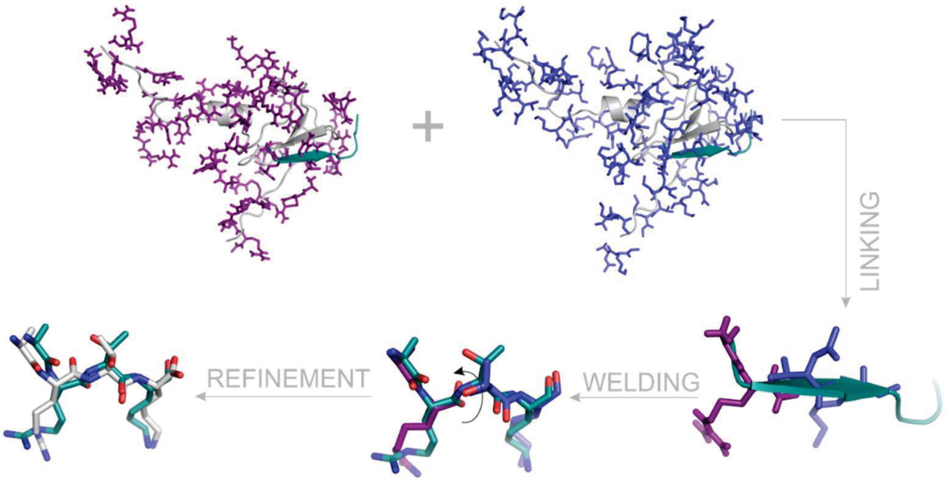


Figure 7. The process of linking, welding and refinement represented on the example of a fragment pair AR-NHMe and Ac-TK-NHMe derived from the histone H3 peptide ligand of System 2ke1. For comparison, the crystallographic ligand conformation is represented in teal cartoon on the first tree images, and with teal sticks on the last two images. The AIRE PHD target is in grey cartoon. Calculated binding modes of AR-NHMe and Ac-TK-NHMe after wrapping of the target are represented with purple and blue sticks, respectively. After linking, binding modes matching with the crystallographic H3 conformation are shown. During welding, the capping groups are removed and the distance between the terminal (C_T and N_T) atoms is minimized via intra-molecular rotation (arrow). Refinement with molecular mechanics minimization helps formation of the proper bound structure of the ARTK fragment (grey sticks).

3. Methods

3.1. Wrapping

3.1.1. Preparation of Targets

Target structures were obtained from the Protein Databank (PDB) entries of the complexes as listed in Table 1. In case of missing atoms of the amino acids, Swiss-PdbViewer was used to complete the structure [26]. Water and ion molecules were removed from the target structure. Prior to docking, energy minimization was carried out on 2pvc, 3qla, 4lk9, 5tdw. A two-step energy minimization was done using Gromacs 5.0.6 [27]. In the first step, a steepest descent was performed, followed by conjugated gradient. The structure optimization was done in AMBER99SB-ILDN force field [28] with TIP3P explicit water model [29]. The target structure was placed in the centre of a cubic box. Distance between the box and the solute atoms was set to 10 Å. The simulation box was filled with water molecules and counter-ions in order to neutralize the total charge of the system. The Particle Mesh Ewald method was used for long-range electrostatics. The van der Waals and Coulomb cut-offs were set to 11 Å. Convergence threshold of the first step (steepest descent) was set to 10^3 kJ mol^{-1} nm^{-2}, in the second step (conjugant gradient) minimization it was set to 10 kJ mol^{-1} nm^{-2}. Position restraints were applied on the heavy atoms with a force constant of 10^3 kJ mol^{-1} nm^{-2} during the energy minimizations.

3.1.2. Preparation of Ligands

The non-modified peptide sequences were built using Tinker program package [30], with the protein, newton and xyzpdb commands. The optimization of the constructed ligand structures was performed using the Amber99 force field [31]. A 10^{-4} kcal/mol gradient was set to the newton program for minimization. Methylated peptide sequences were prepared with Schrödinger Maestro program package [32] by capping of the N- and C-terminal of the fragmented regions, and by adding the hydrogens. The obtained ligand structures, were optimized by Open Babel [33] using a steepest descent optimization, with 10^4 steps. The next step was a conjugate gradient minimization, with a maximum of 10^4 steps, and the convergence threshold was set to 10^{-7} kcal mol^{-1} Å^{-1}. MMFF94 force field [34] was used in both steps. The minimized target and ligand structures were used as inputs for docking, after preparation with AutoDock Tools 1.5.7 [35]. Gasteiger–Marsili [36] partial charges were added for both, the minimized ligand and target atoms as well. All default active torsions are kept for the ligand, but the target is treated rigidly, without active torsions.

3.1.3. Grid maps and Blind Docking Parameters

The grid boxes were generated around the entire protein target with AutoGrid 4.2 [25]. Grid boxes were automatically centred on the target, and grid maps of 200 grid points along all axes, with 0.375 Å spacing were generated. The AutoDock 4.2. [25] program package was used with Lamarckian Genetic Algorithm (LGA), AutoGrid 4.2 was used for calculation of grid maps of the target molecule with pre-calculated energy values. One hundred BD runs were performed, in each cycle, with 20 million maximum number of energy evaluations and the docked ligand structures are collected in a log file. Docking parameters were used as described in the previous study [18].

3.1.4. Wrapping Cycles

For each BD cycle, 100 docked ligand copies were generated [17]. Docked ligand conformations were clustered and ranked based on their intermolecular energy (E_{AD4}), calculated by the AutoDock 4.2 scoring function (1st energy component of estimated free energy of binding in the log file), and closest distance between each heavy atom of the ligand copies (d_{min}). In the initial clustering phase, the 100 docked ligand conformations were sorted according to the $E_{AD4.}$ Ligand conformation of the lowest E_{AD4} from among the 100 docked ligand copies were selected as the representative of Cluster 1. Ligand conformation of the 2nd lowest E_{AD4} was selected as a representative of a new Cluster 2 if d_{min}>drnk, where drnk is a ranking tolerance, a measure of separation of clusters from each other. This wrapping method achieves the coverage of target surface with a monolayer of N ligand copies ending up in a target–ligand$_N$ complex. Ligand copies and interacting target surface elements are excluded from successive BD cycles via assignation of "excluded" atom type as detailed in our previous publication [19]. Further details on structural and physical chemistry of the Wrapper algorithm can also be found in the original publication of Wrap 'n' Shake [19].

After the complete coverage of the target surface, a trimming was performed, where excess ligand copies not interacting with the target were removed after the final cycle and the results were written into a single PDB file. This trimmed PDB file was further used to link the obtained ligand copies of the fragmented segments.

Root mean squared deviation (RMSD) between the calculated (C) and the experimental reference (R) ligand conformations was calculated according to Equation (1).

$$RMSD = \sqrt{\frac{1}{NH_L}\sum_{i=1}^{NH_L}|\boldsymbol{C}_i - \boldsymbol{R}_i|^2} \tag{1}$$

where, NH_L is the number of ligand heavy atoms, R is the space vector of the i^{th} heavy atom of the experimental reference ligand molecule, C is the space vector of the i^{th} heavy atom of the calculated ligand conformation as resulted from docking.

3.2. Analysis of Target–Ligand Interactions

3.2.1. Preparation of the Target–Ligand Complexes

The X-ray structure of the target–ligand complexes were subjected to structure optimization, using the same GROMACS and force field parameters as detailed in Section Wrapping (Preparation of targets). The only exception from the above-mentioned protocol, was the use of position restraints on the heavy atoms with a force constant of 10^4 kJmol^{-1} nm^{-2} during the energy minimizations.

3.2.2. Number of Intermolecular Interactions (N_{inter})

Target residues with a closest atomic distance below 3.5 Å measured from the H3 peptide ligand were collected and counted. Only heavy atoms were considered during the distance calculations. The list of interacting target residues can be found in Supplementary Table S2, and N_{inter} values are presented in Figure 5.

3.2.3. Target–Ligand Intermolecular Interaction Energy (E_{inter})

The energy-minimized target–ligand complexes were also subjected to calculation of intermolecular interaction energies expressed as the sum of Lennard–Jones (LJ) and screened Coulomb potentials [37] (Equation (2)). For both the LJ and Coulomb potentials, Amber99sb-ildn force field parameters were used [28].

$$\begin{gathered}
E_{inter} = \sum_{i,j}^{N_T N_L}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0\ \varepsilon_r r_{ij}}\right] \\
A_{ij} = \varepsilon_{ij} R_{ij}^{12} \\
B_{ij} = 2\varepsilon_{ij} R_{ij}^{6} \\
R_{ij} = R_i + R_j \\
\varepsilon_{ij} = \sqrt{\varepsilon_i\ \varepsilon_j} \\
\varepsilon_r = C + \frac{D}{1+ke^{-\lambda B r_{ij}}} \\
C = \varepsilon_0 - D;\ \varepsilon_0 = 78.4;\ D = -8.5525;\ k = 7.7839;\ \lambda = 0.003627
\end{gathered} \tag{2}$$

where, ε_{ij} is the potential well depth at equilibrium between the i^{th} (ligand) and j^{th} (target) atoms;

ε_0 is the dielectric constant of bulk water at 25 °C; R_{ij} is the inter-nuclear distance at equilibrium between i^{th} (ligand) and j^{th} (target) atoms; q is the partial charge of an atom, used in AMBER99SB-ILDN force field; r_{ij} is the actual distance between the i^{th} (ligand) and j^{th} (target) atoms; N_T is the number of target atoms; N_L is the number of ligand atoms.

3.3. Linking and Welding

Algorithms of linking and welding were scripted in java using JDK version 1.8. into a single code FragmentMerge. The script can be run as described in Supplementary Table S4 and uses a set of PDB files as resulted by Wrapping. There is also an input text file containing the name of the system (for report purposes) and the fragments line by line (Supplementary Table S4). The algorithm uses a class hierarchy. The fragments, the atoms and the bonds in the fragments, the PDB files have own class to represent them. For the rotations of welding, the atoms are stored in a molecule graph in the memory, it helps to calculate the new coordinates during the rotation process. The outputs are saved in a separate directory for pairs, triads, tetrads, pentads, etc. They are PDB files including the linked fragments. A report file after the linking and welding process contains all information about inputs, outputs, parameters and access paths.

The number of the fragment and the name of the input files are listed in REMARK lines and the name of the files refers to the content. The welding algorithm also needs connection information between the atoms. For this, coordinate files are converted by Open Babel [33] into PDB with connectivity lists.

4. Conclusions

In the present study, a new method, FBD, was introduced and tested on the examples of complexes of reader and writer proteins with histone H3 peptide fragments. Heuristic search engines of present fast docking methods cannot handle peptide ligands with numerous internal rotations [12,19]. The large size and flexibility of peptide ligands together with the shallow binding surface of the targeted proteins impose a big challenge on experimental structure determination methods, as well. Moreover, interaction of the C-terminal section of the histone peptides with their targets is often weak and even not visible in the experimentally determined structures. We showed that fragmenting the ligands into small peptides provide reasonable solutions even if the entire protein surface was targeted in blind docking runs. Notably, fragmenting has been used in previous fast docking studies focusing on a known binding pocket. The present study provided the first application of fragmenting in a blind docking context with no restriction of the search space. Thus, even the approximate knowledge of location of binding pocket was not necessary in our successful examples. Despite the above challenges, N-terminal anchoring fragments were correctly positioned and linked using the results of our systematic blind docking search (Wrapper). All-in-all, FBD benefited from the philosophy of its parent methods, fragment and blind docking. Present limitations and mis-docked examples of FBD come from the simplified docking force field and the lack of an explicit water model. However, these limitations can be improved by molecular dynamics simulations in many cases as it was described previously [19]. The systematic approach of FBD will improve the efficiency of structure determination of problematic complexes with large ligands such as histone peptides.

Supplementary Materials: The following supplementary materials are available online at http://www.mdpi.com/1422-0067/20/2/422/s1.

Author Contributions: Conceptualization, M.B. and C.H.; Data curation, M.B., I.H. and N.M.; Formal analysis, N.M.; Funding acquisition, C.H.; Investigation, M.B., I.H, N.M. and C.H.; Methodology, M.B., I.H. and C.H.; Project administration, C.H.; Resources, C.H.; Software, I.H.; Supervision, C.H.; Validation, M.B.; Visualization, M.B.; Writing–original draft, M.B., I.H. and C.H.

Funding: This work was supported by the Hungarian National Research, Development and Innovation Office (K123836). We acknowledge the grant of computer time from CSCS Swiss National Supercomputing Centre, and the Governmental Information Technology Development Agency, Hungary. The University of Pécs is acknowledged for a support by the 17886-4/23018/FEKUTSTRAT excellence grant. C.H.'s work was supported by a grant co-financed by Hungary and the European Union (EFOP-3.6.2-16-2017-00008).

Conflicts of Interest: The authors declare no conflict of interest

References

1. Desjarlais, R.; Tummino, P.J. The Role of Histone-modifying Enzymes and their Complexes in Regulation of Chromatin Biology. *Biochemistry* **2016**, *55*, 1584–1599. [CrossRef] [PubMed]
2. Portela, A.; Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **2010**, *28*, 1057–1068. [CrossRef] [PubMed]
3. Rodenhiser, D.; Mann, M. Epigenetics and human disease: Translating basic biology into clinical applications. *Can. Med. Assoc. J.* **2006**, *174*, 341. [CrossRef] [PubMed]
4. Taverna, S.D.; Li, H.; Ruthenburg, A.J.; Allis, C.D.; Patel, D.J. How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1025–1040. [CrossRef] [PubMed]
5. Catherine, A. Musselman, Marie-Eve Lalonde, Jacques Côté2 and TGK. Readers, Perceiving the epigenetic landscape through histone Catherine. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1218–1227.
6. Strahl, B.D.; Allis, C.D. The language of covalent histone modifications. *Nature* **2000**, *403*, 41–45. [CrossRef] [PubMed]
7. Jenuwein, T.; Allis, C.D. Translating the Histone Code. *Science* **2001**, *293*, 1074–1081. [CrossRef]
8. Davis, A.M.; Teague, S.J.; Kleywegt, G.J. Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design. *Angew. Chem.* **2003**, *42*, 2718–2736. [CrossRef]
9. Blundell, T.L.; Jhoti, H.; Abell, C. High-Throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002**, *1*, 45–54. [CrossRef]
10. Saladin, A.; Rey, J.; Thévenet, P.; Zacharias, M.; Moroy, G.; Tufféry, P. PEP-SiteFinder: A tool for the blind identification of peptide binding sites on protein surfaces. *Nucleic Acids Res.* **2014**, *42*, W221–W226. [CrossRef]
11. Yan, C.; Xu, X.; Zou, X. Fully Blind Docking at the Atomic Level for Protein-Peptide Complex Structure Prediction. *Structure* **2016**, *24*, 1842–1853. [CrossRef] [PubMed]
12. Rentzsch, R.; Renard, B.Y. Docking small peptides remains a great challenge: An assessment using AutoDock Vina. *Brief Bioinform.* **2015**, *16*, 1045–1056. [CrossRef]
13. Joseph-McCarthy, D.; Campbell, A.J.; Kern, G.; Moustakas, D. Fragment-Based Lead Discovery and Design. *J. Chem. Inf. Model.* **2014**, *54*, 693–704. [CrossRef] [PubMed]
14. Liao, J.; Wang, Y.-T.; Lin, C.S. A fragment-based docking simulation for investigating peptide–protein bindings. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10436–10442. [CrossRef] [PubMed]
15. Taylor, R.D.; Jewsbury, P.J.; Essex, J.W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **2002**, *16*, 151–166. [CrossRef]
16. Hetényi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **2002**, *11*, 1729–1737. [CrossRef] [PubMed]
17. Hetényi, C.; Van Der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett.* **2006**, *580*, 1447–1450. [CrossRef]
18. Hetényi, C.; Van Der Spoel, D. Toward prediction of functional protein pockets using blind docking and pocket search algorithms. *Protein Sci.* **2011**, *20*, 880–893. [CrossRef]
19. Bálint, M.; Jeszenői, N.; Horváth, I.; van der Spoel, D.; Hetényi, C. Systematic exploration of multiple drug binding sites. *J. Cheminform.* **2017**, *9*, 65. [CrossRef]
20. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
21. Chignola, F.; Gaetani, M.; Rebane, A.; Mollica, L.; Zucchelli, C.; Spitaleri, A.; Mannella, V.; Peterson, P.; Musco, G. The solution structure of the first PHD finger of autoimmune regulator in complex with non-modified histone H3 tail reveals the antagonistic role of H3R2 methylation. *Nucleic Acids Res.* **2009**, *37*, 2951–2961. [CrossRef] [PubMed]
22. Li, H.; Lu, L.; Chen, R.; Quan, L.; Xia, X.; Lü, Q. PaFlexPepDock: Parallel Ab-initio docking of peptides onto their receptors with full flexibility based on Rosetta. *PLoS ONE* **2014**, *9*, e94769. [CrossRef] [PubMed]
23. Verschueren, E.; Vanhee, P.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Protein-peptide complex prediction through fragment interaction patterns. *Structure* **2013**, *21*, 789–797. [CrossRef] [PubMed]
24. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662. [CrossRef]

25. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *28*, 73–86. [CrossRef] [PubMed]
26. Guex, N.; Peitsch, M.C.; Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* **2009**, *30* (Suppl. 1), 162–173. [CrossRef] [PubMed]
27. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
28. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 1950–1958. [CrossRef]
29. Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236. [CrossRef]
30. Pappu, R.V.; Hart, R.K.; Ponder, J.W. Analysis and Application of Potential Energy Smoothing and Search Methods for Global Optimization. *J. Phys. Chem. B* **1998**, *102*, 9725–9742. [CrossRef]
31. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.J.; Luo, R.; Duan, Y. Development of polarizable models for molecular mechanical calculations. 4. van der waals parametrization. *J. Phys. Chem. B* **2012**, *116*, 7088–7101. [CrossRef] [PubMed]
32. Maestro, Schrödinger LLC: New York, NY, USA, 2017. Available online: https://www.schrodinger.com/maestro (accessed on 17 December 2018).
33. Boyle, N.M.O.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 1–14. [CrossRef] [PubMed]
34. Halgren, T.A. Merck Molecular Force Field, I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519. [CrossRef]
35. Huey, R. GMM AutoDock Tools 1.5.7. Available online: http://mgltools.scripps.edu/downloads (accessed on 17 December 2018).
36. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228. [CrossRef]
37. Mehler, E.L.; Solmajer, T. Electrostatic effects in proteins: Comparison of dielectric and charge models. *Prot. Eng.* **1991**, *4*, 903–910. [CrossRef]

D12

International Journal of *Molecular Sciences*

MDPI

Article

Construction of Histone–Protein Complex Structures by Peptide Growing

Balázs Zoltán Zsidó †, Bayartsetseg Bayarsaikhan †, Rita Börzsei and Csaba Hetényi *

Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti Út 12, 7624 Pécs, Hungary; zsido.balazs@pte.hu (B.Z.Z.); bayartsetseg704@yahoo.com (B.B.); rita.borzsei@aok.pte.hu (R.B.)

* Correspondence: hetenyi.csaba@pte.hu; Tel.: +36-72-536-000 (ext. 38201)

† These authors contributed equally to this work.

Abstract: The structures of histone complexes are master keys to epigenetics. Linear histone peptide tails often bind to shallow pockets of reader proteins via weak interactions, rendering their structure determination challenging. In the present study, a new protocol, PepGrow, is introduced. PepGrow uses docked histone fragments as seeds and grows the full peptide tails in the reader-binding pocket, producing atomic-resolution structures of histone–reader complexes. PepGrow is able to handle the flexibility of histone peptides, and it is demonstrated to be more efficient than linking pre-docked peptide fragments. The new protocol combines the advantages of popular program packages and allows fast generation of solution structures. AutoDock, a force-field-based program, is used to supply the docked peptide fragments used as structural seeds, and the building algorithm of Modeller is adopted and tested as a peptide growing engine. The performance of PepGrow is compared to ten other docking methods, and it is concluded that in situ growing of a ligand from a seed is a viable strategy for the production of complex structures of histone peptides at atomic resolution.

Keywords: docking; histone; peptide; ligand; fragment; growing

check for updates

Citation: Zsidó, B.Z.; Bayarsaikhan, B.; Börzsei, R.; Hetényi, C. Construction of Histone–Protein Complex Structures by Peptide Growing. *Int. J. Mol. Sci.* **2023**, *24*, 13831. https://doi.org/10.3390/ijms241813831

Academic Editor: Wolfgang Sippl

Received: 21 August 2023
Revised: 4 September 2023
Accepted: 5 September 2023
Published: 7 September 2023

1. Introduction

Histones have a diverse interaction profile [1] and play a key role in epigenetic regulation via interactions with the DNA in the chromatin [2,3], as well as various protein partners [4,5]. Readers are important proteins that distinguish between the combinatorial numbers of post-translationally modified histone molecules commonly called as the "histone code" [6]. The atomic-resolution structures of histone–reader complexes are key to understanding the "histone code" and designing new drugs that affect epigenetic regulation [6–8]. The present study is focused on consisting of histone H3 peptides and their reader proteins, which play an important role in the pathophysiology of various autoimmune diseases, intellectual disabilities, cancer development, such as breast cancer, colorectal carcinoma and hematopoietic cancers, autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy, meiotic defects in spermatocytes, breast, prostate and colorectal cancers, and leukemia (Table S1 [9–18]). These pathophysiological involvements render histone reader proteins such as bromodomains [19] and the eleven–nineteen leukemia protein (ENL [20]) attractive targets for drug design purposes.

While knowledge of the structures of histone H3–reader complexes is necessary for understanding the pathomechanism of epigenetic diseases and designing new drugs to act against them, the determination of their atomic-resolution structures can be rather challenging [21]. Experimental difficulties are presented by the creation of well-diffracting and stable crystals in X-ray crystallography [22], the computational processing of noisy images in cryo-electron microscopy [23], and the isotopic labeling of proteins in NMR [24]. Histones are particularly problematic ligands for structural determination, as they have a linear N-terminal tail with a high degree of conformational flexibility [25–27] that sticks out of the nucleosome structure (Figure 1). The protruding N-terminal tails of histones may interact with histone readers (like the proteins in Table S1) or with DNA [28]. Thus, the binding of the N-terminal tail of histone H3 with DNA may compete with the binding of histone N-terminal tails to histone reader proteins [29], which is further supported by the increased accessibility of histone H3 during nucleosome disassembly during transcription [30]. Like all peptides, histones are also extensively hydrated, which further complicates the determination of their interactions [5,31]. Moreover, there are shallow binding pockets on the reader side that result in the histone–reader complexes possessing moderate stability [32,33], with micromolar binding constants (see K_d values in Table S1 for examples). Long peptides such as histone tails are well-known problematic cases for fast computational docking [31,34], due to the inappropriateness of the scoring schemes [35–37] of their binding modes (position, orientation, and conformation) and the lack of explicit water models [38]. The complexes presented in Table S1 are good representatives for investigations of the above structural challenges.

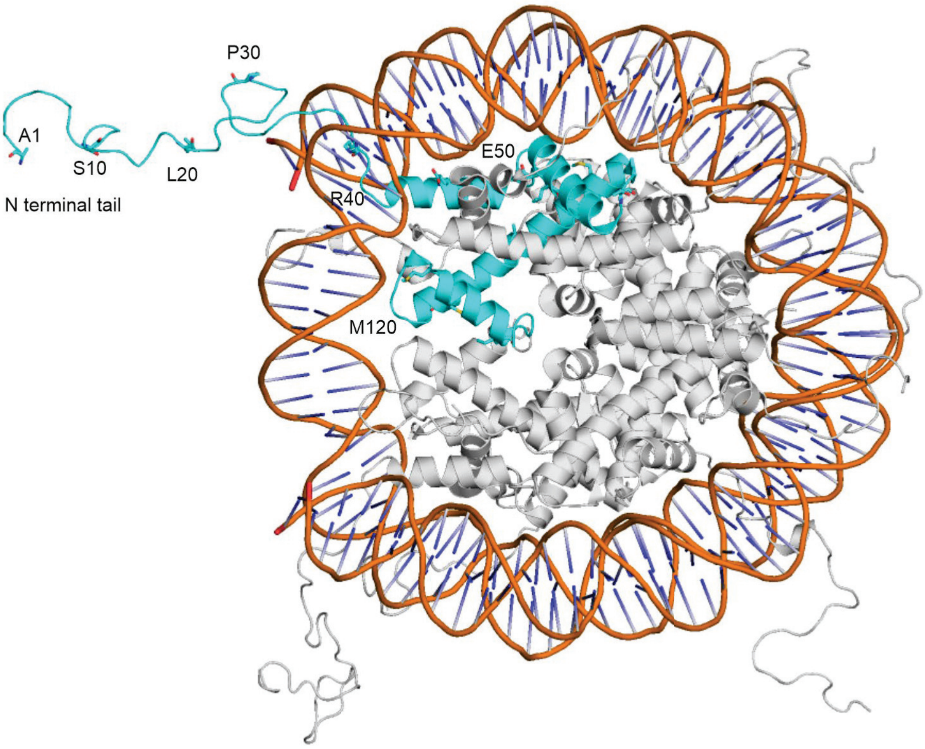


Figure 1. The terminal tails of histone proteins (teal and grey) stick out of the nucleosome core unit and have a flexible structure. The DNA backbone is colored in orange, base pairs are shown as dark blue sticks. A histone H3 protein is highlighted in teal. Every 10th amino acid of the histone H3 (teal) is marked. The figure was prepared from the PDB structure [39] 1kx5 using PyMol v2.0 [40].

The recognition of the above structural and methodological challenges accelerated the development of numerous fast docking methods for peptide ligands. At least three branches can be distinguished among the different methods: physico-chemical approaches, knowledge-based approaches, and their hybrid [41]. Physico-chemical approaches [42–44] calculate energy (scoring) values directly from the atomic positions of the molecules, without conducting further training or experiments. Knowledge-based methods [34,41] are relatively fast and are often restricted by their training set of known structures. Their scores are often based on similarities to the training set [45] and lack physical meaning, which hampers the interpretation of the results (validity problems).

Comprehensive reviews [31,34] and tests [46,47] have shown that the available approaches still have serious limitations with respect to the docking of peptide ligands.

Fragment-based docking is a popular and widely used approach in drug design [48–52], and is based on the linking of docked fragments into the whole bound ligand structure. The number of fragment-based docking methods applied for peptide ligands is still limited. The covalent linking of fragments [48,53] is a critical step in fragment docking, and its success largely depends on the actual steric situation, including the shape-wise matching and the gap between the two docked fragments. Thus, the available methods have multiple limitations, including the lack of full automation, and their dependency on the diversity and selection of linkers and anchoring fragments [54–61]. Inappropriate steric situations of the fragments often necessitate time-consuming follow-up efforts [62] to achieve a new and appropriate covalent bond between two fragments. Further details of the limitations of covalent linking approaches are summarized in Table S2.

In the present study, a new protocol, PepGrow, is introduced and tested for the docking of histone H3 peptide tails to their target reader proteins. PepGrow aims to overcome the limitations of fragment-based docking techniques described above by replacing the fragment linking steps with a growing procedure. Thus, the new protocol is based on the in situ growing of a fragment seed of the peptide ligand in the binding pocket of the reader protein. In drug design, growing steps have been applied for the attachment of small functional groups to ligands [63,64], so as to increase the strength of target–ligand interactions [52]. On the other hand, the growing of a full peptide ligand structure from a small fragment seed is a more difficult task than that handled in the present study. We report the answers to the above challenges, and present a description and validation of PepGrow, comparing its performing with that of of ten other docking methods.

2. Results and Discussion

2.1. Histone Systems and Benchmark Methods

Ten complexes of histone H3 peptides and reader proteins (Table S1) of physiological importance, a complete N-terminal end, and available apo forms of the reader proteins were collected from the Protein Data Bank (PDB [65]) as test systems for the development and evaluation of PepGrow. Due to problems regarding their structural determination (see Section 1), there are relatively few complexes in the PDB with histone ligands of a complete N-terminal end, that is, starting with the first amino acid. Notably, the use of apo target structures allowed a truly unbiased test, excluding any help of the ligand-bound conformation of the pre-formed target-binding sites that may be present in the holo structures.

Histone H3 peptides contain up to ca. 50 rotatable bonds (Table S1), that is a challenge of computational docking.. The challenges are further increased by the unique binding pattern of histones. Reader proteins often have a shallow binding surface, as in the case of the UHRF1 PHD finger (System 3sou, Figure 2A) [66–68]. A considerable part of the linear [69,70] N-terminal region of histone H3 is not able to find anchor points on this shallow target, tending rather to remain unbound in the bulk (Figure 2A). Quantitative analyses of the per-residue interaction energy (E_{inter}; see Section 3) distribution of the experimental holo structures in Table S1 show (Figure 2B) that mostly the first five amino acids of the N-terminal of histone H3 are involved in the interaction with the target, while the C-terminal end is exposed to the bulk, and often has a high degree of conformational freedom, which is also reflected by the large atomic B-factors (red in Figure 2A). This finding also emphasizes that only complexes with a full histone tail (i.e., a complete N-terminal end) are useful as test systems.

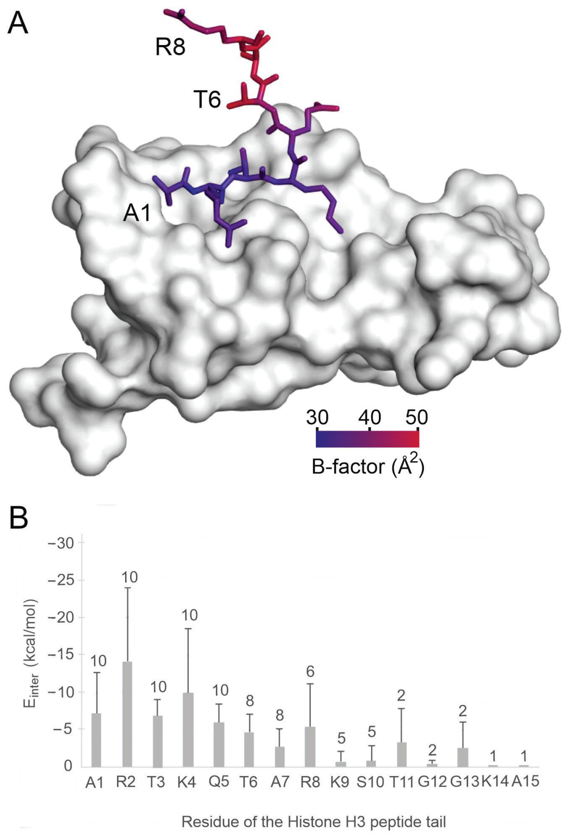


Figure 2. Per-residue energetic and structural analysis of histone H3 peptide ligands bound to their reader proteins. (**A**) The experimental structure of reader UHRF1 PHD finger (grey surface, PDB ID 3sou) in complex with a histone H3 peptide (sticks, colored by Cα B-factors). (**B**) The mean (columns) and standard deviations (error bars) of E_{inter} values for the respective residues calculated for the energy-minimized experimental histone complexes presented in Table S1. The numbers on top of the error bars show the number of systems used for calculation of the averages. The numbers are smaller than the maximum of 10 if histone peptides shorter than 15 amino acids in length were measured experimentally.

Besides the PepGrow protocol, a benchmark set of ten available docking methods (Table S3 [41–45,71–77]) was assembled for the present study. Physico-chemical and hybrid (i.e., incorporating knowledge-based elements into their algorithms; see Section 1) methods were included in the benchmark. The same target and ligand structures were used as inputs for the PepGrow and the benchmark methods.

2.2. The PepGrow Protocol

The PepGrow protocol builds the structures of target–peptide complexes at atomic resolutions (Figure 3) without prior knowledge of the binding site residues of the target. PepGrow starts with the selection of a seed molecule that is a fragment of the ligand peptide. As the ligand used in our cases is the same histone H3 tail (Table S1), the selection of an appropriate seed needs to be carried out only once. For the seed selection procedure, the use of only one holo complex structure (2ke1) proved to be sufficient to pick the best dipeptide fragment from among all of the possible dipeptides (Figure 4A) derived from the H3 peptide (Table S4). In the case of histone H3, Fragment 1 (AR) produced the best results (Figure 4B), and therefore, it was selected as the seed for H3 peptide docking for all complexes except for System 2fuu, for which Fragment 4 (KQ) was used. The selection of Fragments 1 (AR) and 4 (KQ) as seeds is also reflected by the per-residue E_{inter} plot (Figure 2B), where R2 and K4 have the largest E_{inter} contribution among the residues of histone H3. (Thus, the fast, per-residue E_{inter} scoring (Figure 2B) plot of a single strong complex is also applicable for seed selection in PepGrow).

In the next step, the seed was docked on the target protein using a fast method utilizing AutoDock 4.2.6, focusing on the peptide binding area [78], which resulted in several binding modes (where the binding mode refers to the position, orientation, and conformation of a ligand). The binding modes were ranked according to the calculated free energy of their binding and their structural similarity. The representative binding modes were produced for all ranks (see Tables S5 and S6 for a list of the rank counts of all systems). All representative binding modes then proceeded to the fragment growing step, which was accomplished using the builder routine of the homology modeling program Modeller [79]. The experimental target structure with the docked peptide fragment (seed) served as a starting template for growing fragments in the binding pocket. In this way, all docking ranks were used to generate thousands of target–peptide complex models in a matter of minutes, resulting in a large enough pool of peptide binding modes (see Tables S5 and S6 for a list of the binding mode counts of all systems). The complex models of the pool were scored and ranked based on the target–ligand intermolecular interaction energy (E_{inter}, Section 3) values calculated for the full peptide and for the five N-terminal amino acid residues, respectively. The representative peptide structure with coordinates closest to the average coordinates calculated for the peptide structures ranked in the top 1% (Rank 1) according to E_{inter} was selected as the solution. It was observed that in many cases, the top 1% of solution structures contained the best one, but not necessarily the best E_{inter} of all of the structures. Thus, it was reasonable to consider a structure that was representative of the top 1%, rather than a single top structure. Technical details of the PepGrow protocol are provided in the Section 3. Example in- and output files and computational details of the PepGrow protocol are available in the Public Repository files Protocol.pdf and Protocol.tgz (see Data Availability Statement).

2.3. Performance

The structural accuracies of PepGrow and 10 other docking methods are expressed as the root mean square deviation (RMSD; see Section 3) measured between the docked and the experimental (reference) ligand-binding modes. As the experimental complexes mostly show stable (reference) conformations at the first five amino acids of histone H3 (Figure 2), RMSD values were calculated for the full ligand and for the first five amino acids of the N-terminal, respectively. The lowest RMSD of all docked binding modes is referred to as $RMSD_{best}$. The statistics (mean and standard deviation) for the $RMSD_{best}$ values of docking results to the apo targets for all systems in Table S1 are presented in Figure 5. Due to the high mobility (and structural uncertainty) of peptide ligands outside the binding interface, it is common to use only the interfacial (strongly bound core) amino acids [75] for RMSD calculation. In the case of the histone H3 ligand, this core region corresponds to (see Section 2.1) the first five amino acids (full bars in Figure 5). For comparison, the RMSD values measured for all amino acids (empty bars in Figure 5) of the docked histone H3 ligands are also shown. In general, the $RMSD_{best}$ values calculated for the first five amino acids of H3 reflect a much better performance for all methods than the $RMSD_{best}$ values calculated for the full ligand (Figure 5A), due to the natural flexibility of the extended C-terminal region described above (Figure 2).

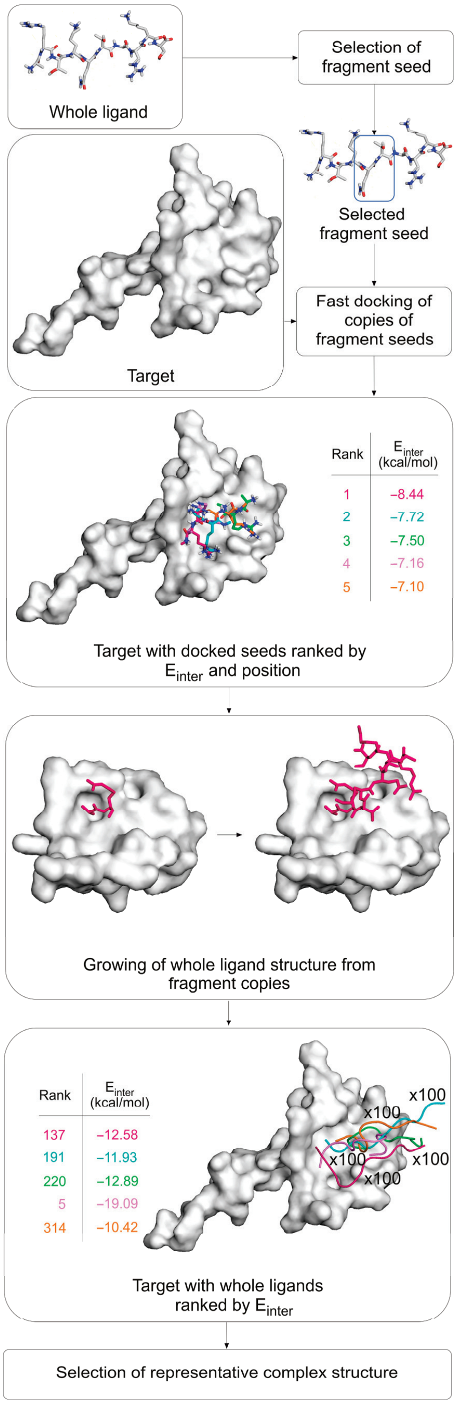

Whole ligand
Selection of fragment seed
Selected fragment seed
Target
Fast docking of copies of fragment seeds
Rank
Einter (kcal/mol)
1 −8.44
2 −7.72
3 −7.50
4 −7.16
5 −7.10
Target with docked seeds ranked by Einter and position
Growing of whole ligand structure from fragment copies
Rank
Einter (kcal/mol)
137 −12.58
191 −11.93
220 −12.89
5 −19.09
314 −10.42
x100
x100
x100
x100
x100
Target with whole ligands ranked by Einter
Selection of representative complex structure

Figure 3. The flow chart of the PepGrow protocol. The different fragment colors correspond to different fragment seed ranks acquired during the fast-docking and seed ranking steps. A close-up of the growth of Rank 1 fragments (purple) only during the growth step is shown for clarity.

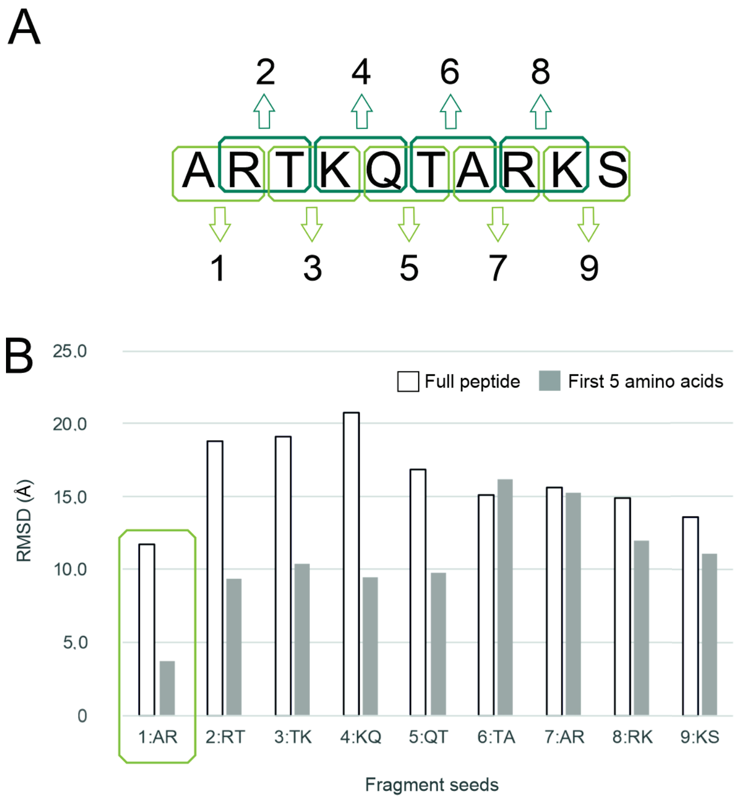


Figure 4. Seed selection. (**A**) All possible (nine) dipeptide fragments were produced from the histone H3 peptide N terminal sequence. Note that Fragment 1 (AR) was capped with an N-methyl group (-NHMe) at the R residue, and Fragment 7 (AR) was capped with an additional acetyl group at the A residue. The capping of the other fragments (2–9) was performed on both ends. (**B**) The PepGrow results for each fragment for System 2ke1. The fragment with the lowest RMSD$_{top}$ is marked with a green frame. See Table S4 for details.

The statistics regarding the apo targets show that PepGrow outperformed all of the other fast docking approaches (Figure 5A), with an RMSD$_{best}$ of 5.36 (±1.47) Å being calculated for all of the amino acids of the docked histone H3 peptide fragments. Furthermore, PepGrow achieved an excellent RMSD$_{best}$ of 4.09 (±1.18) Å, when calculated for the first five amino acids, as well. The per-system analysis of the PepGrow results (Figure 5B) indicates that the best performance was obtained in the case of the target human BAZ2A PHD zinc finger (System 4qf2). Here, the AR-NHMe dipeptide seed was accurately docked (Figure 6), providing a good starting point for ligand growing. The docking of such dipeptides can be accomplished precisely [80] using fast docking techniques. Thus, they provide a good starting point for growing peptide ligands, which is a better alternative than the problematic linking of several, often inadequately docked large-peptide fragments. The accurately docked dipeptide seeds also have the best E$_{inter}$ values (Figure 2), determining the success of PepGrow.

Target flexibility poses a great challenge for docking methods [81]. To check the sensitivity of the investigated docking methods to target conformation, all docking calculations were repeated for the holo structures of the target molecules. As the holo structures have a pre-formed conformation that is ideal for binding to a certain ligand, large differences between the results when docking to the apo and when docking to the holo forms may indicate a high (unwanted) sensitivity to target conformation and moderate robustness of the method. In the case of PepGrow, no significant differences could be detected

(Figure 5 vs. Figure S1) between the results on the apo and holo targets, indicating the robustness of the method.

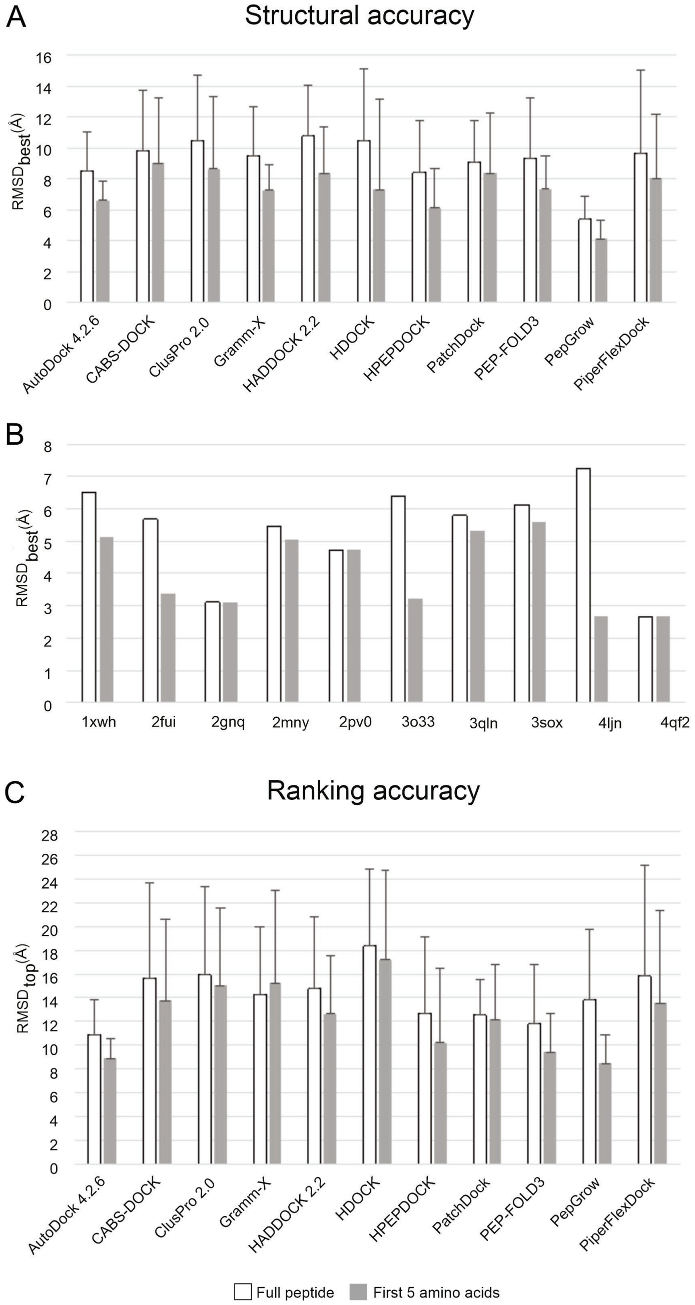


Figure 5. The statistics of docking results obtained for all test systems of Table S1 using all apo target structures. (**A**) Columns represent the mean RMSD$_{best}$ values (of all test systems) calculated for ligand-binding modes supplied by PepGrow and the 10 benchmark methods. Error bars represent standard deviations (see also Table S7a). (**B**) Structural performance of PepGrow on the individual test systems (see also Table S5). (**C**) Columns represent the mean RMSD$_{top}$ values (of all test systems) calculated for ligand-binding modes supplied by PepGrow and the 10 benchmark methods. Error bars represent standard deviations (Table S7a).

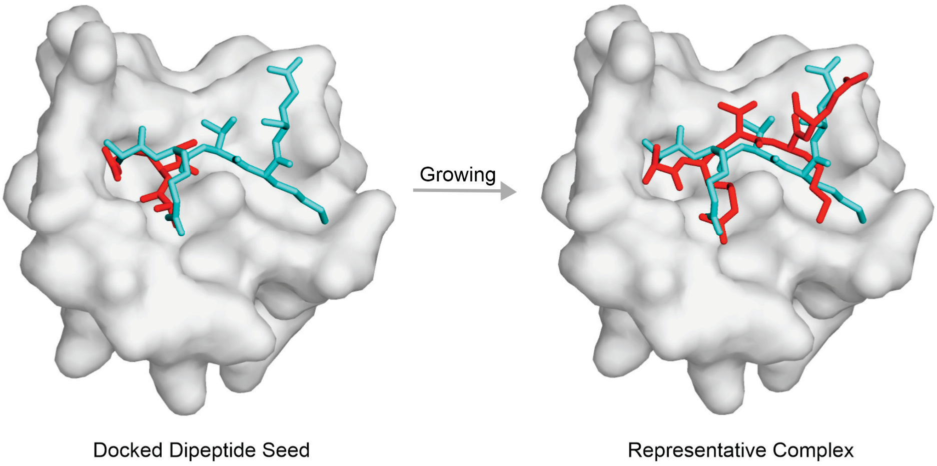


Figure 6. Fragment growing of the fast-docked seed for the complex of the human BAZ2A PHD zinc finger reader (grey surface)–histone H3 peptide (sticks, System 4qf2). The fast-docked seed AR-NHMe of an RMSD of 3.79 Å is shown as red sticks (**left**), representing a good basis of peptide growing. The ligand structure corresponding to an $\mathrm{RMSD_{best}}$ of 2.67 Å is shown as red sticks (**right**), representing the results of the growing. The crystallographic ligand-binding mode is shown as teal sticks for comparison.

The acceptable level of $\mathrm{RMSD_{best}}$ was concluded to be 4.0 ± 3.0 Å on the basis of data (Table S8) collected from publications related to the benchmark methods (Table S3), in which RMSD was calculated only for the peptide backbone. Notably, side-chain atoms were also included in the RMSD calculations in the present study. Thus, the above performance of PepGrow can be considered to be as good as or above average when compared to the RMSD values produced by the benchmark methods (Figure 5).

Besides the structural accuracy of the methods, their ranking performance was also measured on the basis of their respective RMSD values. The docked-ligand-binding modes were ranked by the default scoring functions of the respective methods (Table S3, Supplementary Materials). The RMSD value of the ligand with the best score (representative of the first rank) is referred to as $\mathrm{RMSD_{top}}$. In the case of a method with perfect scoring and ranking, $\mathrm{RMSD_{top}}$ is equal to $\mathrm{RMSD_{best}}$ per definitionem. Unfortunately, such an ideal situation was not observed with the methods investigated, as $\mathrm{RMSD_{top}}$ considerably exceeded $\mathrm{RMSD_{best}}$ in all dockings to the apo targets (Figure 5C), and the same trend was observed in the cases of holo forms (Figure S1). A comparison of the ranking performance of all of the methods (Figure 5C) shows that PepGrow achieved the best results when compared to the benchmark methods. Thus, the $\mathrm{E_{inter}}$-based representative selection method of PepGrow is a viable ranking alternative. Notably, the separate components of $\mathrm{E_{inter}}$ (Lennard-Jones and Coulomb terms, respectively) showed a drop in performance (Table S9), and therefore, $\mathrm{E_{inter}}$ including both terms (see Section 3) was used in the ranking throughout the present study.

The above results indicate that the structural (Figure 5A) and ranking (Figure 5C) performances of PepGrow are better than/comparable to those of the 10 benchmark methods presented in Table S3. PepGrow can also be considered a physico-chemical method, with energy-based scoring and ranking of the ligand-binding mode (Section 2). In theory, physico-chemical methods are generally applicable for any ligand type with appropriate molecular mechanics parametrization. The efficient sampling of the conformational space of flexible peptide ligands [82] like histone H3 tails is a common problem for all fast docking methods. Knowledge-based and hybrid methods (Table S3) attempt to solve this problem using a training set of experimentally determined structures as templates for achieving the correct bound ligand conformation. However, their performance is limited by the availability and reliability of templates for use in training.

In addition to the above sampling problem, the scoring functions of fast docking methods (Table S3) tend to maximize the interactions of the entire ligand with the target, and therefore cannot handle non-interacting parts (see Section 1). Fragment docking methods may provide a solution for this scoring problem by docking only short fragments instead of the entire ligand. This may be a divide-and-conquer strategy for addressing the limitation of linking fragments (see Section 1). For example, PIPER-FlexPepDock is a fragment-based, hybrid approach in which an ensemble of short peptide fragments is collected from experimentally determined structures with a high degree of sequence and (predicted) secondary structure similarity to the actual ligand. However, such methods are also limited by the lack of structures of peptide fragments of large size and/or unusual conformations. Similar to PIPER-FlexPepDock, PepGrow utilizes the potential of physico-chemical methods to accurately dock small peptide fragments, but instead of all possible fragments in the peptide, it focuses on the anchoring fragment of a good E_{inter} (see Section 2, Figure 2) and grows the remaining part of the peptide in situ in the binding pocket. Thus, PepGrow addresses both the sampling and scoring (ranking) problems via its fragment docking strategy and the focused growing of a ligand from the docked seed the strongest interaction with the target.

Data files of the performance tests of PepGrow and the benchmark methods are available in the Public Repository files PepGrow.tgz and Benchmark.tgz (see Data Availability Statement).

3. Materials and Methods

3.1. Selection of Test Systems and Benchmark Methods

All atomic coordinates of the targets were acquired from the PDB. Apart from their physiological relevance, histone-target systems were preferentially selected that exhibited high resolution (<4 Å) and the availability of a non-covalently bound histone H3 N-terminal peptide tail, starting from the first amino acid (A). The availability of both complexed (holo) and apo forms was a selection criterion, as well. For the benchmark methods, fast docking engines were selected that were designed to model interactions in protein–peptide or macromolecular complexes (except AutoDock) and had previously been evaluated on protein–peptide complexes. A further selection criterion was their free availability for academic purposes via web servers or as standalone programs. The investigated docking engines can be roughly sorted into knowledge-based, physico-chemical, and hybrid categories (Table S3).

3.2. Performance Metrics

Both structural and ranking performance are expressed in terms of root mean square deviation (RMSD), a commonly used measure for the comparison of the conformational match of two molecules. In the present study, the bound conformation of a peptide ligand produced by PepGrow (P) was compared to the bound conformation of the same ligand in the experimental complex (E) structure used as a reference (Equation (1)).

$$RMSD = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |\boldsymbol{P}_n - \boldsymbol{E}_n|^2} \quad (1)$$

N is the number of ligand heavy atoms, $\boldsymbol{E}$ is the space vector of the nth heavy atom of the experimental reference ligand molecule, and $\boldsymbol{P}$ is the space vector of the nth heavy atom of the PepGrow-calculated ligand conformation. Crystallographic structures were mostly used as references (Table S1). In 3 cases, NMR structures were also employed, where the first model was selected as a reference. RMSD values were calculated after superimposition of the target parts (Table S10).

3.3. Application of Benchmark Methods

The general and specific settings, and the preparation of targets and ligands are detailed for all benchmark methods in the Supplementary Materials Methods [83–89].

3.4. PepGrow

Target preparation. The atomic coordinate structure files for the selected target protein (Table S1) were downloaded from the PDB. All non-protein parts (ligands, waters, etc.) were removed from all selected target structures prior to docking. If the structure was a homo-oligomer, then only one selected chain was used (the first protein chain in the PDB file). The rest of the target molecule was equipped with polar hydrogen atoms and Gasteiger–Marsilli [90] partial charges in AutoDock Tools [44].

Ligand preparation. An initial fragmenting step was used to create dipeptide-sized fragments of the original histone H3 peptide. The fragments were built using the Tinker program package [91] with the protein, newton and xyzpdb commands. The cut was made between the carbon and nitrogen atoms of the amide bond, acetyl (Ac-) and N-methyl (-NHMe) groups were used to block the N- and C-terminal cut ends (the 1:AR fragment was not capped at the N-terminal end, but the 7:AR fragment was capped at both ends). These blocking groups were added in Tinker [91]. The acquired ligand structures were then energy minimized using Open Babel [92] with the Amber99 force field [93] using the steepest descent optimization with 10^4 steps; the convergence threshold was set to 10^3 kJ mol^{-1} nm^{-1}. The next step was conjugate gradient minimization; a maximum of 10^4 steps was used, and the convergence threshold was set to 10 kJ mol^{-1} nm^{-1}. Gasteiger–Marsili charges [90] were added to the fragments with AutoDock Tools [44]

Fragment docking. The fragment docking was performed using AutoDock 4.2.6 [44]. The previously prepared target was handled as a rigid body. All active torsions were allowed on the prepared ligand fragments. All ligand structures were docked to the interacting site defined by the experimental structure, where the docking box was set to a size that would fit the whole peptide inside. The number of grid points was set to 60 × 60 × 60, with a grid spacing of 0.375 Å; the middle of the box was set to the center of the respective experimental full ligand conformation in a manner similar to the procedure used for the benchmark methods. The Lamarckian genetic algorithm was used to perform a global search. Ten docking runs were performed, and the resulting fragment conformations were ranked [94], and representatives of each rank were used in Step 4.

Fragment growing with a homology modelling tool. All docked fragment copies were processed using Modeller 9.22 [79], a homology modeling program. The template structure was the experimental structure of the target protein with the docked (previous step) fragment seed of the ligand peptide. The query sequence was the respective sequence of each system and the histone H3 peptide tail matching the sequence length seen in the corresponding experimental structures (Table S1). The target and ligand sequences were taken from the UniProt database. The alignment between the template structure and the query sequence was manually optimized if necessary to obtain identical regions that correctly matched each other. This was necessary when fitting the sequence of the docked dipeptide seed to the sequence of the whole ligand. The Modeller 9.22 software package was applied to generate 100 models per step, following the final PepGrow protocol. Explicit manual restraints were not added to access additional energy calculation features. During the method development phase of the present work, restraints, energy calculating features, and seed number variation steps were evaluated thoroughly (Table S11). When the rapid generation of 100 models with default building settings was compared with the generation of fewer models (20) with slower refinement, the results were similar, so the faster method (with 100 models) was selected as the main PepGrow protocol step. The robustness of the building procedure was further challenged by changing the random seed number, which did not affect the results (Table S11). For System 2fuu, fragment 4:KQ was selected, due to the special interaction of the trimethylated K4 with the target. In addition, fragment 4:KQ had the second-best performance (after 1:AR) when compared with the other seeds (Figure 4).

Scoring. To extend the use of the method to apo structures with previously unknown N-terminal histone tail ligand positions, it is important to apply a scoring function that is able to select the bound ligand conformation closest to the real structure. The discrete optimized protein energy (DOPE [79,95]), the Modeller probability density function (molpdf [79,95]), and the Lennard-Jones, Coulomb and E_{inter} interaction energy scores (Equation (2)) of each model were calculated. The E_{inter} interaction energy score calculated for the first five amino acids was the basis of the representative model selection (Tables S7a,b and S12). Table S13 details the scoring functions of the benchmark methods; the differences between the physico-chemical, knowledge-based and hybrid methods were determined based on these scoring functions. Notably, the DOPE and molpdf scores were developed on a benchmark set containing only single-chain proteins, according to the User's Manual of Modeller 9.22 [79]; there is no guarantee of their applicability to multi-chain structures. The calculated DOPE and molpdf scores were therefore only used to test the effect of changing the random seed number for model generation during the initial steps of testing Modeller, as these two scores are the default scoring functions of the software (Table S11).

3.5. Calculation of E_{inter} and Energy Analyses

Experimental, Modeller-built, and energy-minimized experimental structures were subjected to per-residue interaction energy scoring. The missing atoms of all crystallographic targets were modeled using SWISS-Model [96]; for a detailed list of the missing atoms and residues, please see the respective pdb structure files. However, these missing atoms did not affect the binding site. The experimental structures were equipped with polar hydrogen atoms and Gasteiger–Marsilli partial charges [90] using Open Babel 2.4.0 [92], and were converted from pdb files to mol2 files. The mol2 files were then subjected to per-residue interaction energy calculation using Equation (2), implemented in an energy calculator program, which is available as a binary version, downloadable as PepGrow.tgz (see Data Availability Statement). Lennard-Jones and Coulomb energies were calculated and summarized to obtain the total E_{inter} for each residue, and the whole ligand according to Equation (2). The Coulomb term was calculated with a distance-dependent dielectric function of Mehler and Solmajer [97] (Equation (3)), and Amber 2012 van der Waals parameters and atom types were used [98].

$$\begin{gathered} E_{inter} = E_{LJ} + E_{Coulomb} = \sum_{i,j}^{N_T N_L} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \right) \\ A_{ij} = \varepsilon_{ij} R_{ij}^{12} \\ B_{ij} = 2\varepsilon_{ij} R_{ij}^{6} \\ R_{ij} = R_i + R_j \\ \varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \end{gathered} \tag{2}$$

where ε_{ij} is the potential well depth at equilibrium between the ith (ligand) and jth (target) atoms; ε_0 is the permittivity of vacuum; ε_r is the distance-dependent relative permittivity (Equation (3)); R_{ij} is the inter-nuclear distance at equilibrium between the ith (ligand) and jth (target) atoms; q is the partial charge of an atom; r_{ij} is the actual distance between the ith (ligand) and jth (target) atoms; N_T is the number of target atoms; N_L is the number of ligand atoms.

$$\varepsilon_r = A + \frac{B}{1 + ke^{-\lambda Br}} \tag{3}$$

where B = $\varepsilon_0 - A$, ε_0 is the dielectric constant of water at 25 °C, and A, λ and k are constants [97].

4. Conclusions

Although fast docking methods have proven successful in the design of small-molecule ligands [99,100], they face persistent challenges [99–102]. While long peptides are often used as templates for the development of new drugs [103–105], they are especially challenging ligands due to their high degree of flexibility and hydration, which cannot properly be handled by fast docking methods. In the present study, a popular fast docking method, AutoDock 4.2.6, and the fast model building function of the widely used program Modeller were combined into a new protocol PepGrow.

A comparison of the results with those obtained using ten other benchmark methods showed that PepGrow offers a real alternative for the construction of histone complexes. The relatively good performance of PepGrow is based on at least two key components of the algorithm. Firstly, the docking of very short and strongly interacting (di)peptide seeds can be reliably achieved [80] using currently available fast docking methods like AutoDock 4.2.6 (unlike large peptide ligands, where fast docking presents problems [31]). Secondly, instead of the problematic linking step of all fragments of the ligand, a robust ligand growing step is implemented.

PepGrow constructs the complex structures of histone H3 peptides of various lengths with various targets. While the number of such complexes is expected to be very high (histone code), only a small number of structures have been determined. Thus, PepGrow can help to accelerate the structural exploration of the histone code, as well as the prediction of the outcome of the reader–DNA binding competition mentioned in the Introduction. The disordered nature of histone peptides presented a real challenge for all eleven methods compared. The structural performance of PepGrow was better than that of the other methods, the ranking of such large ligands still remains [34,37] a challenging task for all methods. Our results also indicate that physico-chemical scores like E_{inter} are a necessary component of the ranking and selection of representative structures. The histone complexes selected for the present work can be recommended as a particularly challenging test set for future method development studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms241813831/s1.

Author Contributions: Conceptualization, C.H.; writing—original draft preparation, B.Z.Z., B.B., R.B. and C.H.; writing—review and editing, B.Z.Z., B.B., R.B. and C.H.; visualization, B.Z.Z., B.B., R.B. and C.H.; supervision, C.H.; project administration, B.Z.Z. and C.H.; funding acquisition, B.Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Supported by the PTE ÁOK KA-2022-26 grant. The project was supported by the European Union, co-financed by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data files are available at https://zenodo.org/record/8251745 (access date: 20 August 2023). A compressed data file PepGrow.tgz contains the in- and output files of PepGrow for both the holo and apo systems, the scripts and the programs necessary to produce them, and a Table of Contents file with detailed folder descriptions. A portable document file Protocol.pdf contains detailed description how to perform the PepGrow protocol. A compressed data file Protocol.tgz contains example folders of PepGrow on a system with in- and output files, scripts and the programs necessary to produce them, and a Table of Contents file with detailed folder descriptions. A compressed data file Benchmark.tgz contains the in- and output files of the 10 benchmark methods for both holo and apo systems, and a Table of Contents file with detailed folder descriptions.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research were achieved using the DECI resource Archer2, based in the UK at the National Supercomputing Service, with support from the PRACE aisbl.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shvedunova, M.; Akhtar, A. Modulation of Cellular Processes by Histone and Non-Histone Protein Acetylation. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 329–349. [CrossRef]
2. Enetics, E.P.I.G.; Gamblin, S.J.; Wilson, J.O.N.R. A Key to Unlock Chromatin. *Nature* **2019**, *573*, 355–356.
3. Izzo, L.T.; Wellen, K.E. Histone Lactylation Links Metabolism and Gene Regulation. *Nature* **2019**, *574*, 492–493. [CrossRef] [PubMed]
4. Org, T.; Chignola, F.; Hetényi, C.; Gaetani, M.; Rebane, A.; Liiv, I.; Maran, U.; Mollica, L.; Bottomley, M.J.; Musco, G.; et al. The Autoimmune Regulator PHD Finger Binds to Non-Methylated Histone H3K4 to Activate Gene Expression. *EMBO Rep.* **2008**, *9*, 370–376. [CrossRef] [PubMed]
5. Zsidó, B.Z.; Hetényi, C. Molecular Structure, Binding Affinity, and Biological Activity in the Epigenome. *Int. J. Mol. Sci.* **2020**, *21*, 4134. [CrossRef]
6. Strahl, B.D.; Allis, C.D. The Language of Covalent Histone Modifications. *Nature* **2000**, *403*, 41–45. [CrossRef] [PubMed]
7. Musselman, C.A.; Lalonde, M.E.; Côté, J.; Kutateladze, T.G. Perceiving the Epigenetic Landscape through Histone Readers. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1218–1227. [CrossRef]
8. Arrowsmith, C.H.; Bountra, C.; Fish, P.V.; Lee, K.; Schapira, M. Epigenetic Protein Families: A New Frontier for Drug Discovery. *Nat. Rev. Drug Discov.* **2012**, *11*, 384–400. [CrossRef]
9. Bortoluzzi, A.; Amato, A.; Lucas, X.; Blank, M.; Ciulli, A. Structural Basis of Molecular Recognition of Helical Histone H3 Tail by PHD Finger Domains. *Biochem. J.* **2017**, *474*, 1633–1651. [CrossRef]
10. Ruthenburg, A.J.; Wang, W.; Graybosch, D.M.; Li, H.; Allis, C.D.; Patel, D.J.; Verdine, G.L. Histone H3 Recognition and Presentation by the WDR5 Module of the MLL1 Complex. *Nat. Struct. Mol. Biol.* **2006**, *13*, 704–712. [CrossRef]
11. Ooi, S.K.T.; Qiu, C.; Bernstein, E.; Li, K.; Jia, D.; Yang, Z.; Erdjument-Bromage, H.; Tempst, P.; Lin, S.P.; Allis, C.D.; et al. DNMT3L Connects Unmethylated Lysine 4 of Histone H3 to de Novo Methylation of DNA. *Nature* **2007**, *448*, 714–717. [CrossRef] [PubMed]
12. Iwase, S.; Xiang, B.; Ghosh, S.; Ren, T.; Lewis, P.W.; Cochrane, J.C.; Allis, C.D.; Picketts, D.J.; Patel, D.J.; Li, H.; et al. ATRX ADD Domain Links an Atypical Histone Methylation Recognition Mechanism to Human Mental-Retardation Syndrome. *Nat. Struct. Mol. Biol.* **2011**, *18*, 769–776. [CrossRef] [PubMed]
13. Rajakumara, E.; Wang, Z.; Ma, H.; Hu, L.; Chen, H.; Lin, Y.; Guo, R.; Wu, F.; Li, H.; Lan, F.; et al. PHD Finger Recognition of Unmodified Histone H3R2 Links UHRF1 to Regulation of Euchromatic Gene Expression. *Mol. Cell* **2011**, *43*, 275–284. [CrossRef] [PubMed]
14. Tsai, W.W.; Wang, Z.; Yiu, T.T.; Akdemir, K.C.; Xia, W.; Winter, S.; Tsai, C.Y.; Shi, X.; Schwarzer, D.; Plunkett, W.; et al. TRIM24 Links a Non-Canonical Histone Signature to Breast Cancer. *Nature* **2010**, *468*, 927–932. [CrossRef]
15. Chignola, F.; Gaetani, M.; Rebane, A.; Org, T.; Mollica, L.; Zucchelli, C.; Spitaleri, A.; Mannella, V.; Peterson, P.; Musco, G. The Solution Structure of the First PHD Finger of Autoimmune Regulator in Complex with Non-Modified Histone H3 Tail Reveals the Antagonistic Role of H3R2 Methylation. *Nucleic Acids Res.* **2009**, *37*, 2951–2961. [CrossRef]
16. Zhang, Y.; Yang, H.; Guo, X.; Rong, N.; Song, Y.; Xu, Y.; Lan, W.; Zhang, X.; Liu, M.; Xu, Y.; et al. The PHD1 Finger of KDM5B Recognizes Unmodified H3K4 during the Demethylation of Histone H3K4me2/3 by KDM5B. *Protein Cell* **2014**, *5*, 837–850. [CrossRef] [PubMed]
17. Li, H.; Ilin, S.; Wang, W.; Duncan, E.M.; Wysocka, J.; Allis, C.D.; Patel, D.J. Molecular Basis for Site-Specific Read-out of Histone H3K4me3 by the BPTF PHD Finger of NURF. *Nature* **2006**, *442*, 91–95. [CrossRef] [PubMed]
18. Dreveny, I.; Deeves, S.E.; Fulton, J.; Yue, B.; Messmer, M.; Bhattacharya, A.; Collins, H.M.; Heery, D.M. The Double PHD Finger Domain of MOZ/MYST3 Induces α-Helical Structure of the Histone H3 Tail to Facilitate Acetylation and Methylation Sampling and Modification. *Nucleic Acids Res.* **2014**, *42*, 822–835. [CrossRef]
19. Sanchez, R.; Meslamani, J.; Zhou, M.-M. The Bromodomain: From Epigenome Reader to Druggable Target. *Biochim. Biophys. Acta BBA Gene Regul. Mech.* **2014**, *1839*, 676–685. [CrossRef]
20. Li, X.; Yao, Y.; Wu, F.; Song, Y. A Proteolysis-Targeting Chimera Molecule Selectively Degrades ENL and Inhibits Malignant Gene Expression and Tumor Growth. *J. Hematol. Oncol.* **2022**, *15*, 41. [CrossRef]
21. Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* **2013**, *10*, 47–53. [CrossRef]
22. Srivastava, A.; Nagai, T.; Srivastava, A.; Miyashita, O.; Tama, F. Role of Computational Methods in Going beyond X-Ray Crystallography to Explore Protein Structure and Dynamics. *Int. J. Mol. Sci.* **2018**, *19*, 3401. [CrossRef]
23. Frank, J. Electron Microscopy Applied to Molecular Machines. *Biopolymers* **2013**, *99*, 832–836. [CrossRef]
24. Verardi, R.; Traaseth, N.J.; Masterson, L.R.; Vostrikov, V.V.; Veglia, G. *Isotope Labeling in Biomolecular NMR.*; Springer: Dordrecht, The Netherlands, 2012; Volume 992, ISBN 978-94-007-4953-5.
25. Antunes, D.A.; Devaurs, D.; Kavraki, L.E. Understanding the Challenges of Protein Flexibility in Drug Design. *Expert Opin. Drug Discov.* **2015**, *10*, 1301–1313. [CrossRef]
26. Du, X.; Li, Y.; Xia, Y.L.; Ai, S.M.; Liang, J.; Sang, P.; Ji, X.L.; Liu, S.Q. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **2016**, *17*, 144. [CrossRef]
27. Hauser, A.S.; Windshügel, B. LEADS-PEP: A Benchmark Data Set for Assessment of Peptide Docking Performance. *J. Chem. Inf. Model.* **2016**, *56*, 188–200. [CrossRef] [PubMed]

28. Lehmann, K.; Felekyan, S.; Kühnemuth, R.; Dimura, M.; Tóth, K.; Seidel, C.A.M.; Langowski, J. Dynamics of the Nucleosomal Histone H3 N-Terminal Tail Revealed by High Precision Single-Molecule FRET. *Nucleic Acids Res.* **2020**, *48*, 1551–1571. [CrossRef] [PubMed]
29. Morrison, E.A.; Bowerman, S.; Sylvers, K.L.; Wereszczynski, J.; Musselman, C.A. The Conformation of the Histone H3 Tail Inhibits Association of the BPTF PHD Finger with the Nucleosome. *eLife* **2018**, *7*, e31481. [CrossRef]
30. Morrison, E.A.; Baweja, L.; Poirier, M.G.; Wereszczynski, J.; Musselman, C.A. Nucleosome Composition Regulates the Histone H3 Tail Conformational Ensemble and Accessibility. *Nucleic Acids Res.* **2021**, *49*, 4750–4767. [CrossRef] [PubMed]
31. Rentzsch, R.; Renard, B.Y. Docking Small Peptides Remains a Great Challenge: An Assessment Using AutoDock Vina. *Brief. Bioinform.* **2015**, *16*, 1045–1056. [CrossRef] [PubMed]
32. Peach, C.J.; Mignone, V.W.; Arruda, M.A.; Alcobia, D.C.; Hill, S.J.; Kilpatrick, L.E.; Woolard, J. Molecular Pharmacology of VEGF-A Isoforms: Binding and Signalling at VEGFR2. *Int. J. Mol. Sci.* **2018**, *19*, 1264. [CrossRef]
33. Weaver, T.M.; Morrison, E.A.; Musselman, C.A. Reading More than Histones: The Prevalence of Nucleic Acid Binding among Reader Domains. *Molecules* **2018**, *23*, 2614. [CrossRef] [PubMed]
34. Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–Peptide Docking: Opportunities and Challenges. *Drug Discov. Today* **2018**, *23*, 1530–1537. [CrossRef] [PubMed]
35. Lee, A.C.L.; Harris, J.L.; Khanna, K.K.; Hong, J.H. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *Int. J. Mol. Sci.* **2019**, *20*, 2383. [CrossRef] [PubMed]
36. Peterson, L.X.; Roy, A.; Christoffer, C.; Terashi, G.; Kihara, D. Modeling Disordered Protein Interactions from Biophysical Principles. *PLoS Comput. Biol.* **2017**, *13*, e1005485. [CrossRef]
37. Xiong, G.-L.; Ye, W.-L.; Shen, C.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Improving Structure-Based Virtual Screening Performance via Learning from Scoring Function Components. *Brief. Bioinform.* **2021**, *22*, bbaa094. [CrossRef]
38. Zsidó, B.Z.; Hetényi, C. The Role of Water in Ligand Binding. *Curr. Opin. Struct. Biol.* **2021**, *67*, 1–8. [CrossRef]
39. Richmond, T.J.; Davey, C.A. The Structure of DNA in the Nucleosome Core. *Nature* **2003**, *423*, 145–150. [CrossRef]
40. DeLano, W.L. *The PyMOL Molecular Graphics System, Version 2.0*; Schrödinger, LLC.: New York, NY, USA, 2021.
41. Dominguez, C.; Boelens, R.; Bonvin, A.M.J.J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737. [CrossRef]
42. Alam, N.; Goldstein, O.; Xia, B.; Porter, K.A.; Kozakov, D.; Schueler-Furman, O. High-Resolution Global Peptide-Protein Docking Using Fragments-Based PIPER-FlexPepDock. *PLoS Comput. Biol.* **2017**, *13*, e1005905. [CrossRef]
43. Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-Dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Nucleic Acids Res.* **2015**, *43*, W419–W424. [CrossRef]
44. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef] [PubMed]
45. Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P. PEP-FOLD3: Faster de Novo Structure Prediction for Linear Peptides in Solution and in Complex. *Nucleic Acids Res.* **2016**, *44*, W449–W454. [CrossRef]
46. Castro-Alvarez, A.; Costa, A.M.; Vilarrasa, J. The Performance of Several Docking Programs at Reproducing Protein-Macrolide-like Crystal Structures. *Molecules* **2017**, *22*, 136. [CrossRef] [PubMed]
47. Hetényi, C.; Körtvélyesi, T.; Penke, B. Mapping of Possible Binding Sequences of Two Beta-Sheet Breaker Peptides on Beta Amyloid Peptide of Alzheimer's Disease. *Bioorg. Med. Chem.* **2002**, *10*, 1587–1593. [CrossRef] [PubMed]
48. Bian, Y.; Xie, X.Q. Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *AAPS J.* **2019**, *20*, 59. [CrossRef]
49. Evans, D.J.; Yovanno, R.A.; Rahman, S.; Cao, D.W.; Beckett, M.Q.; Patel, M.H.; Bandak, A.F.; Lau, A.Y. Finding Druggable Sites in Proteins Using TACTICS. *J. Chem. Inf. Model.* **2021**, *61*, 2897–2910. [CrossRef]
50. Aguayo-Ortiz, R.; Dominguez, L. Unveiling the Possible Oryzalin-Binding Site in the α-Tubulin of *Toxoplasma Gondii*. *ACS Omega* **2022**, *7*, 18434–18442. [CrossRef]
51. Aguayo-Ortiz, R.; Guzmán-Ocampo, D.C.; Dominguez, L. Insights into the Binding of Morin to Human ΓD-Crystallin. *Biophys. Chem.* **2022**, *282*, 106750. [CrossRef]
52. Pires, D.E.V.; Portelli, S.; Rezende, P.M.; Veloso, W.N.P.; Xavier, J.S.; Karmakar, M.; Myung, Y.; Linhares, J.P.V.; Rodrigues, C.H.M.; Silk, M.; et al. A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods. *Methods Mol. Biol.* **2020**, *2112*, 91–106.
53. Lamoree, B.; Hubbard, R.E. Current Perspectives in Fragment-Based Lead Discovery (FBLD). *Essays Biochem.* **2017**, *61*, 453–464. [CrossRef]
54. de Beauchene, I.C.; de Vries, S.J.; Zacharias, M. Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS Comput. Biol.* **2016**, *12*, e1004697. [CrossRef]
55. Liao, J.M.; Wang, Y.T.; Lin, C.L.S. A Fragment-Based Docking Simulation for Investigating Peptide-Protein Bindings. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10436–10442. [CrossRef] [PubMed]
56. Budin, N.; Majeux, N.; Caflisch, A. Fragment-Based Flexible Ligand Docking by Evolutionary Optimization. *Biol. Chem.* **2001**, *382*, 1365–1372. [CrossRef]
57. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S.B.; Johnson, A.P. EHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graph. Model.* **2007**, *26*, 198–212. [CrossRef]

58. Thompson, D.C.; Denny, R.A.; Nilakantan, R.; Humblet, C.; Joseph-McCarthy, D.; Feyfant, E. CONFIRM: Connecting Fragments Found in Receptor Molecules. *J. Comput. Aided Mol. Des.* **2008**, *22*, 761–772. [CrossRef]
59. Samsonov, S.A.; Zacharias, M.; de Chauvot Beauchene, I. Modeling Large Protein–Glycosaminoglycan Complexes Using a Fragment-Based Approach. *J. Comput. Chem.* **2019**, *40*, 1429–1439. [CrossRef] [PubMed]
60. Cross, S.S.J. Improved FlexX Docking Using FlexS-Determined Base Fragment Placement. *J. Chem. Inf. Model.* **2005**, *45*, 993–1001. [CrossRef] [PubMed]
61. Bálint, M.; Horváth, I.; Mészáros, N.; Hetényi, C. Towards Unraveling the Histone Code by Fragment Blind Docking. *Int. J. Mol. Sci.* **2019**, *20*, 422. [CrossRef]
62. Hoffer, L.; Muller, C.; Roche, P.; Morelli, X. Chemistry-Driven Hit-to-Lead Optimization Guided by Structure-Based Approaches. *Mol. Inform.* **2018**, *37*, e1800059. [CrossRef]
63. Yuan, Y.; Pei, J.; Lai, L. LigBuilder V3: A Multi-Target de Novo Drug Design Approach. *Front. Chem.* **2020**, *8*, 142. [CrossRef] [PubMed]
64. Perez, C.; Soler, D.; Soliva, R.; Guallar, V. FragPELE: Dynamic Ligand Growing within a Binding Site. A Novel Tool for Hit-To-Lead Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 1728–1736. [CrossRef]
65. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef] [PubMed]
66. Bálint, M.; Jeszenoi, N.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Systematic Exploration of Multiple Drug Binding Sites. *J. Cheminform.* **2017**, *9*, 65. [CrossRef] [PubMed]
67. Jeszenoi, N.; Bálint, M.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Exploration of Interfacial Hydration Networks of Target-Ligand Complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148–158. [CrossRef]
68. Zhao, S.; Yang, M.; Zhou, W.; Zhang, B.; Cheng, Z.; Huang, J.; Zhang, M.; Wang, Z.; Wang, R.; Chen, Z.; et al. Kinetic and High-Throughput Profiling of Epigenetic Interactions by 3D-Carbene Chip-Based Surface Plasmon Resonance Imaging Technology. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E7245–E7254. [CrossRef] [PubMed]
69. Van Roey, K.; Uyar, B.; Weatheritt, R.J.; Dinkel, H.; Seiler, M.; Budd, A.; Gibson, T.J.; Davey, N.E. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **2014**, *114*, 6733–6778. [CrossRef]
70. Davis, A.M.; Teague, S.J.; Kleywegt, G.J. Application and Limitations of X-Ray Crystallographic Data in Structure-Based Ligand and Drug Design. *Angew. Chem. Int. Ed.* **2003**, *42*, 2718–2736. [CrossRef]
71. Kozakov, D.; Hall, D.R.; Xia, B.; Porter, K.A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein–Protein Docking. *Nat. Protoc.* **2017**, *12*, 255–278. [CrossRef]
72. Tovchigrechko, A.; Vakser, I.A. GRAMM-X Public Web Server for Protein-Protein Docking. *Nucleic Acids Res.* **2006**, *34*, W310–W314. [CrossRef]
73. van Zundert, G.C.P.; Rodrigues, J.P.G.L.M.; Trellet, M.; Schmitz, C.; Kastritis, P.L.; Karaca, E.; Melquiond, A.S.J.; van Dijk, M.; de Vries, S.J.; Bonvin, A.M.J.J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720–725. [CrossRef] [PubMed]
74. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H.J. PatchDock and SymmDock: Servers for Rigid and Symmetric Docking. *Nucleic Acids Res.* **2005**, *33*, W363–W367. [CrossRef] [PubMed]
75. Zhou, P.; Jin, B.; Li, H.; Huang, S.-Y. HPEPDOCK: A Web Server for Blind Peptide–Protein Docking Based on a Hierarchical Algorithm. *Nucleic Acids Res.* **2018**, *46*, W443–W450. [CrossRef] [PubMed]
76. Yan, Y.; Zhang, D.; Zhou, P.; Li, B.; Huang, S.-Y. HDOCK: A Web Server for Protein–Protein and Protein–DNA/RNA Docking Based on a Hybrid Strategy. *Nucleic Acids Res.* **2017**, *45*, W365–W373. [CrossRef]
77. Kozakov, D.; Brenke, R.; Comeau, S.R.; Vajda, S. PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials. *Proteins Struct. Funct. Bioinform.* **2006**, *65*, 392–406. [CrossRef] [PubMed]
78. Wang, P.; Wu, R.; Jia, Y.; Tang, P.; Wei, B.; Zhang, Q.; Wang, V.Y.-F.; Yan, R. Inhibition and Structure-Activity Relationship of Dietary Flavones against Three Loop 1-Type Human Gut Microbial β-Glucuronidases. *Int. J. Biol. Macromol.* **2022**, *220*, 1532–1544. [CrossRef]
79. Fiser, A.; Do, R.K.G.; Šali, A. Modeling of Loops in Protein Structures. *Protein Sci.* **2000**, *9*, 1753–1773. [CrossRef]
80. Hetényi, C.; van der Spoel, D. Efficient Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Protein Sci.* **2009**, *11*, 1729–1737. [CrossRef]
81. Basciu, A.; Callea, L.; Motta, S.; Bonvin, A.M.J.J.; Bonati, L.; Vargiu, A.V. No Dance, No Partner! A Tale of Receptor Flexibility in Docking and Virtual Screening. *Annu. Rep. Med. Chem.* **2022**, *59*, 43–97.
82. Li, C.; Sun, J.; Li, L.-W.; Wu, X.; Palade, V. An Effective Swarm Intelligence Optimization Algorithm for Flexible Ligand Docking. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 2672–2684. [CrossRef]
83. The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef] [PubMed]
84. Kolinski, A. Protein Modeling and Structure Prediction with a Reduced Representation. *Acta Biochim. Pol.* **2004**, *51*, 349–371. [CrossRef]
85. Huang, S.-Y.; Zou, X. MDockPP: A Hierarchical Approach for Protein-Protein Docking and Its Application to CAPRI Rounds 15–19. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 3096–3103. [CrossRef] [PubMed]

86. Yan, Y.; Zhang, D.; Huang, S.-Y. Efficient Conformational Ensemble Generation of Protein-Bound Peptides. *J. Cheminform.* **2017**, *9*, 59. [CrossRef]
87. Huang, S.-Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins Struct. Funct. Bioinform.* **2006**, *66*, 399–421. [CrossRef]
88. Gront, D.; Kulp, D.W.; Vernon, R.M.; Strauss, C.E.M.; Baker, D. Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS ONE* **2011**, *6*, e23294. [CrossRef]
89. Raveh, B.; London, N.; Schueler-Furman, O. Sub-Angstrom Modeling of Complexes between Flexible Peptides and Globular Proteins. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 2029–2040. [CrossRef]
90. Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228. [CrossRef]
91. Rackers, J.A.; Wang, Z.; Lu, C.; Laury, M.L.; Lagardère, L.; Schnieders, M.J.; Piquemal, J.-P.; Ren, P.; Ponder, J.W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289. [CrossRef]
92. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef]
93. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef]
94. Hetényi, C.; Van Der Spoel, D. Blind Docking of Drug-Sized Compounds to Proteins with up to a Thousand Residues. *FEBS Lett.* **2006**, *580*, 1447–1450. [CrossRef]
95. Shen, M.; Sali, A. Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci.* **2006**, *15*, 2507–2524. [CrossRef]
96. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. [CrossRef]
97. Mehler, E.L.; Solmajer, T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Eng. Des. Sel.* **1991**, *4*, 903–910. [CrossRef] [PubMed]
98. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations. 4. van Der Waals Parametrization. *J. Phys. Chem. B* **2012**, *116*, 7088–7101. [CrossRef]
99. Ferreira, L.; dos Santos, R.; Oliva, G.; Andricopulo, A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20*, 13384–13421. [CrossRef] [PubMed]
100. Vitoria, M.; Granich, R.; Gilks, C.F.; Gunneberg, C.; Hosseini, M.; Were, W.; Raviglione, M.; De Cock, K.M. The Global Fight Against HIV/AIDS, Tuberculosis, and Malaria. *Am. J. Clin. Pathol.* **2009**, *131*, 844–848. [CrossRef] [PubMed]
101. Torres, P.H.M.; Sodero, A.C.R.; Jofily, P.; Silva, F.P., Jr. Key Topics in Molecular Docking for Drug Design. *Int. J. Mol. Sci.* **2019**, *20*, 4574. [CrossRef]
102. de Ruyck, J.; Brysbaert, G.; Blossey, R.; Lensink, M. Molecular Docking as a Popular Tool in Drug Design, an in Silico Travel. *Adv. Appl. Bioinform. Chem.* **2016**, *9*, 1–11. [CrossRef]
103. Schreiber, G.; Fleishman, S.J. Computational Design of Protein–Protein Interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 903–910. [CrossRef] [PubMed]
104. Grosdidier, S.; Fernandez-Recio, J. Protein-Protein Docking and Hot-Spot Prediction for Drug Discovery. *Curr. Pharm. Des.* **2012**, *18*, 4607–4618. [CrossRef] [PubMed]
105. Bienstock, R.J. Computational Drug Design Targeting Protein-Protein Interactions. *Curr. Pharm. Des.* **2012**, *18*, 1240–1254. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

D13

hetenyi.csaba_83_23

International Journal of
Molecular Sciences

MDPI

Article

Exploration of Somatostatin Binding Mechanism to Somatostatin Receptor Subtype 4

Rita Börzsei [1,2], Balázs Zoltán Zsidó [1,2], Mónika Bálint [1,2], Zsuzsanna Helyes [1,2,3,4], Erika Pintér [1,2,3,4] and Csaba Hetényi [1,2,*]

[1] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, 7624 Pécs, Hungary; rita.borzsei@gmail.com (R.B.); zsido.balazs@pte.hu (B.Z.Z.); monibalint18@gmail.com (M.B.); helyes.zsuzsanna@pte.hu (Z.H.); erika.pinter@aok.pte.hu (E.P.)
[2] János Szentágothai Research Centre & Centre for Neuroscience, University of Pécs, 7624 Pécs, Hungary
[3] Algonist Gmbh, 1030 Vienna, Austria
[4] PharmInVivo Ltd., 7624 Pécs, Hungary
* Correspondence: hetenyi.csaba@pte.hu

Abstract: Somatostatin (also named as growth hormone-inhibiting hormone or somatotropin release-inhibiting factor) is a regulatory peptide important for the proper functioning of the endocrine system, local inflammatory reactions, mood and motor coordination, and behavioral responses to stress. Somatostatin exerts its effects via binding to G-protein-coupled somatostatin receptors of which the fourth subtype (SSTR4) is a particularly important receptor mediating analgesic, anti-inflammatory, and anti-depressant effects without endocrine actions. Thus, SSTR4 agonists are promising drug candidates. Although the knowledge of the atomic resolution-binding modes of SST would be essential for drug development, experimental elucidation of the structures of SSTR4 and its complexes is still awaiting. In the present study, structures of the somatostatin–SSTR4 complex were produced using an unbiased, blind docking approach. Beyond the static structures, the binding mechanism of SST was also elucidated in the explicit water molecular dynamics (MD) calculations, and key binding modes (external, intermediate, and internal) were distinguished. The most important residues on both receptor and SST sides were identified. An energetic comparison of SST binding to SSTR4 and 2 offered a residue-level explanation of receptor subtype selectivity. The calculated structures show good agreement with available experimental results and indicate that somatostatin binding is realized via prerequisite binding modes and an induced fit mechanism. The identified binding modes and the corresponding key residues provide useful information for future drug design targeting SSTR4.

Keywords: pocket; site; peptide; interaction; selectivity; dynamics

check for updates

Citation: Börzsei, R.; Zsidó, B.Z.; Bálint, M.; Helyes, Z.; Pintér, E.; Hetényi, C. Exploration of Somatostatin Binding Mechanism to Somatostatin Receptor Subtype 4. *Int. J. Mol. Sci.* **2022**, *23*, 6878. https://doi.org/10.3390/ijms23136878

Academic Editors: Ian A. Nicholls and Vladimir N. Uversky

Received: 24 May 2022
Accepted: 16 June 2022
Published: 21 June 2022

1. Introduction

Somatostatin is a cyclic neuropeptide, widely expressed in both peripheral and central tissues. SST has two active forms, the 14 amino acid-long (referred to as SST throughout this study), and an *N*-terminally extended isoform of 28 amino acids [1–4]. Both forms are expressed in the same tissue areas, but it is not clear whether the same cells can produce them. SST is internally stabilized by a disulfide bridge between cysteine residues in positions 3 and 14 (Figure 1A).

SST inhibits the release of several endocrine hormones such as growth hormone, prolactin, thyrotropin, gastrin, insulin, secretin, and glucagon [3,5–7], and the local inflammatory reaction at the periphery [8,9]. As a neurotransmitter, SST plays role in many mechanisms centrally, such as pain transmission, mood coordination, and learning and behavioral responses to stress [3,10–13]. It has emerging therapeutic relevance for the diagnosis and/or the treatment of numerous diseases, such as type 2 diabetes mellitus, Cushing disease, Alzheimer's disease, acromegaly, several neuroendocrine tumors, pain-associated conditions (inflammation, neuropathy, rheumatoid arthritis), and depression [6,14–17]. The native form of SST does not have clinical importance because of its short plasma half-life of 3 min [3] and various actions.

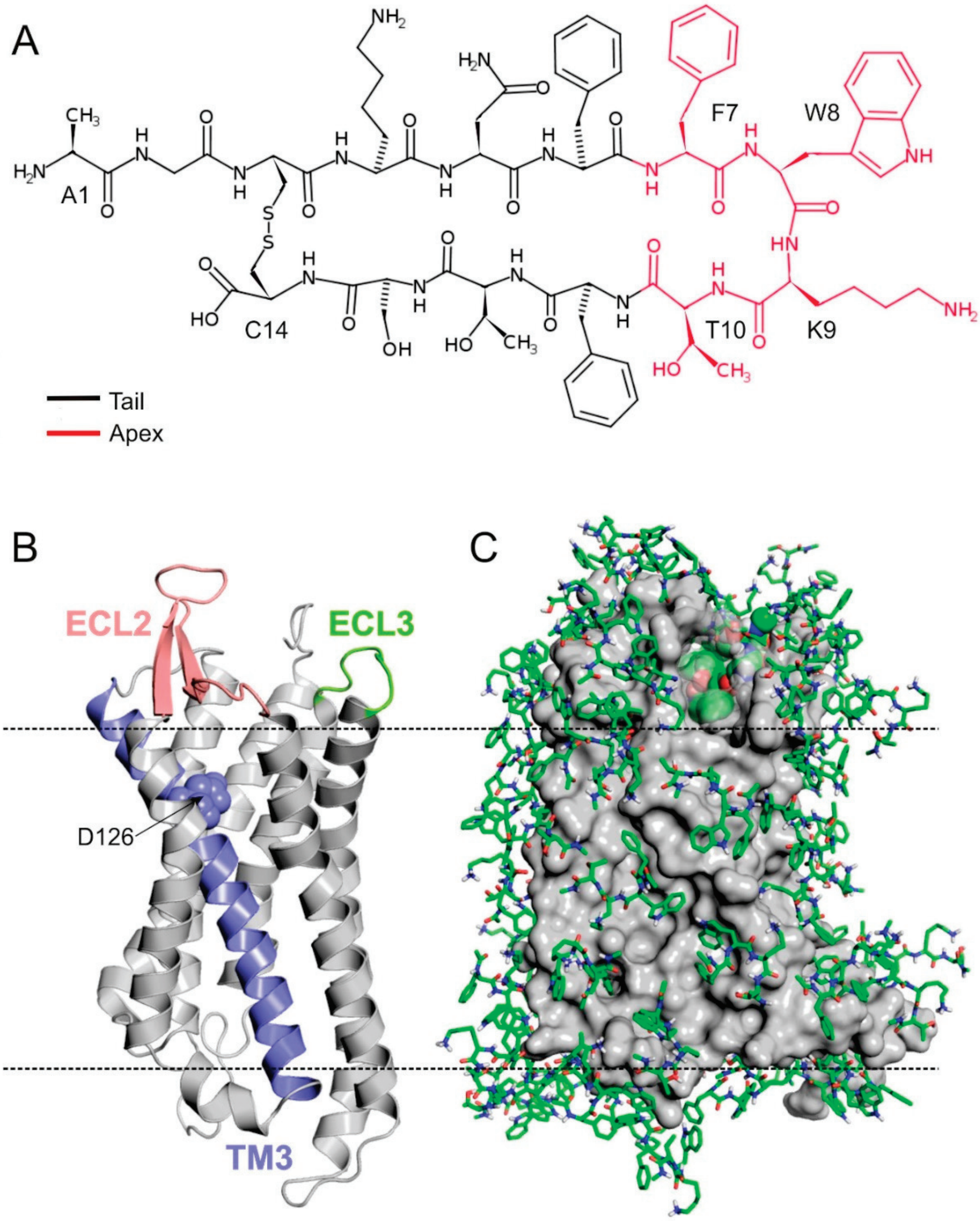


Figure 1. (**A**) Lewis structure of SST highlighting the apical FWKT region with red. (**B**) Homology model of SSTR4 in cartoon representation. D126 (spheres) on TM3 (teal), ECL2 (salmon), and ECL3 (green) are proved to be important in ligand binding and receptor activation. (**C**) SSTR4 (grey, surface) covered with monolayer of tetrapeptide fragment (Ace–FWKT–NHMe) copies (green, all atom, sticks) at the end of the 7th docking cycle. The best energy fragment is highlighted with spheres (green, all atom).

SST exerts its diverse biological effects via modulating somatostatin receptors (SSTRs). The therapeutic potential of SST–SSTR interactions is not fully utilized, and there is current pursuit for receptor-selective, orally-administrable drug candidates in several research groups and pharmaceutical companies [18–25]. SSTRs belong to the rhodopsin-like G-protein coupled receptor (GPCR) superfamily and contain seven transmembrane (TM) helices (Figure 1B) and extracellular (ECL) and intracellular (ICL) loops. ECL2 was suggested to play a major role in ligand binding and receptor activation. It was supported by mutational analysis and receptor chimera examinations, with the result that the ligand-binding pocket involves residues of TMs 3–7 and ECL2 that are responsible for high affinity ligand binding in all SSTR subtypes [26–28]. There are five SSTR subtypes named as SSTR1–SSTR5 with more than 50% sequence identity. The binding of SST is not SSTR subtype selective according to competitive radio-ligand measurements [29–33].

In this study, we focus on SSTR4 that proved to be a promising target in the treatment of inflammation and pain-associated conditions (neuropathic pain, neurogenic inflammation, bronchial asthma, rheumatoid arthritis), Alzheimer's disease [34,35], and depression [36,37]. SST elicits anti-inflammatory and anti-nociceptive actions and can be released from the capsaicin-sensitive sensory nerve endings. This is mediated through the activation of SSTR4 [38]. Centrally, SSTR4 is involved in learning and memory processes [10] and anxiety and depression-like behavior [37]. Thus, SSTR4 agonists would be promising drug candidates with analgesic, anti-inflammatory, and anti-depressant actions. However, there is no potent SSTR4 selective, orally administrable drug on the market [39–42]. Several SST analogs are under development [43], and many of them are used in therapy, such as pasireotide, octreotide, dopastatin, lanreotide, or in diagnostics [44–46]. Most of the studies [31,47–55] investigated either the binding mode of several exogenous peptidergic SST analogs as drug candidates or the residues of SST taking part in ligand binding. There are only a few studies [26,56,57] that examined the binding properties of endogenous ligands to SSTRs, which might be explained with the lack of atomic resolution experimental structures of SSTRs.

The target-based rational design of new agonists necessitates the atomic resolution structure of SSTR4 and its complex with the native (endogenous) ligand SST. Indirect experimental [26,31,57,58] and theoretical [48,50–53,59] information has been accumulated on the approximate binding sites of SST on SSTR4. While the atomic resolution structures of subtype SSTR2 and its complexes were measured recently [60], experimental determination of the atomic resolution structure of SSTR4 has not been published.

In the present study, we investigate the binding mechanism of SST to SSTR4. Atomic resolution structures of the SST–SSTR4 complex are produced using an unbiased, blind docking approach, and the binding mechanism is explored using molecular dynamics simulations in an explicit water model. We investigate if SST follows a "lock and key" or rather an induced fit mechanism and if it adopts prerequisite binding modes while approaching the final binding pocket on SSTR4.

2. Results and Discussion

2.1. The Structure of SSTR4

The experimental determination of the atomic resolution structure of SSTR4 has not been accomplished yet (Introduction). Homology modelling is an alternative method of choice [40,41,48,51–53,61,62] for producing SSTR structures. Building a good SSTR model necessitates the selection of a template protein of good sequential agreement with the receptor. The first homology modeling study of SSTR4 [51] used the active form of the β_2 adrenergic receptor (PDB code: 3p0g) as a structural template. In recent years, new template structures have emerged, and our BLAST [63] (Methods) search resulted in a list of new template proteins (Table S1) in the Protein Databank (PDB, [64]). A comparison of the homology models built from the templates led to the selection of the active form of the µ-opioid receptor (PDB code 5c1m) as a new template of SSTR4, also used in a previous study as a template of receptor subtypes SSTR2 [48]. The homology models generated from the old (3p0g) and new (5c1m, Figure 1B) templates showed overall similarity (Table S2) in the position of TM helices and differences in ECL2 possibly involved in ligand binding.

2.2. The External Binding Mode

Following generation of the homology model of SSTR4, a modified fragment blind docking (FBD) approach [65,66] was applied to locate the binding pocket of SST targeting the entire surface of receptor. The approach allows an unbiased (blind) detection of anchoring points of SST without prior information on the location of the binding pocket. Structure–activity relationship studies have shown [31,33,55,67] that central amino acids F7W8K9T10 (Figure 1A) play a pivotal role in SST binding activation of the SSTRs. The disulfide bond between C3 and C14 (Figure 1A) largely determines the positioning of FWKT in the apical β-turn region of the SST structure [31,68–70]. Accordingly, this FWKT fragment was used as a seed during FBD to locate the binding mode of the central region of SST on SSTR4. The entire surface of the (3p0g-based) homology model of SSTR4 was covered by a mono-layer of the copies of the blocked tetrapeptide fragment (Ace–FWKT–NHMe) using several wrapping cycles ([65] Section 3). After seven cycles, 74 copies covered the entire surface of the SSTR4 target (Figure 1C). The docked binding mode of the best interaction energy (E_{inter}, Table S3) found an extracellular binding cleft formed by the ECL1–3 (Figure 1C) regions of SSTR4.

The tetrapeptide–SSTR4 complex structure was used to construct the full length SST molecule in the binding cleft. This was achieved by a somewhat unusual application of the popular homology modelling program Modeller [71]. The program was instructed to grow the remaining ten amino acids of SST (Methods) in the binding cleft of the SSTR4 target structure (Methods), extending the tetrapeptide seed (Figure 2A). The resulting SST (full length)–SSTR4 complexes were energy-minimized, and the corresponding interaction energies were calculated (Methods). The SST structure in the raw complex with the best E_{inter} after the growing step (Figure 2B) did not adopt the above-mentioned β-turn structure at the FWKT region and resided at the extracellular surface of SSTR4 (see Section 3 for identifying criteria of a β-turn structure). SST also did not form a salt bridge with D126, a key residue involved in SSTR4 activation [26,28,56–58,70,72]. D126 is located deeper in the transmembrane region of TM3 (Figure 1B) and expectedly formed a salt bridge with the apical K9 in the final binding mode of SST.

Due to the apparent disagreement of the raw SST–SSTR4 complex with the above-mentioned literature data, it was subjected to further refinement in a 350 ns-long MD simulation (Section 3). The expected [31,33,55,67] β-turn structure of SST appeared for longer periods during the MD simulation. Migration of SST was also observed towards the transmembrane region, as indicated by the slight decrease of the distance of the expected SST:K9-SSTR4:D126 salt bridge (d_{SB}) from 21 (Figure 2B) to 18.5 Å (Figure 2C). In the MD-refined structure, the interaction of SST:K9 with ECL3 was broken down, while the connection of the tail regions of SST with ECL2 and ECL3 remained (Figure 2C,D).

During the MD refinement, movement H-bonds of SST:K9 with the target residues on ECL3 were broken down, and instead of the apical K9, backbone oxo groups of SST formed anchoring salt bridges with positively charged amino acids (R188, R191) of ECL2 (Figure 2D, Table S4). Interactions between SST and ECL2 were reasonable, as ECL2 is known [27] to have a lid function in SST association. Thus, the external binding mode identified at the ECL2 lid (Figure 2C,D) is certainly a prerequisite state en route to the internal binding mode.

2.3. The Internal Binding Mode

To construct the final, internal binding mode, the SSTR4–tetrapeptide complex (Figure 2A) was subjected next to a 100 ns-long MD simulation, where the tetrapeptide and the ECLs moved freely, but position restraints were applied on the TMs (Methods). It was expected that the tetrapeptide would find the internal binding mode faster than the full length SST due to its higher translational and conformational mobility. As can be seen in Figure 3A,

d_{SB} decreased from the initial 18.5 Å to about 10 Å (red squares in Figure 3A) several times, which may indicate the presence of a stable intermediate conformation of SST between its external and internal binding modes. From the 83rd ns, the fluctuation of d_{SB} decreased, reaching the lowest distance (5.2 Å) by 98.2 ns.

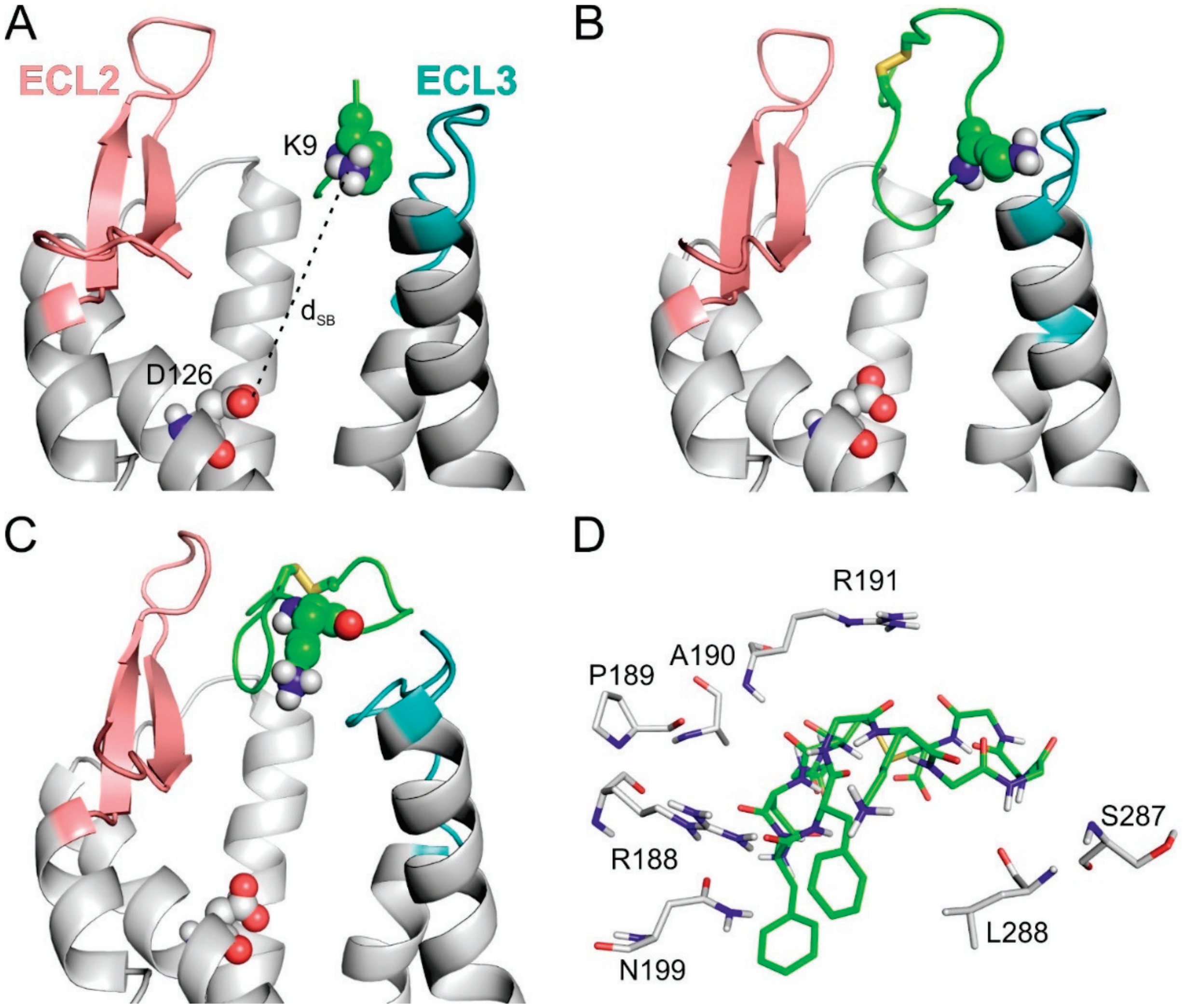


Figure 2. (**A**) SSTR4 with the best energy tetrapeptide fragment at the end of WNS. (**B**) The energy-minimized SSTR4–SST complex built from the "seed" of the fragment by homology modelling. (**C**) The SSTR4–SST complex in the prerequisite external binding mode after MD refinement. In (a, b, and c) ECL2 (salmon), ECL3 (teal), and D126 (spheres) are highlighted on SSTR4 (grey, cartoon) K9 of SST, and its fragment (green, cartoon) is in spheres representation. (**D**) The close-up view of the external binding mode of SST (green, sticks, all atom) with the target residues (grey, sticks, all atom) being within 3.5 Å distance of the ligand.

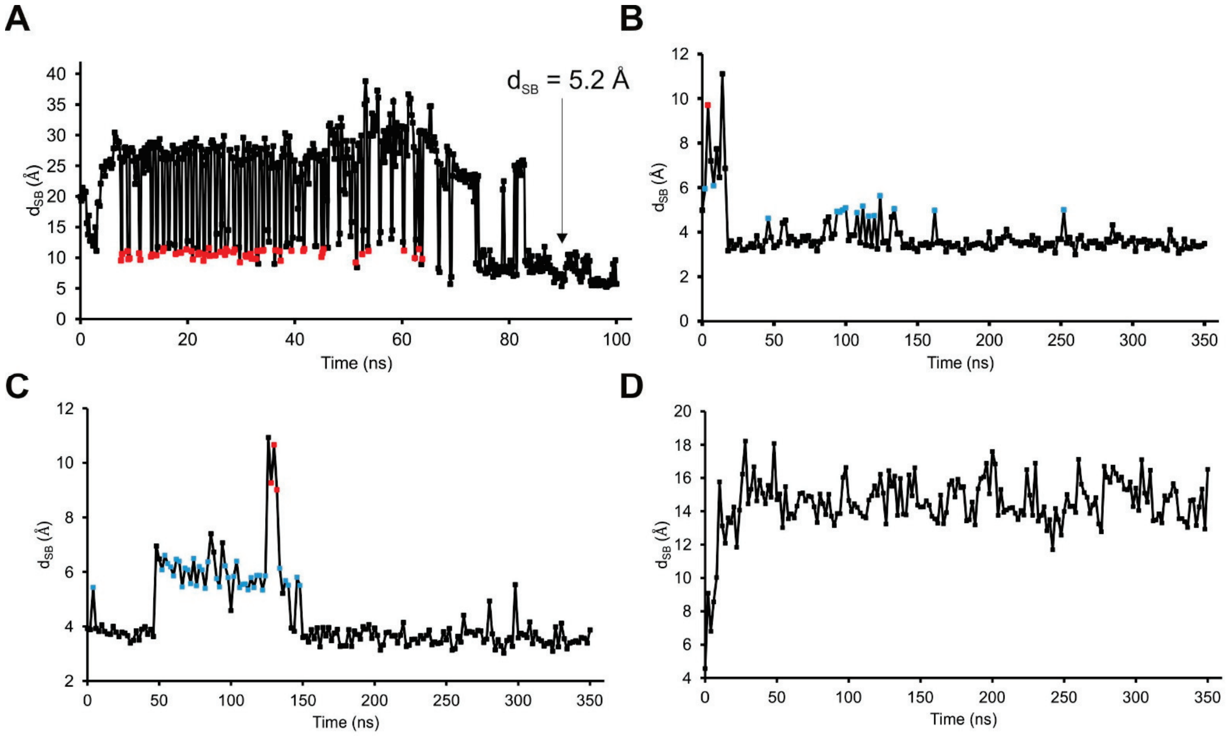


Figure 3. (**A**) D_{SB} plot of SSTR4–tetrapeptide fragment complex MD simulation for exploring the internal binding cleft of SST; (**B**–**D**) D_{SB} plots of the MD refinement of the three SSTR4–SST models containing the ligand in the internal binding cleft. In two of these simulations (**B**,**C**), SST was able to create a salt bride with D126 (d_{SB} = 3.1 Å and 3.0 Å), but in the third one (**D**), the dissociation of SST could be observed. Intermediate states are colored with red (d_{SB} = ~10 Å) and blue (d_{SB} = ~5 Å) points.

Similarly to the previous section, the full length SST molecule was grown from amino acid K9 (of aFWK9Tm) as a seed, with the lowest d_{SB} of 5.2 Å observed during MD (Figure 4A). The growing process (described in detail in Methods) resulted in three full length SST–SSTR4 complex structures subjected to three, respective, 350 ns-long MD simulations. In two of the MD simulations, SST:K9 reached a d_{SB} of 3.1 Å (Figures 3B and 4B) and 3.0 Å (Figures 3B and 4B), respectively. The third MD resulted in a backward movement of SST towards the external binding mode (increasing d_{SB} in Figure 3D). The interaction patterns (Table S5) of the internal binding modes described in Figure 4B,C were similar, and the one with a d_{SB} of 3.0 Å was selected as an internal binding mode for further description. In the internal binding mode, the position of SST was stabilized by salt bridges, and H-bonds formed with SSTR4 residues, including D126, N199, D289, and Y301 (Figure 4D).

2.4. The Binding Mechanism

The MD simulations of the previous section shed light on the association of SST with SSTR4 and its movement back to the external binding mode. Both associative MDs indicated that there were two highly occupied intermediate binding modes at a d_{SB} of 5–6 Å and 10 Å (Figure 3B,C), respectively. Notably, the intermediate at 10 Å was also identified in the simulation of the tetrapeptide–SSTR4 complex (Figure 3A). The steps of the associative movements were visualized (Figure 5, Video S1) and showed a considerable conformational change of SST during the binding process. The conformational flexibility of SST was the most pronounced at its apical region, which showed a large flip between the internal and external binding modes (Figure 6A).

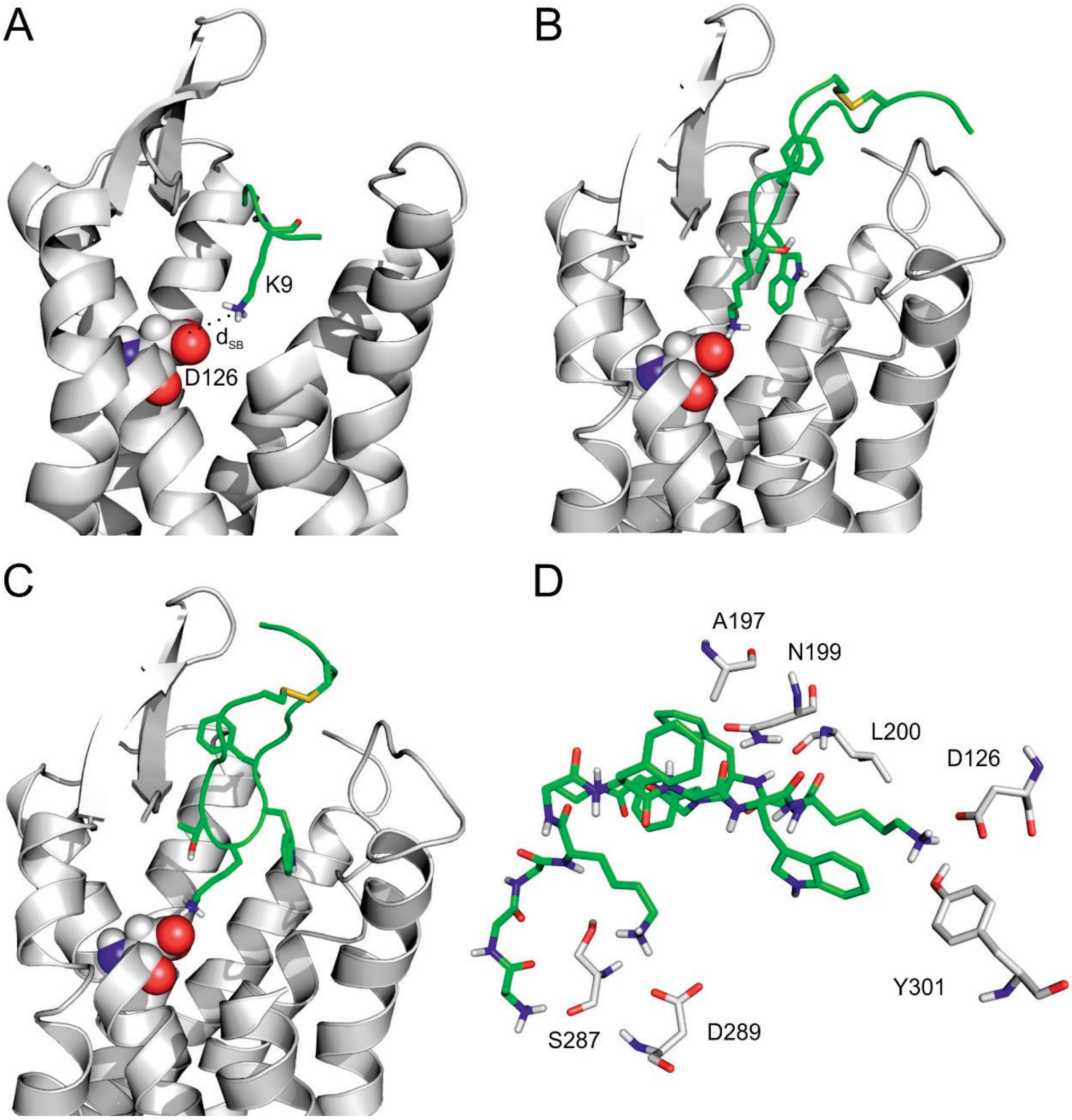


Figure 4. (**A**) Complex of SSTR4 (grey, cartoon, D126 highlighted by spheres) and the tetrapeptide fragment (green, K9 highlighted by sticks) with the smallest d_{SB} in the 100 ns-long MD simulation. (**B**,**C**) Internal binding mode of SST with 3.1 Å (**B**) and 3.0 Å d_{SB} determined in separate 350 ns-long MD simulations. (**D**) The close-up view of SST (green, sticks) in the internal binding mode surrounded with target residues (grey, sticks) within a 3.5 Å distance from the ligand.

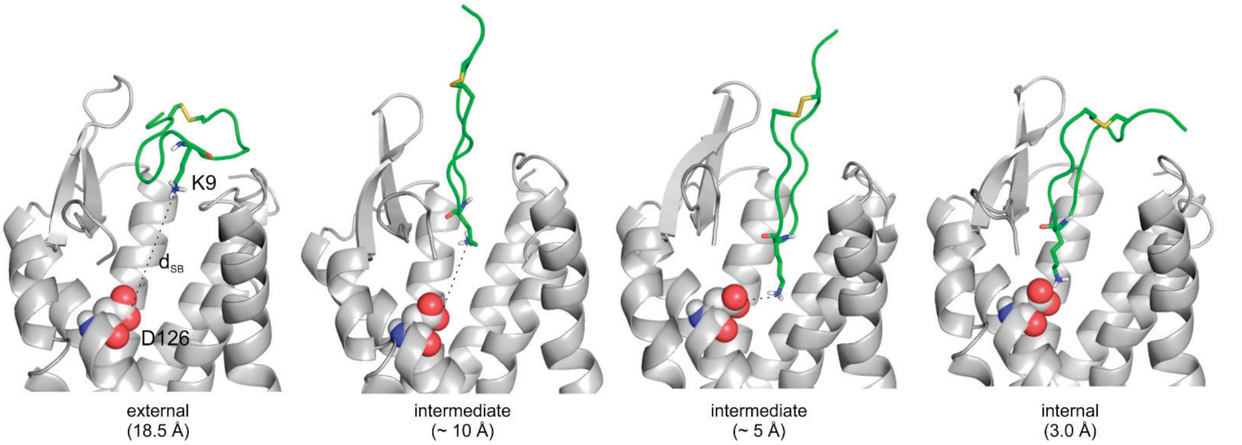


Figure 5. Main steps of the binding mechanism of SST (green, cartoon, K9 highlighted with sticks, all atoms) including external, intermediate (~10 Å and ~5 Å), and internal binding modes on SSTR4 (grey, cartoon). D126 is highlighted with spheres. This binding mechanism is illustrated in Video S1.

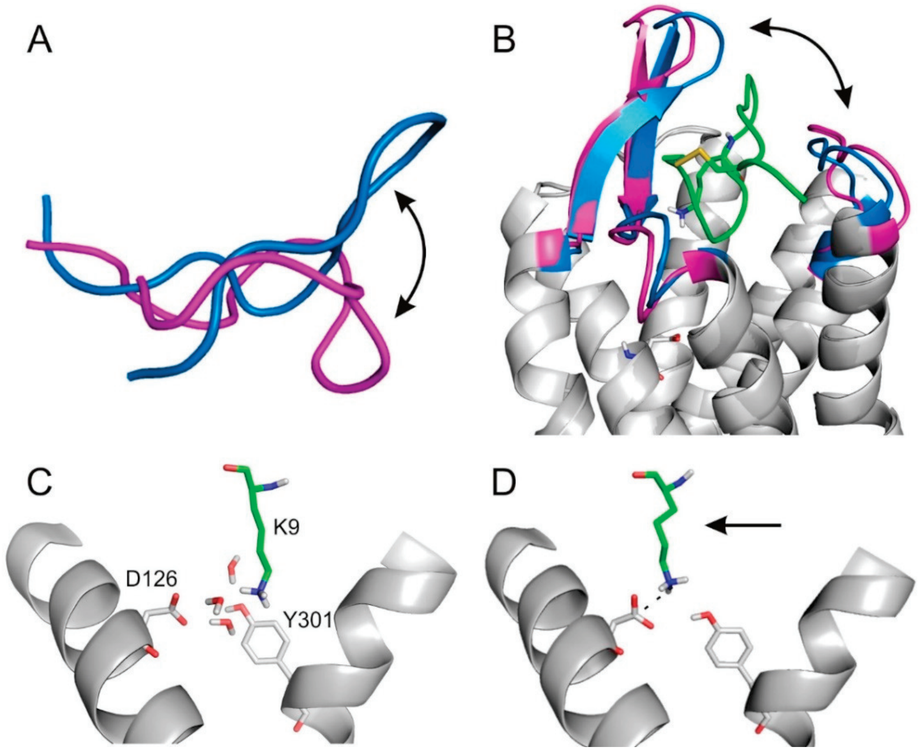


Figure 6. (**A**) The conformational change of SST during binding to SSTR4: internal (marine) and external (magenta) binding conformation of SST (cartoon) aligned by their tail regions. (**B**) Opening (magenta) and closing (marine) movements of the lid including ECL2 and ECL3 during the binding and dissociation of SST (green, cartoon, K9 highlighted with sticks, all atom). (**C**) The close-up view of the intermediate state (~5 Å) (Figure 5). The three water (grey, sticks, all atom) molecules help the connection of SST and SSTR4. (**D**) The close-up view of the final internal binding mode of SST (Figure 5) on SSTR4 after dehydration and movement (arrow) of SST:K9. In (**C**,**D**) K9:SST and SSTR4 are in green, sticks, all atom and grey, cartoon, D126, Y301 highlighted with sticks, all atom representation, respectively.

Similarly, SSTR4 also underwent a conformational change when SST moved from the intermediate state (d_{SB} = 10 Å) to the external binding position. The gap formed by ECL2 and ECL3 of SSTR4 increased to let the ligand dissociate from the receptor (Figures 3D and 6B). In agreement with our findings, oligopeptides such as SST are known to activate their receptor via an induced fit mechanism very common in similar receptor activation processes involving considerable conformational changes on both the target [73–75] and the ligand [74,76,77] sides.

Furthermore, the intermediate state at d_{SB} = 5–6 Å (Figure 5) was stabilized by a network of water molecules in the interface and linked the apical region of SST to SSTR4, as shown in a close-up (Figure 6C). There were three water molecules connecting D126, Y301, and SST:K9 via a H-bonding network (Figure 6C). The role of such networks has been described by recent studies [78]. However, the internal binding mode was finally stabilized only by the SST:K9-SSTR4:D126 salt bridge (d_{SB} = 3.0 Å) without the above interfacial water molecules (Figure 6D), indicating that a de-hydration process took place in the final binding step. Several target amino acids (A197, C198, N199, and D289) were involved in both the external and internal binding modes. These residues assist the transition movement of SST from the external towards the internal binding mode (see also Table S5).

All-in-all, the binding mechanism of SST to SSTR4 involves a migration between external and internal binding modes via intermediate states stabilized by water networks. The binding involves a conformational flip in the apical β-turn region of SST.

2.5. Comparison of SST Subtype Binding

Recent determination of the atomic resolution structure of the SSTR2–SST complex [60] allowed for comparison of the binding modes of SST on SSTR2 and SSTR4. The internal binding mode of SST on SSTR4 (Section 4) was used for this comparison. A per-residue energy analysis of the SSTR-SST interaction energy (E_{inter}) showed that residues D122(D126), S279(S287), Y302(Y301) are important for binding of SST to both SSTR2(SSTR4) receptor subtypes (Figure 7). D122 proved to be essential in receptor activation [56,57,72]. The E_{inter} pattern on the SST side (Figure 6B) showed that residues A1, G2, K4, K9, and C14 are important in the interaction with both receptor subtypes. A difference could be observed at N5 and F6 (I284, V280, Y205, E200, R184) positions, preferring SSTR2, while F7 and F11 (D289, T286, L200, N199) are involved in the SSTR4 complex. The role of Fs and K4 was also suggested by previous alanine scanning studies [29].

An overall ca. 180° flip of the binding conformation of SST (Figure S1) could be observed between the internal binding mode on SSTR4 if compared with that of SSTR2 [60]. An E_{inter} analysis was also performed for the alternative binding mode of SST (observed in [60]) on SSTR4 (see Methods for details of construction of the complex). The E_{inter} plots (Figure S2) showed that K4, K9, and C14 (SST) and D126, S287, and Y301 (SSTR4) are important in all binding modes. F6 has importance only in case of SSTR2. W8 and N5 (on the SST side) and L283 and Q201 (on the receptor side) were identified as important residues only in the alternative binding mode. The above differences in SST binding to SSTR2 and 4 may serve as a good starting point in the design of subtype-selective SST analogues.

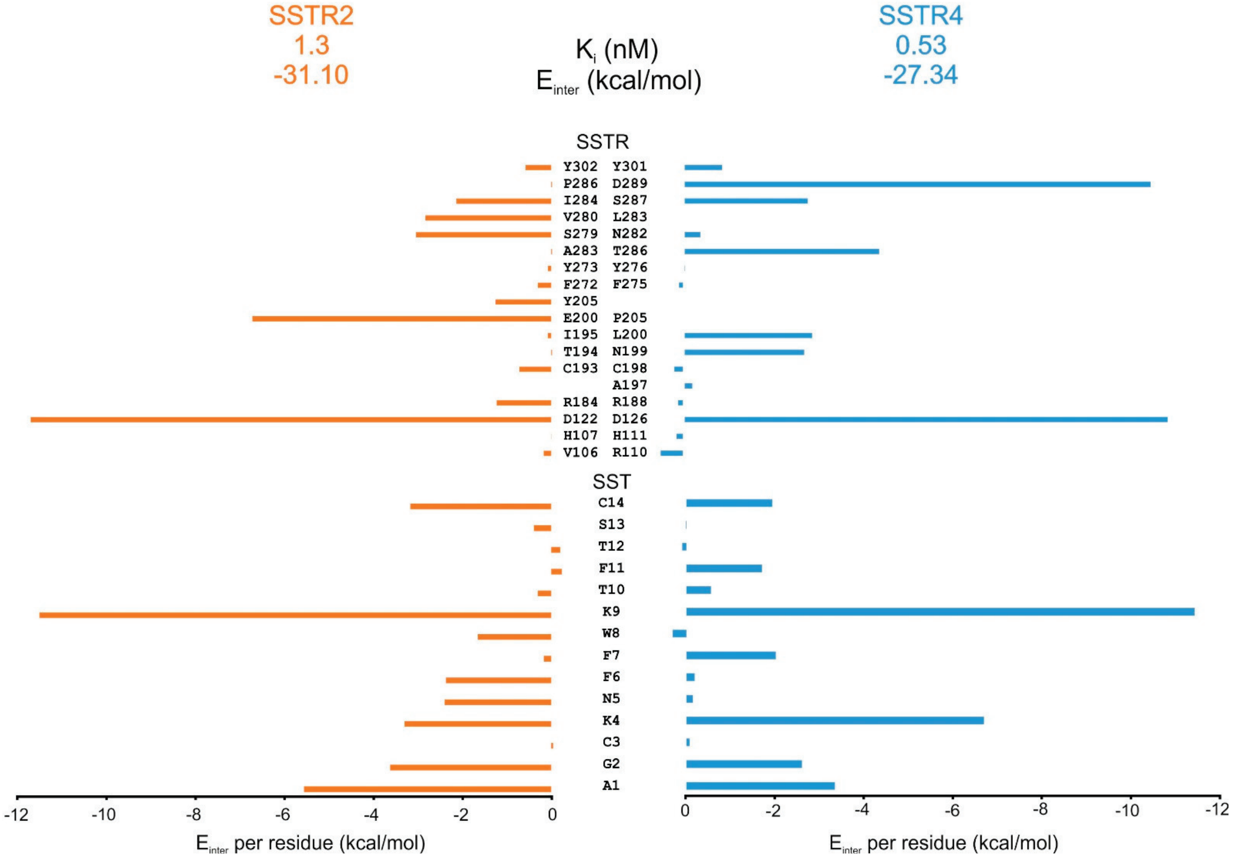


Figure 7. Per-residue E_{inter} contributions of SSTR2–SST (orange) and SSTR4–SST (blue) complexes shown for both the receptor (SSTR) and the somatostatin (SST) sides. Note that different amino acids may appear at identical positions in the sequences of SSTR2 and 4 after sequence alignment, as listed in the SSTR-based analysis (top). Source of K_i values is [32].

3. Methods

3.1. The Structure of SSTR4

A BLAST (Basic Local Alignment Search Tool) [63] search with Blosum 62 substitution matrix using a conditional compositional score matrix adjustment at NBCI [79] against the PDB Database [64] was applied to identify the template candidates for model building. The BLAST search resulted in 100 PDB codes. They were ranked according to their total scores. The best ranked template candidates were 4n6h, 4rwa, 6dde, and 5c1m (Table S1). The structure of the δ-opioid receptor bound to a bifunctional peptide (PDB code: 4rwa) was excluded. Structures 5c1m and 6dde represent the crystal structures of agonist binding μ-opioid receptors, and 5c1m had a better resolution (2.1 Å compared to 3.5 Å for 6dde). The A chain of both the human δ-opioid receptor (4n6h) and the active form of the μ-opioid receptor (5c1m), and, furthermore, the active form of the β_2-adrenergic receptor (3p0g) used in a previous study were employed for model building described in the paper of Liu et al. [51]. SSR4 sequence was taken from the UniProt database (P31391 (37-330)) the not-aligned *N* and *C* terminals were cut). After the sequence alignment using the Modeller program package [71], ten models were generated from each template, and models with the lowest Discrete Optimized Protein Energy (DOPE) score were further investigated (Table S6). The RSMD value of CA atoms for the best models was calculated (Table S2). The models were superimposed, and their structures were compared. Due to the high similarity of the opioid receptor-derived models, only the μ-opioid receptor (5c1m)- and β-adrenergic receptor (3p0g)-based homology models were used for further investigations.

3.2. The External Binding Mode

3.2.1. Fragment of SST

The NMR structure of SST dissolved in 5% D-mannitol is known (PDB code: 2mi1). The apical region of SST, F7-W8-K9-T10, was extracted, and its *N* and *C* terminals were capped with acetyl and *N*-methyl groups (Ace–FWKT–NHMe) to neutralize the terminal charges. This Ace–FWKT–NHMe was used for docking calculations.

3.2.2. Energy Minimization

A uniform two-step energy minimization process in AMBER99SB-ILDN force field by GROMACS [80] was used prior to MD simulations. Molecules were placed in the center of a cubic box with the distance of 10 Å between the box and the solute atoms. The simulation box was filled with TIP3P explicit [81] water molecules and counter ions to neutralize the total charge of the system. The convergence thresholds of the first (steepest descent) and second (conjugant gradient) steps of minimization were set to 100 and 10 kJ mol^{-1} nm^{-2}, respectively.

3.2.3. Docking Calculations

The energy-minimized target structures were used in docking calculations. The Wrapper module of the WnS method [65] was applied for Fragment Blind Docking (FBD) during which the entire surface of the target (3p0g) was covered by a mono-layer of the Ace–FWKT–NHMe copies by a series of blind docking cycles performed by AutoDock and AutoGrid [82]. Docking parameters were used for FBD, as described in our previous studies [65,83]. Wrapping the target into ligand copies allows systematic mapping of all possible binding modes of a ligand. At the end of wrapping, the fragment bound with the lowest E_{inter} was chosen as the best ligand position (Table S3). The resulting docked complex was superimposed on the receptor structure of 5c1m and used in the next growing step. The distance between the amino *N* atom of K9:SST and the carboxylate *C* atom of D126:SSTR4 (d_{SB}) was determined. The docking calculations were not focused on a selected region of the protein, and the ligand could navigate without positional or torsional constraints during docking. Thus, the blind docking calculations were unbiased without the use of previous knowledge of the binding site. The binding modes of the ligand covered the entire surface of the protein after blind docking with Wrapper (Figure 1C). The binding mode with the most favorable calculated E_{inter} was selected for further homology modelling steps.

3.2.4. Growing of SST into the Binding cleft

The full-length ligand was built into the receptor using the fragment as a seed by the homology modelling approach. SSTR4-FWKT (Ace–FWKT–NHMe without the capping groups) structures were used as templates, and the query sequence was the sequence of the receptor and the full length ligand together taken from UniProt database (SSTR4: P31391 (37–330), like the homology models, SST: P61278 (103–116)). Structure alignment was manually optimized to obtain the identical regions correctly under each other. The Modeller [71] program package was applied to build ten models for each template. Explicit manual restraint was added to generate the disulfide bond in SST. As the DOPE score was very similar for all generated models, the E_{inter} (Table S6) values were calculated [84] (Lennard–Jones energy, Amber parameters [85,86]) and applied for model selection.

3.2.5. Molecular Dynamics Simulation

For identifying the internal binding mode of SST and investigating its binding mechanism on SSTR4, a series of MD simulations was applied in the TIP3P explicit water model with the AMBER99SB-ILDN force field using the GROMACS program package following two step energy-minimization (described in Section 3.2.2). In all cases, the target was treated as a rigid body, except the ECL regions (37–42; 109–225; 184–208; 284–294), to allow the entrance of the ligand into the receptor. Position restraints were applied on the heavy atoms of TMs with a force constant of 100 kJ/mol^{-1} nm^{-2}. For temperature coupling, the velocity rescale and the Parrinello–Rahman algorithm were used. Solute and solvent were coupled separately with a reference temperature of 310.15 K and a coupling time constant of 0.1 ps. The protonation states of amino acids were set according to pH 7.4. Pressure was coupled by the Parrinello–Rahman algorithm and a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1}, and reference pressure of 1 bar. Particle Mesh–Ewald summation was used for long range electrostatics. Van der Waals and Coulomb interactions had a cut-off at 11 Å. Periodic boundary conditions were treated after the finish of the calculations. After each trajectory, the periodic boundary effects were handled, the system was centered in the box, and target molecules in subsequent frames were fit on the top of the first frame. The final trajectory including all atomic coordinates of all frames were converted to portable xdr-based xtc binary files.

3.2.6. MD Refinement in the External Binding Cleft

The SSTR4–SST complex was submitted to a 350 ns-long MD simulation described above, and d_{SB} was calculated throughout the MD simulation using the gmx distance modul of GROMACS. The structure with the smallest d_{SB} was determined as the external binding mode of SST on SSTR4. Interacting target residues within a 3.5 Å distance of SST were determined and listed (Table S4).

3.3. The Internal Binding Mode

3.3.1. Molecular Simulation for Exploring the Internal Binding Cleft

The exploration of the internal binding mode of SST was performed similarly to the external one; however, the location/position of the SST tetrapeptide seed was determined by a 100 ns-long MD simulation instead of docking. After the two-step energy minimization process, the SSTR4–Ace–FWKT–NHMe complex (5c1m-based model with the superimposed aFWKTm) was submitted to a 100 ns-long MD simulation.

3.3.2. Growing the Full Length SST into the Receptor

After the 100 ns-long MD, the structure with the smallest d_{SB} was used to build the full length SST into the receptor similarly to the external binding mode. After generating the homologies (similar to the method of external binding cleft), many close contacts occurred in the structures that remained also after the two step energy minimization procedure. Thus, in this case, instead of the whole apical FWKT region, only the K9:SST was used as a “seed” for building the ligand. Models (3 × 10) were generated using no, 5 Å, and 6 Å distance restraints on d_{SB}, respectively, and in all homologies, d_{SB} was determined again (Table S8). Models with the smallest d_{SB} from each group were further investigated using MD simulation.

3.3.3. MD Refinement in the Internal Binding Cleft

The energy-minimized models with the smallest d_{SB} distance from each group (Table S8) were submitted to a separate 350 ns-long MD simulation to investigate the associative and dissociative movements of the ligand. Calculation of d_{SB} was performed throughout each MD simulation, and structures having the smallest one were determined as the internal binding position of SST, and the interacting target residues within 3.5 Å distance from the ligand were determined (Table S5).

3.4. Comparison of SST Subtype Binding

3.4.1. Determination of Interacting Energy per Residues in SSTR4/SSTR2–SST Complexes

Following the two step energy minimization procedure, Coulomb intermolecular interaction energies were calculated [84] with a distance-dependent dielectric function [87] and Amber partial charges [85,86] globally and per residues for both SSTR2–SST and internal SSTR4–SST complexes. Comparison of the per residue interacting energies was based on the sequence alignment of the targets created by EMBOSS Needle [88].

3.4.2. Energy Analysis of Alternative Binding Mode

There was a ca. 180° flip of the binding conformation of SST (Figure S1) in the internal binding mode on SSTR4 compared with that of SSTR2. Thus, the SSTR4–SST complex with this alternative binding mode was constructed by superimposing the targets. Following a two-step energy-minimization, a global and per residue E_{inter} analysis was also performed for this structure.

4. Conclusions

The present study investigated the binding mechanism of SST to SSTR4. While the SST–SSTR2 structure was recently published, the atomic level complex of SST and SSTR4 has not been determined yet. As SSTR4 also plays an important role in the pathobiochemistry of various diseases (Introduction), we thus focused on the calculation of SST–SSTR4 complex structures. Beyond the complex structures, the dynamics of the binding mechanism of SST was also elucidated, and key binding modes (external, intermediate, and internal) were distinguished. The role of induced fit and hydration was discussed. The most important residues on both receptor and SST sides were identified. Finally, an energetic comparison of SST binding to SSTR2 and 4 offered a residue-level explanation of receptor subtype selectivity. In good agreement with experimental results, we found that the extracellular regions of helices and loops play an important role in SST binding, and structural differences in these regions are important in receptor subtype selectivity. The detailed structural comparison of SST binding to SSTR2 and 4 helps in the development of new, subtype, and disease-selective SST analogues.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23136878/s1.

Author Contributions: Conceptualization, C.H. and R.B.; methodology, R.B., C.H. and M.B.; investigation, R.B, M.B. and B.Z.Z.; resources, C.H.; data curation, R.B.; writing—original draft preparation, R.B., C.H. and B.Z.Z.; writing—review and editing, R.B, C.H., Z.H. and E.P.; visualization, R.B.; supervision, C.H.; project administration, C.H.; funding acquisition, C.H. and E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Hungarian National Research, Development, and Innovation Office (K123836 and K134214). Project no. TKP2021-EGA-16 has been implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the EGA-16 funding scheme. Project no. TKP2021-EGA-13 has been implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the EGA-13 funding scheme and the Lorand Eotvos research Network. This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The work was supported by the ÚNKP-21-5 and ÚNKP-21-3-II New National Excellence Program of the Ministry for Innovation and Technology and by the PTE ÁOK-KA 2021/KA-2021-39. The project has been supported by the European Union, co-financed by the European Social Fund, project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research have been achieved using the DECI resource Archer2 based in the UK at the National Supercomputing Service with support from the PRACE aisbl.

Conflicts of Interest: E. Pinter and Zs. Helyes are co-founders and shareholders of Algonist GmBH Austria and PharmInVivo Ltd., Hungary, focusing on drug discovery for (among others) chronic pain and services in pain models, respectively. They declare no conflicts of interest with the present work.

Abbreviation

SST	somatostatin
TM	transmembrane
ECL	extracellular loop
ICL	intracellular loop
SSTR1–5	somatostatin receptor subtype 1–5
GPCR	G-protein coupled receptor
WNS	Wrap ‘n’ Shake
MD	molecular dynamics
PDB	Protein Data Bank
d_{SB}	Distance between the amino N atom of SST:K9 and the carboxylate C atom of SSTR4:D126
BLAST	Basic Local Alignment Search Tool
FBD	Fragment blind docking

References

1. Esch, F.; Böhlen, P.; Ling, N.; Benoit, R.; Brazeau, P.; Guillemin, R. Primary Structure of Ovine Hypothalamic Somatostatin-28 and Somatostatin-25. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 6827–6831. [CrossRef] [PubMed]
2. Hoyer, D.; Bell, G.I.; Berelowitz, M.; Epelbaum, J.; Feniuk, W.; Humphrey, P.P.A.; O’Carroll, A.-M.; Patel, Y.C.; Schonbrunn, A.; Taylor, J.E.; et al. Classification and Nomenclature of Somatostatin Receptors. *Trends Pharmacol. Sci.* **1995**, *16*, 86–88. [CrossRef]
3. Patel, Y.C. Somatostatin and Its Receptor Family. *Front. Neuroendocr.* **1999**, *20*, 157–198. [CrossRef] [PubMed]
4. Pradayrol, L.; Jörnvall, H.; Mutt, V.; Ribet, A. N-Terminally Extended Somatostatin: The Primary Structure of Somatostatin-28. *FEBS Lett.* **1980**, *109*, 55–58. [CrossRef]
5. Olias, G.; Viollet, C.; Kusserow, H.; Epelbaum, J.; Meyerhof, W. Regulation and Function of Somatostatin Receptors. *J. Neurochem.* **2004**, *89*, 1057–1091. [CrossRef]
6. Rai, U.; Thrimawithana, T.R.; Valery, C.; Young, S.A. Therapeutic Uses of Somatostatin and Its Analogues: Current View and Potential Applications. *Pharm. Ther.* **2015**, *152*, 98–110. [CrossRef]
7. Shamsi, B.H.; Chatoo, M.; Xu, X.K.; Xu, X.; Chen, X.Q. Versatile Functions of Somatostatin and Somatostatin Receptors in the Gastrointestinal System. *Front. Endocrinol.* **2021**, *12*, 652363. [CrossRef]
8. Helyes, Z.; Pintér, E.; Németh, J.; Kéri, G.; Thán, M.; Oroszi, G.; Horváth, A.; Szolcsányi, J. Anti-Inflammatory Effect of Synthetic Somatostatin Analogues in the Rat. *Br. J. Pharmacol.* **2001**, *134*, 1571–1579. [CrossRef]
9. Pintér, E.; Helyes, Z.; Szolcsányi, J. Inhibitory Effect of Somatostatin on Inflammation and Nociception. *Pharmacol. Ther.* **2006**, *112*, 440–456. [CrossRef]
10. Gastambide, F.; Lepousez, G.; Viollet, C.; Loudes, C.; Epelbaum, J.; Guillou, J.-L. Cooperation between Hippocampal Somatostatin Receptor Subtypes 4 and 2: Functional Relevance in Interactive Memory Systems. *Hippocampus* **2010**, *20*, 745–757. [CrossRef]
11. Prasoon, P.; Kumar, R.; Gautam, M.; Sebastian, E.K.; Reeta, K.H.; Ray, S.B. Role of Somatostatin and Somatostatin Receptor Type 2 in Postincisional Nociception in Rats. *Neuropeptides* **2015**, *49*, 47–54. [CrossRef]
12. Schuelert, N.; Just, S.; Kuelzer, R.; Corradini, L.; Gorham, L.C.J.; Doods, H. The Somatostatin Receptor 4 Agonist J-2156 Reduces Mechanosensitivity of Peripheral Nerve Afferents and Spinal Neurons in an Inflammatory Pain Model. *Eur. J. Pharmacol.* **2015**, *746*, 274–281. [CrossRef]
13. Yeo, X.Y.; Cunliffe, G.; Ho, R.C.; Lee, S.S.; Jung, S. Potentials of Neuropeptides as Therapeutic Agents for Neurological Diseases. *Biomedicines* **2022**, *10*, 343. [CrossRef] [PubMed]
14. Boehm, B.O.; Lustig, R.H. Use of Somatostatin Receptor Ligands in Obesity and Diabetic Complications. *Best Pract. Res. Clin. Gastroenterol.* **2002**, *16*, 493–509. [CrossRef] [PubMed]
15. Chanson, P. Medical Treatment of Acromegaly with Dopamine Agonists or Somatostatin Analogs. *Neuroendocrinology* **2016**, *103*, 50–58. [CrossRef] [PubMed]
16. Hofland, L.J. Somatostatin and Somatostatin Receptors in Cushing’s Disease. *Mol. Cell. Endocrinol.* **2008**, *286*, 199–205. [CrossRef]
17. Sun, L.; Coy, D.H. Somatostatin and Its Analogs. *Curr. Drug. Targets* **2016**, *17*, 529–537. [CrossRef]
18. Abdel-Magid, A.F. Treating Pain with Somatostatin Receptor Subtype 4 Agonists. *ACS Med. Chem. Lett.* **2015**, *6*, 110–111. [CrossRef]
19. Clinical Development Pipeline | Science | Eli Lilly and Company. Available online: https://www.lilly.com/discovery/clinical-development-pipeline#/ (accessed on 3 May 2021).

20. Banno, Y.; Sasaki, S.; Kamata, M.; Kunitomo, J.; Miyamoto, Y.; Abe, H.; Taya, N.; Oi, S.; Watanabe, M.; Urushibara, T.; et al. Design and Synthesis of a Novel Series of Orally Active, Selective Somatostatin Receptor 2 Agonists for the Treatment of Type 2 Diabetes. *Bioorganic Med. Chem.* **2017**, *25*, 5995–6006. [CrossRef]
21. Fattah, S.; Brayden, D.J. Progress in the Formulation and Delivery of Somatostatin Analogs for Acromegaly. *Ther. Deliv.* **2017**, *8*, 867–878. [CrossRef]
22. Giovannini, R.; Cui, Y.; Doods, H.; Ferrara, M.; Just, S.; Kuelzer, R.; Lingard, I.; Mazzaferro, R.; Rudolf, K. New Somatostatin Receptor Subtype 4 (Sstr4) Agonists. Patent International Publication Number WO 2014/184275 A1, 20 November 2014.
23. Johnson, M.L.; Meyer, T.; Halperin, D.M.; Fojo, A.T.; Cook, N.; Blaszkowsky, L.S.; Schlechter, B.L.; Yao, J.C.; Jemiai, Y.; Kriksciukaite, K.; et al. First in Human Phase 1/2a Study of PEN-221 Somatostatin Analog (SSA)-DM1 Conjugate for Patients (PTS) with Advanced Neuroendocrine Tumor (NET) or Small Cell Lung Cancer (SCLC): Phase 1 Results. *J. Clin. Oncol.* **2018**, *36*, 4097. [CrossRef]
24. Liu, W.; Shao, P.P.; Liang, G.-B.; Bawiec, J.; He, J.; Aster, S.D.; Wu, M.; Chicchi, G.; Wang, J.; Tsao, K.-L.; et al. Discovery and Pharmacology of a Novel Somatostatin Subtype 5 (SSTR5) Antagonist: Synergy with DPP-4 Inhibition. *ACS Med. Chem. Lett.* **2018**, *9*, 1082–1087. [CrossRef] [PubMed]
25. Mansi, R.; Abid, K.; Nicolas, G.P.; Del Pozzo, L.; Grouzmann, E.; Fani, M. A New 68Ga-Labeled Somatostatin Analog Containing Two Iodo-Amino Acids for Dual Somatostatin Receptor Subtype 2 and 5 Targeting. *EJNMMI Res.* **2020**, *10*, 90. [CrossRef] [PubMed]
26. Kaupmann, K.; Bruns, C.; Raulf, F.; Weber, H.P.; Mattes, H.; Lübbert, H. Two Amino Acids, Located in Transmembrane Domains VI and VII, Determine the Selectivity of the Peptide Agonist SMS 201-995 for the SSTR2 Somatostatin Receptor. *EMBO J.* **1995**, *14*, 727–735. [CrossRef]
27. Leu, F.P.; Nandi, M. GPCR Somatostatin Receptor Extracellular Loop 2 Is a Key Ectodomain for Making Subtype-Selective Antibodies with Agonist-like Activities in the Pancreatic Neuroendocrine Tumor BON Cell Line. *Pancreas* **2010**, *39*, 1155–1166. [CrossRef] [PubMed]
28. Liapakis, G.; Fitzpatrick, D.; Hoeger, C.; Rivier, J.; Vandlen, R.; Reisine, T. Identification of Ligand Binding Determinants in the Somatostatin Receptor Subtypes 1 and 2. *J. Biol. Chem.* **1996**, *271*, 20331–20339. [CrossRef]
29. Bruns, C.; Raulf, F.; Hoyer, D.; Schloos, J.; Lübbert, H.; Weckbecker, G. Binding Properties of Somatostatin Receptor Subtypes. *Metabolism* **1996**, *45*, 17–20. [CrossRef]
30. Bruns, C.; Lewis, I.; Briner, U.; Meno-Tetang, G.; Weckbecker, G. SOM230: A Novel Somatostatin Peptidomimetic with Broad Somatotropin Release Inhibiting Factor (SRIF) Receptor Binding and a Unique Antisecretory Profile. *Eur. J. Endocrinol.* **2002**, *146*, 707–716. [CrossRef]
31. Lewis, I.; Bauer, W.; Albert, R.; Chandramouli, N.; Pless, J.; Weckbecker, G.; Bruns, C. A Novel Somatostatin Mimic with Broad Somatotropin Release Inhibitory Factor Receptor Binding and Superior Therapeutic Potential. *J. Med. Chem.* **2003**, *46*, 2334–2344. [CrossRef]
32. Patel, Y.C.; Srikant, C.B. Subtype Selectivity of Peptide Analogs for All Five Cloned Human Somatostatin Receptors (Hsstr 1-5). *Endocrinology* **1994**, *135*, 2814–2817. [CrossRef]
33. Rohrer, S.P.; Birzin, E.T.; Mosley, R.T.; Berk, S.C.; Hutchins, S.M.; Shen, D.-M.; Xiong, Y.; Hayes, E.C.; Parmar, R.M.; Foor, F.; et al. Rapid Identification of Subtype-Selective Agonists of the Somatostatin Receptor Through Combinatorial Chemistry. *Science* **1998**, *282*, 737–740. [CrossRef] [PubMed]
34. Sandoval, K.E.; Farr, S.A.; Banks, W.A.; Niehoff, M.L.; Morley, J.E.; Crider, A.M.; Witt, K.A. Chronic Peripheral Administration of Somatostatin Receptor Subtype-4 Agonist NNC 26-9100 Enhances Learning and Memory in SAMP8 Mice. *Eur. J. Pharm.* **2011**, *654*, 53–59. [CrossRef] [PubMed]
35. Sandoval, K.E.; Witt, K.A.; Crider, A.M.; Kontoyianni, M. Somatostatin Receptor-4 Agonists as Candidates for Treatment of Alzheimer's Disease. In *Drug Design and Discovery in Alzheimer's Disease*; Bentham Science Publishers: Sharjah, United Arab Emirates, 2014; pp. 566–597. ISBN 978-0-12-803959-5. [CrossRef]
36. Kecskés, A.; Pohóczky, K.; Kecskés, M.; Varga, Z.V.; Kormos, V.; Szőke, É.; Henn-Mike, N.; Fehér, M.; Kun, J.; Gyenesei, A.; et al. Characterization of Neurons Expressing the Novel Analgesic Drug Target Somatostatin Receptor 4 in Mouse and Human Brains. *Int. J. Mol. Sci.* **2020**, *21*, 7788. [CrossRef] [PubMed]
37. Scheich, B.; Gaszner, B.; Kormos, V.; László, K.; Ádori, C.; Borbély, É.; Hajna, Z.; Tékus, V.; Bölcskei, K.; Ábrahám, I.; et al. Somatostatin Receptor Subtype 4 Activation Is Involved in Anxiety and Depression-like Behavior in Mouse Models. *Neuropharmacology* **2016**, *101*, 204–215. [CrossRef] [PubMed]
38. Helyes, Z.; Pintér, E.; Németh, J.; Sándor, K.; Elekes, K.; Szabó, Á.; Pozsgai, G.; Keszthelyi, D.; Kereskai, L.; Engström, M.; et al. Effects of the Somatostatin Receptor Subtype 4 Selective Agonist J-2156 on Sensory Neuropeptide Release and Inflammatory Reactions in Rodents. *Br. J. Pharm.* **2006**, *149*, 405–415. [CrossRef] [PubMed]
39. Lilly Clinical Development Pipeline. Available online: https://www.lilly.com/discovery/pipeline (accessed on 6 August 2020).
40. Kántás, B.; Börzsei, R.; Szőke, É.; Bánhegyi, P.; Horváth, Á.; Hunyady, Á.; Borbély, É.; Hetényi, C.; Pintér, E.; Helyes, Z. Novel Drug-Like Somatostatin Receptor 4 Agonists Are Potential Analgesics for Neuropathic Pain. *Int. J. Mol. Sci.* **2019**, *20*, 6245. [CrossRef]

41. Szőke, É.; Bálint, M.; Hetényi, C.; Markovics, A.; Elekes, K.; Pozsgai, G.; Szűts, T.; Kéri, G.; Őrfi, L.; Sándor, Z.; et al. Small Molecule Somatostatin Receptor Subtype 4 (Sst4) Agonists Are Novel Anti-Inflammatory and Analgesic Drug Candidates. *Neuropharmacology* **2020**, *178*, 108198. [CrossRef]
42. Tigerstedt, N.-M.; Aavik, E.; Aavik, S.; Savolainen-Peltonen, H.; Hayry, P. Vasculoprotective Effects of Somatostatin Receptor Subtypes. *Mol. Cell. Endocrinol.* **2007**, *279*, 34–38. [CrossRef]
43. Dasgupta, P.; Gűnther, T.; Schulz, S. Pharmacological Characterization of Veldoreotide as a Somatostatin Receptor 4 Agonist. *Life* **2021**, *11*, 1075. [CrossRef]
44. Orlewska, E.; Stępień, R.; Orlewska, K. Cost-Effectiveness of Somatostatin Analogues in the Treatment of Acromegaly. *Expert Rev. Pharm. Outcomes Res.* **2019**, *19*, 15–25. [CrossRef]
45. Reynaert, H.; Colle, I. Treatment of Advanced Hepatocellular Carcinoma with Somatostatin Analogues: A Review of the Literature. *Int. J. Mol. Sci.* **2019**, *20*, 4811. [CrossRef] [PubMed]
46. Stueven, A.K.; Kayser, A.; Wetz, C.; Amthauer, H.; Wree, A.; Tacke, F.; Wiedenmann, B.; Roderburg, C.; Jann, H. Somatostatin Analogues in the Treatment of Neuroendocrine Tumors: Past, Present and Future. *Int. J. Mol. Sci.* **2019**, *20*, 3049. [CrossRef] [PubMed]
47. Crider, A.M.; Witt, K.A. Somatostatin Sst4 Ligands: Chemistry and Pharmacology. *Mini. Rev. Med. Chem.* **2007**, *7*, 213–220. [CrossRef] [PubMed]
48. Kumar Nagarajan, S.; Babu, S.; Sohn, H.; Devaraju, P.; Madhavan, T. Toward a Better Understanding of the Interaction between Somatostatin Receptor 2 and Its Ligands: A Structural Characterization Study Using Molecular Dynamics and Conceptual Density Functional Theory. *J. Biomol. Struct. Dyn.* **2019**, *37*, 3081–3102. [CrossRef]
49. Lamberts, S.W.; van der Lely, A.J.; de Herder, W.W.; Hofland, L.J. Octreotide. *N. Engl. J. Med.* **1996**, *334*, 246–254. [CrossRef]
50. Liu, S.; Tang, C.; Ho, B.; Ankersen, M.; Stidsen, C.E.; Crider, A.M. Nonpeptide Somatostatin Agonists with Sst4 Selectivity: Synthesis and Structure-Activity Relationships of Thioureas. *J. Med. Chem.* **1998**, *41*, 4693–4705. [CrossRef]
51. Liu, Z.; Crider, A.M.; Ansbro, D.; Hayes, C.; Kontoyianni, M. A Structure-Based Approach to Understanding Somatostatin Receptor-4 Agonism (Sst4). *J. Chem. Inf. Model.* **2012**, *52*, 171–186. [CrossRef]
52. Nagarajan, S.K.; Babu, S.; Madhavan, T. Theoretical Analysis of Somatostatin Receptor 5 with Antagonists and Agonists for the Treatment of Neuroendocrine Tumors. *Mol. Divers.* **2017**, *21*, 367–384. [CrossRef]
53. Negi, A.; Zhou, J.; Sweeney, S.; Murphy, P.V. Ligand Design for Somatostatin Receptor Isoforms 4 and 5. *Eur. J. Med. Chem.* **2019**, *163*, 148–159. [CrossRef]
54. Veber, D.F.; Freidlinger, R.M.; Perlow, D.S.; Paleveda, W.J.; Holly, F.W.; Strachan, R.G.; Nutt, R.F.; Arison, B.H.; Homnick, C.; Randall, W.C.; et al. A Potent Cyclic Hexapeptide Analogue of Somatostatin. *Nature* **1981**, *292*, 55–58. [CrossRef]
55. Veber, D.F.; Saperstein, R.; Nutt, R.F.; Freidinger, R.M.; Brady, S.F.; Curley, P.; Perlow, D.S.; Paleveda, W.J.; Colton, C.D.; Zacchei, A.G.; et al. A Super Active Cyclic Hexapeptide Analog of Somatostatin. *Life Sci.* **1984**, *34*, 1371–1378. [CrossRef]
56. Chen, L.; Hoeger, C.; Rivier, J.; Fitzpatrick, V.D.; Vandlen, R.L.; Tashjian, A.H. Structural Basis for the Binding Specificity of a SSTR1-Selective Analog of Somatostatin. *Biochem. Biophys. Res. Commun.* **1999**, *258*, 689–694. [CrossRef] [PubMed]
57. Nehring, R.B.; Meyerhof, W.; Richter, D. Aspartic Acid Residue 124 in the Third Transmembrane Domain of the Somatostatin Receptor Subtype 3 Is Essential for Somatostatin-14 Binding. *DNA Cell. Biol.* **1995**, *14*, 939–944. [CrossRef] [PubMed]
58. Greenwood, M.T.; Hukovic, N.; Kumar, U.; Panetta, R.; Hjorth, S.A.; Srikant, C.B.; Patel, Y.C. Ligand Binding Pocket of the Human Somatostatin Receptor 5: Mutational Analysis of the Extracellular Domains. *Mol. Pharm.* **1997**, *52*, 807–814. [CrossRef]
59. Daryaei, I.; Sandoval, K.; Witt, K.; Kontoyianni, M.; Michael Crider, A. Discovery of a 3,4,5-Trisubstituted-1,2,4-Triazole Agonist with High Affinity and Selectivity at the Somatostatin Subtype-4 (Sst4) Receptor. *Medchemcomm* **2018**, *9*, 2083–2090. [CrossRef] [PubMed]
60. Robertson, M.J.; Meyerowitz, J.G.; Panova, O.; Borrelli, K.; Skiniotis, G. Plasticity in Ligand Recognition at Somatostatin Receptors. *Nat. Struct. Mol. Biol.* **2022**, *29*, 210–217. [CrossRef] [PubMed]
61. Kántás, B.; Szőke, É.; Börzsei, R.; Bánhegyi, P.; Asghar, J.; Hudhud, L.; Steib, A.; Hunyady, Á.; Horváth, Á.; Kecskés, A.; et al. In Silico, In Vitro and In Vivo Pharmacodynamic Characterization of Novel Analgesic Drug Candidate Somatostatin SST4 Receptor Agonists. *Front Pharm.* **2021**, *11*, 601887. [CrossRef]
62. Nagarajan, S.K.; Babu, S.; Sohn, H.; Madhavan, T. Molecular-Level Understanding of the Somatostatin Receptor 1 (SSTR1)-Ligand Binding: A Structural Biology Study Based on Computational Methods. *ACS Omega* **2020**, *5*, 21145–21161. [CrossRef]
63. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
64. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
65. Bálint, M.; Jeszenői, N.; Horváth, I.; van der Spoel, D.; Hetényi, C. Systematic Exploration of Multiple Drug Binding Sites. *J. Cheminform.* **2017**, *9*, 65. [CrossRef] [PubMed]
66. Bálint, M.; Horváth, I.; Mészáros, N.; Hetényi, C. Towards Unraveling the Histone Code by Fragment Blind Docking. *Int. J. Mol. Sci* **2019**, *20*, 422. [CrossRef] [PubMed]
67. Bauer, W.; Briner, U.; Doepfner, W.; Haller, R.; Huguenin, R.; Marbach, P.; Petcher, T.J.; Pless, J. SMS 201–995: A Very Potent and Selective Octapeptide Analogue of Somatostatin with Prolonged Action. *Life Sci.* **1982**, *31*, 1133–1140. [CrossRef]

68. Hallenga, K.; Binst, G.V.; Scarso, A.; Michel, A.; Knappenberg, M.; Dremier, C.; Brison, J.; Dirkx, J. The Conformational Properties of the Peptide Hormone Somatostatin (III). *FEBS Lett.* **1980**, *119*, 47–52. [CrossRef]
69. Knappenberg, M.; Michel, A.; Scarso, A.; Brison, J.; Zanen, J.; Hallenga, K.; Deschrijver, P.; Van Binst, G. The Conformational Properties of Somatostatin IV. The Conformers Contributing to the Conformational Equilibrium of Somatostatin in Aqueous Solution as Found by Semi-Empirical Energy Calculations and High-Resolution NMR Experiments. *Biochim. Et Biophys. Acta (BBA)—Protein Struct. Mol. Enzymol.* **1982**, *700*, 229–246. [CrossRef]
70. Simon, Á.; Czajlik, A.; Perczel, A.; Kéri, G.; Nyikos, L.; Emri, Z.; Kardos, J. Binding Crevice for TT-232 in a Homology Model of Type 1 Somatostatin Receptor. *Biochem. Biophys. Res. Commun.* **2004**, *316*, 1059–1064. [CrossRef]
71. Šali, A.; Blundell, T.L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815. [CrossRef]
72. Strnad, J.; Hadcock, J.R. Identification of a Critical Aspartate Residue in Transmembrane Domain Three Necessary for the Binding of Somatostatin to the Somatostatin Receptor SSTR2. *Biochem. Biophys. Res. Commun.* **1995**, *216*, 913–921. [CrossRef]
73. Pernomian, L.; Gomes, M.S.; de Paula da Silva, C.H.T.; Rosa, J.M.C. Reverse Induced Fit-Driven MAS-Downstream Transduction: Looking for Metabotropic Agonists. *Curr. Med. Chem.* **2017**, *24*, 4360–4367. [CrossRef]
74. Van Regenmortel, M.H. Transcending the Structuralist Paradigm in Immunology-Affinity and Biological Activity Rather than Purely Structural Considerations Should Guide the Design of Synthetic Peptide Epitopes. *Biomed Pept. Proteins Nucleic Acids* **1995**, *1*, 109–116.
75. Ye, L.; Van Eps, N.; Zimmer, M.; Ernst, O.P.; Prosser, R.S. Activation of the A2A Adenosine G-Protein-Coupled Receptor by Conformational Selection. *Nature* **2016**, *533*, 265–268. [CrossRef]
76. Rastelli, G.; Pacchioni, S.; Sirawaraporn, W.; Sirawaraporn, R.; Parenti, M.D.; Ferrari, A.M. Docking and Database Screening Reveal New Classes of Plasmodium Falciparum Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **2003**, *46*, 2834–2845. [CrossRef] [PubMed]
77. Zhou, Y.; Hussain, M.; Kuang, G.; Zhang, J.; Tu, Y. Mechanistic Insights into Peptide and Ligand Binding of the ATAD2-Bromodomain via Atomistic Simulations Disclosing a Role of Induced Fit and Conformational Selection. *Phys. Chem. Chem. Phys.* **2018**, *20*, 23222–23232. [CrossRef] [PubMed]
78. Zsidó, B.Z.; Hetényi, C. The Role of Water in Ligand Binding. *Curr. Opin. Struct. Biol.* **2021**, *67*, 1–8. [CrossRef] [PubMed]
79. Sayers, E.W.; Bolton, E.E.; Brister, J.R.; Canese, K.; Chan, J.; Comeau, D.C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2022**, *50*, D20–D26. [CrossRef]
80. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
81. Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105*, 9954–9960. [CrossRef]
82. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef]
83. Hetényi, C.; van der Spoel, D. Toward Prediction of Functional Protein Pockets Using Blind Docking and Pocket Search Algorithms. *Protein Sci.* **2011**, *20*, 880–893. [CrossRef]
84. Horváth, I.; Jeszenői, N.; Bálint, M.; Paragi, G.; Hetényi, C. A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level. *Int. J. Mol. Sci.* **2019**, *20*, 4384. [CrossRef]
85. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins* **2010**, *78*, 1950–1958. [CrossRef]
86. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations IV: Van Der Waals Parameterization. *J. Phys. Chem. B* **2012**, *116*, 7088–7101. [CrossRef] [PubMed]
87. Mehler, E.L.; Solmajer, T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Eng.* **1991**, *4*, 903–910. [CrossRef] [PubMed]
88. Madeira, F.; Pearce, M.; Tivey, A.R.N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and Sequence Analysis Tools Services from EMBL-EBI in 2022. *Nucleic Acids Res.* **2022**, gkac240. [CrossRef] [PubMed]

hetenyi.csaba_83_23

hetenyi.csaba_83_23

D14

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Structural Biology

The role of water in ligand binding

Balázs Zoltán Zsidó and Csaba Hetényi

Exploration of the complex modulatory role of water in ligand–target binding is a current challenge of drug design. This review reports on recent advances of prediction of water structure and function in the context of ligand engineering. The surveyed theoretical approaches showed remarkable progress in the past years. Beyond complementing experiments, they also supplied unmeasurable data. For example, thermodynamic calculations improved ligand binding by the selection of certain water molecules for structural replacement. Molecular dynamics and explicit solvent models remained indispensable to achieve precise results. Topographical analyses of hydration networks proved useful for the prediction of the stabilizing role of interconnected water clusters mediating target–ligand interactions.

Address
Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary

Corresponding author: Hetényi, Csaba (hetenyi.csaba@pte.hu)

Current Opinion in Structural Biology 2021, **67**:1–8

This review comes from a themed issue on **Theory and simulation/ computational methods**

Edited by **Amarda Shehu** and **David van der Spoel**

For a complete overview see the Issue and the Editorial

Available online 14th September 2020

https://doi.org/10.1016/j.sbi.2020.08.002

Introduction

Water plays various roles on both macroscopic [1] and microscopic [2–4] stages of life. The present review focuses on the microscopic roles of water during the binding of a ligand to a target molecule. The precise understanding and prediction of ligand binding are essential in drug design projects. Ligands possess various sizes ranging between small organic compounds [5,6], and large proteins [7,8]. Water molecules mediate the binding of ligands of any sizes, and can be sorted roughly into four functional categories [2,4,9,10] (Figure 1).

Experimental determination of water positions requires atomic resolution techniques. A number of permanent limitations of experimental structure determination [11] impose a challenge on the elucidation of water function in biological complexes. Drug designers answer this 'hydration challenge' with the help of computational approaches surveyed in our present review. We focus mainly on the results of the past two years concerning structure, binding affinity, and networking roles of water in ligand binding.

Water structure

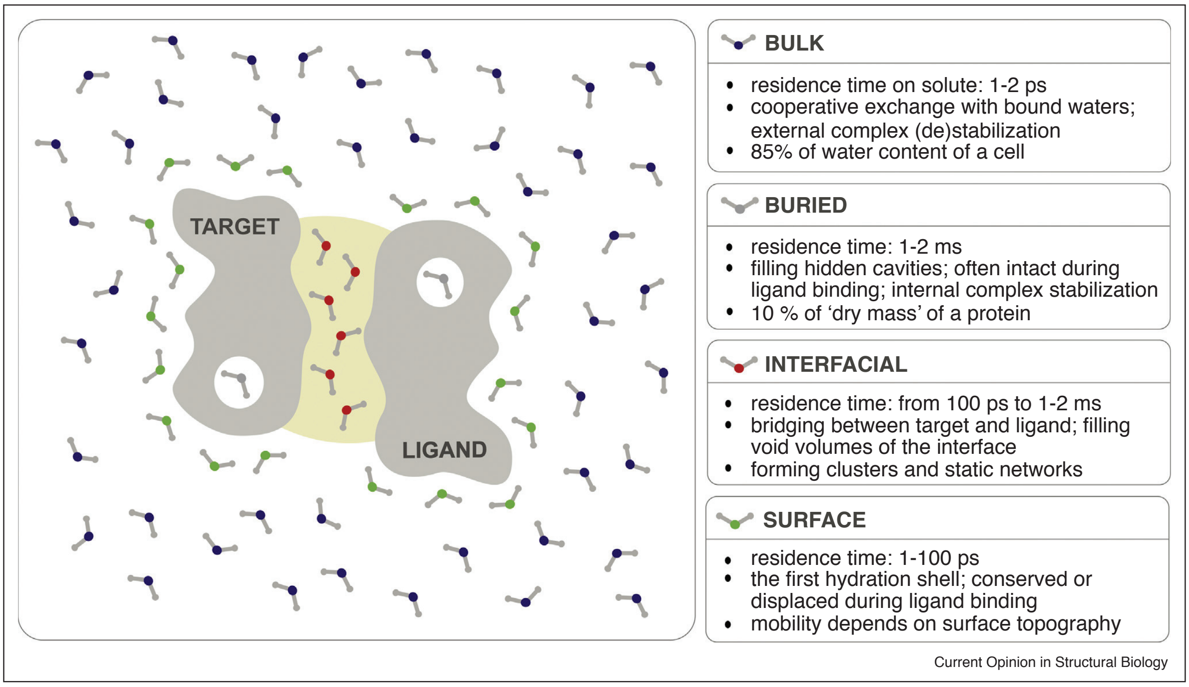
Biomolecular crystallography is the primary experimental source [12,13] of atomic resolution structures of target–ligand complexes. There is a continuous development of X-ray [14] and joint neutron [15] crystallographic methods. Promising combined methods were also introduced [16] with quantum chemical refinements of experimental structures. However, the determination of water positions remains an Achilles' heel [11,17] of crystallography. It is also difficult to assess the quality of assigned water positions. An analysis [18] of 2.3 million experimental water positions concluded that high resolution of a system does not guarantee proper assignation of the hydration structure.

The experimental limitations have motivated the development and application of dynamic and static computational methods for the prediction of water molecules affecting (Figure 1) ligand binding. Dynamic methods supply water positions by clustering snapshots of short molecular dynamics (MD) simulations of explicit water molecules hydrating the solute (target, ligand, or their complex). Static methods use knowledge-based grid maps or geometric rules to build up the hydration structure around a given solute.

Interfacial water molecules can be captured at high precision as they are strongly bound in a relatively tight and buried crevice between the target and the ligand (Figure 1). In a recent report [19], a commercial, dynamic method WaterMap [20] found 90 % of the 41 crystallographic water positions in the interfaces of bromodomain targets and aromatic ligands at a 1.5 Å match level. An open-source software MobyWat [17] showed the same performance on 344 interfacial water molecules in various complexes of peptide and protein ligands [11]. A geometry-based method WarPP [21••] applies an iterative shifting-clustering algorithm. WarPP was validated on almost 20 000 experimental water positions of protein–ligand interfaces of 1500 complexes, and showed a success rate of the above dynamic methods. Other research groups also developed new static approaches, like HydraMap [22] and Splash'Em [23].

The determination of surface waters (Figure 1) is slightly more demanding. An analysis using the EDIA (Electron

Figure 1



BULK

- residence time on solute: 1-2 ps
- cooperative exchange with bound waters; external complex (de)stabilization
- 85% of water content of a cell

BURIED

- residence time: 1-2 ms
- filling hidden cavities; often intact during ligand binding; internal complex stabilization
- 10 % of ‘dry mass’ of a protein

INTERFACIAL

- residence time: from 100 ps to 1-2 ms
- bridging between target and ligand; filling void volumes of the interface
- forming clusters and static networks

SURFACE

- residence time: 1-100 ps
- the first hydration shell; conserved or displaced during ligand binding
- mobility depends on surface topography

Current Opinion in Structural Biology

Functional categories of water in ligand binding.

Density for Individual Atoms) index showed [18] that 90% of insufficiently resolved crystallographic water molecules are positioned on the surface. However, there are conserved surface water molecules of low B-factors which can be captured relatively easily. They remain bound to the target surface during ligand binding [17,24], and make up 77% of bridging waters between the target and the ligand according to an MD study [25]. Other, mobile surface waters of uncertain positions (higher B-factors) cause a plausible drop in overall success rates of prediction methods from 90% (see the previous paragraph) to ca. 80% [17] calculated for all surface waters.

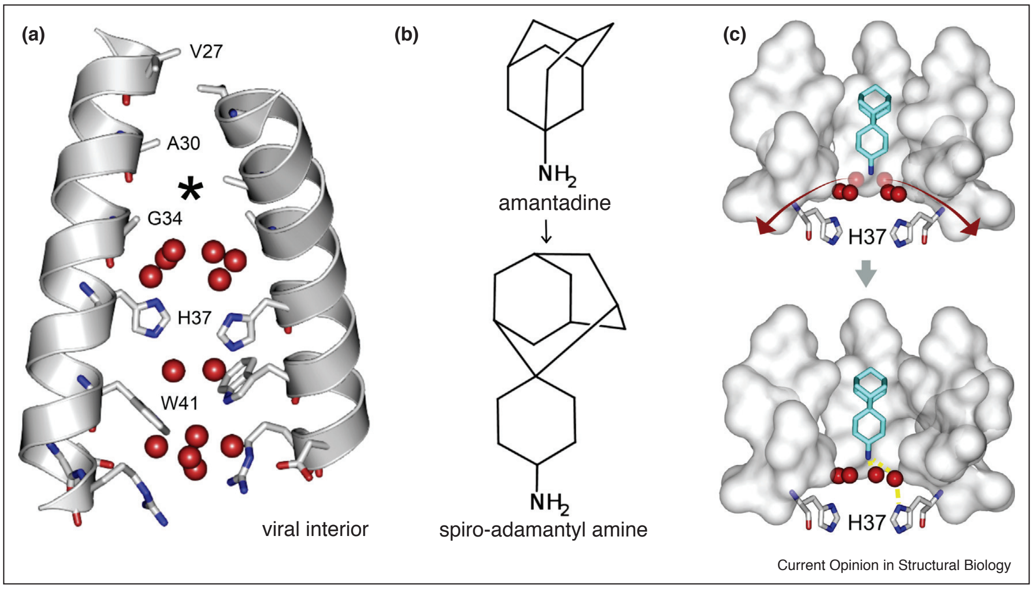
Buried water molecules (Figure 1) occupy hidden binding pockets with a challenging geometry to predict. A combination of the static 3D-RISM [26] and dynamic WATsite [27] methods produced [28•] successful predictions. JAL, an explicit solvent MD-based method also managed to compute buried water positions in tumour suppressor protein p53 and a translation initiation factor [29]. The application of MD can be recommended for the difficult cases of buried waters. Dynamic methods have general applicability in all three categories (interface, surface, and buried) discussed above as they take care of both solute–water and water–water interactions and allow cooperative water exchange [10] with the bulk (Figure 1) [11,29], as well. They provide accurate [30•] and reproducible [31] results, and the necessary MD snapshots can be produced in short simulations by high performance, parallelized open-source software [32].

The generation of water structure during the computational docking of a ligand to a target would be an attractive technique for virtual (high throughput) screening [33,34] projects. While there are several promising advances [35,36] of this direct methodology, its full automation remains a challenge. Another (indirect) approach, the comparison of end-points of ligand binding seems fairly manageable by available tools. The above-mentioned methods supply surface and interfacial water positions for the hydration structures of the initial (apo, ligand-free) and the final (holo, ligand-bound) stages, respectively. Pairwise comparisons of holo and apo structures or holo structures with similar ligands [37••] (Figure 2) help the identification of conserved and displaced waters, and optimization of ligand–target interactions (see also next Section).

Contribution of water molecules to binding affinity

The ligand–target binding affinity is expressed as the free energy change of the binding reaction (ΔG_b). The ΔG_b can

Figure 2



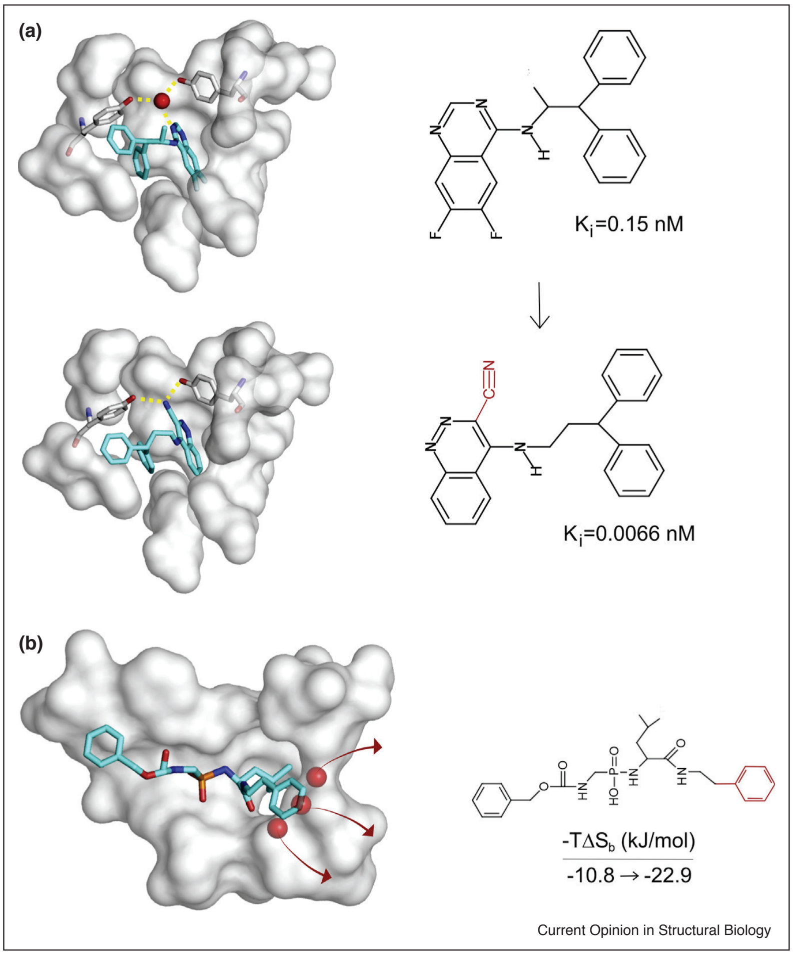
Conserved and displaced water molecules during binding of spiro-adamantyl amine in the influenza M2 transmembrane (TM) proton channel. The structure of the proton channel **(a)** was constructed as a homotetramer of the TM helices (grey cartoon, only a dimer is shown for clarity, PDB code 3lbw) containing residues 22–46. Water molecules and side-chains inside the channel are shown as red spheres, and sticks, respectively. The black asterisk marks the center of the binding pocket of amantadine, a drug in clinical use. Spiro-adamantyl amine preserved **(b)** the main pharmacophores including the amine group and the bulky hydrophobic ring system. Having a larger size than amantadine (b), it displaces some of the water molecules (light red in **(c)**) observed in the amantadine-bound pocket (PDB code 3lbw; at the top in (c)) [37••]. Other water molecules above the H37 side-chains remain conserved in the spiro-adamantyl amine-bound pocket (PDB code 6bmz; at the bottom in (c)) and involved in H-bonding interaction (yellow dashes in (c)) with the amine group of the inhibitor pointing toward the viral interior. To feature the steric conflict with the positions of displaced water molecules, the structure of spiro-adamantyl amine was created in the amantadine-bound pocket (at the top in (c)) by superimposition of PDB structure 6bmz on 3lbw. Programs PyMol [71] and Marvin Sketch [72] were used for drawing of spatial and Lewis structures, respectively.

be engineered via ligand modifications affecting the hydration structure [17,24,30•,34,37••,38–42,43•,44•,45,46,47•,48]. For example, the target–ligand complex can be stabilized by inserting H-bonding functional groups that interact with or replace (Figure 3a) interfacial water molecules resulting in a favourable contribution to binding enthalpy (ΔH_b) [43•]. New functional groups may increase the ligand's ability to expel surface waters into the bulk (Figure 3b) increasing binding entropy (ΔS_b) [46] and affinity.

The determination of thermodynamic stability and the prediction of the contributions of individual water positions to binding affinity (Figure 3) is a key to ligand design. However, experimental methods like isothermal titration calorimetry (ITC) cannot partition ΔG_b values into individual contributions per water molecule [18,39,49]. Theoretical methods with explicit solvent models help to overcome this limitation. For example, the inhomogeneous fluid solvation theory (IFST) [50] has gained application for thermodynamic characterization of individual hydration sites. IFST with explicit solvent MD calculations was used [47•] to investigate various modifications of ligand structures that led to the displacement (see e.g. Figure 2) of binding site water molecules. The IFST calculations were useful [47•] in guiding water replacements in lead optimization but did not improve the prediction of the corresponding differences in ΔG_b. Such differences of ΔG_b were successfully correlated with solvent displacement on sets of similar ligands in another study [51] presenting new functionals for grid inhomogeneous solvation theory (GIST) [52].

Nevertheless, there is some controversy in the literature on the usefulness of the above solvation theories for the prediction of ΔG_b. Initial evaluations of IFST (in WaterMap [20]), and GIST [51] performed better for prediction of ΔG_b than other calculators based on implicit solvent models [53]. Indeed, GBSA (PBSA) methods or their combination with explicit water molecules showed limited [30•] or occasional [35] success for ΔG_b calculations, due to their theoretical limitations [54,55]. However,

Figure 3



Water structure helps enthalpic and entropic optimization of ligands.
The binding free energy (ΔG_b) of ligand molecules can be optimized by modification of the enthalpic (ΔH_b) or entropic (ΔS_b) contributions according to $\Delta G_b = \Delta H_b\text{-}T\Delta S_b$ (where T is the thermodynamic temperature). Targets and ligands are shown in surface and stick representations, respectively. **(a)** Ligand (6,7-difluoro-quinazolin-4-yl)-(1-methyl-2,2-diphenyl-ethyl)-amine shows a good, subnanomolar binding to scytalone-dehydratase stabilized by a bridging water molecule. The isosteric displacement of the bridging water molecule with a nitrile group (red in the Lewis structure) further lowered the K_i and contributed to the enthalpic optimization of ligand binding. The direct hydrogen bonding between the nitrile group of the ligand and the tyrosine residues of the target provides a stronger target–ligand contact (more negative ΔH_b) than the indirect hydrogen bond system with bridging water molecule [43•]. **(b)** The growing of ligand UBTLN46 by addition of a larger phenyl group (red in the Lewis structure), resulted in the displacement of water molecules from the binding pocket of thermolysin. The leaving water molecules increased ΔS_b [46] which resulted in a more favourable negative contribution to ΔG_b.

other reports [19,30•] comparing grid-based SZMAP [56], WaterFLAP [57], 3D-RISM [26], and WaterMap did not show significant improvement of ΔG_b calculations with IFST. Assessment of the general applicability of solvation theories in ΔG_b calculations will require additional validations on large and diverse test sets. At present, the above methods seem more useful [30•] for selecting key waters for planned ligand modifications (Figure 3).

An increasing number of studies suggest that the use of appropriately positioned explicit water molecules is required in binding thermodynamics calculations. For example, relative ΔG_b calculations on small organic ligands showed that the free energy perturbation method [48] is very sensitive to the choice of initial hydration structure possibly due to water molecules trapped in and/or insufficiently filling buried cavities. Another study [58] also involved large peptide ligands and applied a combination of predicted, explicit interfacial water molecules with the COnductor-like Screening MOdel (COSMO) [59] in end-point calculations. The combined water model resulted in good correlations with experimental ΔH_b values at a PM7 semi-empirical quantum mechanics level.

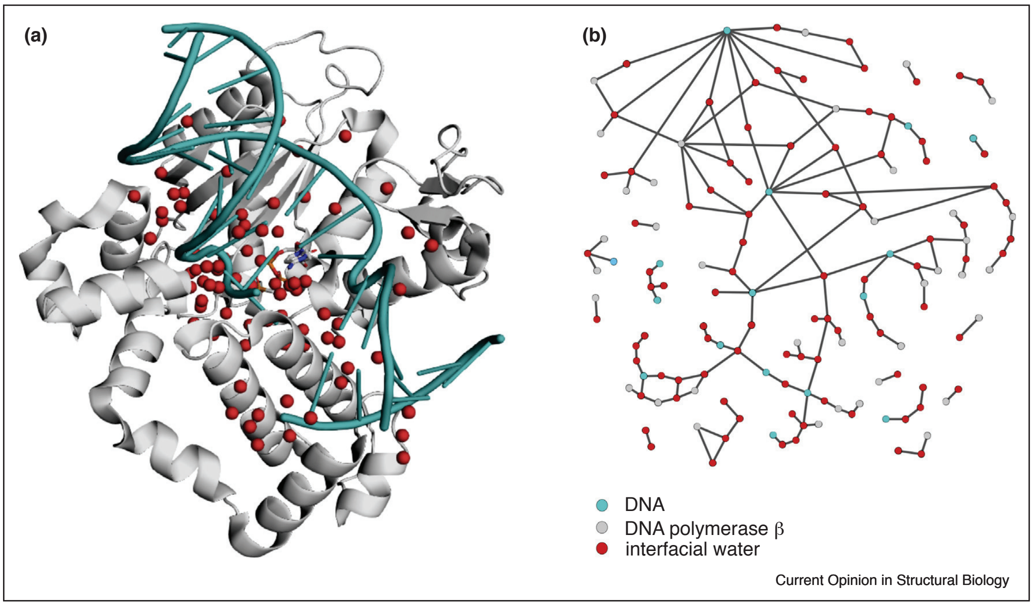
The mobility of water networks

In addition to their individual contributions (see previous Sections), water molecules often participate in molecular networks at various locations (Figure 1). Exploration of networking of waters may open a new pathway of ligand design likewise to discoveries in other complex (data) systems [60]. In the cases of small ligands [61•], network changes may be discovered by manual comparisons of the end-points (Figure 2). In the case of large water networks of for example, protein–protein or protein–DNA [62] (Figure 4) complexes, the comparisons should be automatized using graph representations [11,63•].

However, there are relatively few methods offering graph theoretical approaches of hydration networks. Brysbaert *et al.* analyzed [63•] the changes of residue interaction networks (RIN) of interfaces of protein complexes using RINspector [64]). Adding water molecules to the RIN graphs helped the identification of interface residues involved in the water-mediated binding of the protein partners. The mobility of water nodes was used to distinguish between static and dynamic hydration networks in another graph-based study [11]. Static networks of low mobility contain numerous solute–water and water–water H-bonds stabilizing the target–ligand complex [41,65,66]. Dynamic networks contribute to complex destabilization [11,61•] and binding diverse ligands [67] via cooperative water exchange mechanisms [10] with the bulk.

Ligand binding can be fine-tuned by surrounding water networks. The stabilizing role of static networks was demonstrated by the analysis of the changes in hydration graphs [11] of a histone-chaperone complex [68] following amino acid mutations in the interface region. A similar networking situation was explored [44•] in the case of mutated protein–glycan complexes. The study showed the dominating contribution of a static hydration network of a few, core water molecules to binding thermodynamic signatures. Similarly, only a few stable water positions were identified in ligand binding pockets of G-protein coupled receptors (GPCRs) [67]. Although the conserved GPCR binding pockets are filled mostly by mobile

Figure 4



Current Opinion in Structural Biology

The complexity of the interfacial hydration network of the DNA polymerase β (cyan) in complex with DNA (light blue cartoon).
A small molecule inhibitor dCMPP(CH2)P and crystallographic (PDB code 6w2m) water positions are shown **(a)** as sticks red spheres, respectively. The large polymerase-DNA interface holds numerous water molecules mediating between the binding partners. The two-dimensional graph of the interfacial hydration network **(b)** shows a high complexity due to several water-solute and water–water connections. The graph representation allows quick visualization, automated analysis and comparisons of complex hydration networks between large macromolecules. The graph in panel (b) was generated from the PDB structure using the NetDraw function of program MobyWat [11] with a 3 Å distance cut-off, and visualized by Gephi [73].

waters, the stable waters and conserved water-networks are involved in the binding of structurally diverse ligands.

Reorganization or replacement of water networks can be often observed during ligand binding. Water networks of the binding pocket of human carbonic anhydrase II show fast μs-time-scale dynamics according to NMR measurements combined with MD simulations [69]. The effect of an inhibitor ligand on the disruption of such intra-pocket water–water networking and enzymatic activity was analyzed [69]. Another study combining crystallography and MD [37••] showed how amantadyl-amine ligands disrupt key segments of water networks in the influenza A virus matrix 2 proton channel. An earlier MD study [70] of the same system also suggested the replacement of water clusters for the design of new ligands (see also Figure 2) with an ammonium group mimicking the effect of oxonium ions in proton transport. The effect of the dynamic reorganization of water networks on ligand binding affinity was quantified [61•] involving a crystallographic structure set of Haemophilus influenzae virulence protein SiaP mutants in complex with sialic acid ligands. Relative ΔG_b values were calculated [61•] using B-factors of water molecules involved in the interaction network around the ligand. Although the approach is probably applicable only for similar complexes, further tests with experimental data or extension using calculated B-factors might be interesting.

Some of the above studies [44•,68] report on mutations of target amino acids not directly interacting with the ligand. In these examples, mutations affect ligand binding indirectly, via concerted changes in the interfacial water network. Exploration of such ‘hidden’ features of a complex (Figure 4b) hydration network is a key to the prediction of binding affinity of large ligands.

Conclusions

Drug designers often complain of incomplete experimental hydration structures. They could make good use of quantifying thermodynamic contributions of individual water molecules to the overall binding process which cannot be supplied by experiments. Computational techniques have supplied solutions to these requests and performed well in the calculation of the water structure of biomolecules participating in target–ligand binding. Ligand design has benefited from structure-based thermodynamic calculations comparing hydration structures of the apo and holo stages. Molecular dynamics and explicit solvent models have become the gold standard of simulations accounting for water–water interactions often observed in extended hydration networks of for example, protein ligands. Like any other approach, the surveyed theoretical methods and applications have their technical limitations which can be overcome in the not-too-distant future. Potential improvements of polarizable water models (force fields), new quantum mechanical applications, and topographical analyses of water networks will further increase the efficiency of prediction of the role of water in ligand binding.

Conflict of interest statement

Nothing declared.

Acknowledgements

This work was funded by the Hungarian National Research, Development and Innovation Office (K123836), 2017-1.2.1-NKP-2017-00002 (NAP-2; Chronic Pain Research Group), EFOP-3.6.1.-16-2016-0004 and GINOP 2.3.2-15-2016-00050 ‘PEPSYS’. This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. The work was supported by the ÚNKP-20-3 and ÚNKP-20-5New National Excellence Program of the Ministry for Innovation and Technology.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as

- • of special interest
- •• of outstanding interest

1. Brini E, Fennell CJ, Fernandez-Serra M, Hribar-Lee B, Lukšič M, Dill KA: **How water’s properties are encoded in its molecular structure and energies**. *Chem Rev* 2017, **117**:12385-12414.
2. Ball P: **Water as an active constituent in cell biology**. *Chem Rev* 2008, **108**:74-108.
3. Bellissent-Funel MC, Hassanali A, Havenith M, Henchman R, Pohl P, Sterpone F, Van Der Spoel D, Xu Y, Garcia AE: **Water determines the structure and dynamics of proteins**. *Chem Rev* 2016, **116**:7673-7697.
4. Laage D, Elsaesser T, Hynes JT: **Water dynamics in the hydration shells of biomolecules**. *Chem Rev* 2017, **117**:10694-10725.
5. Trisciuzzi D, Nicolotti O, Miteva MA, Villoutreix BO: **Analysis of solvent-exposed and buried co-crystallized ligands: a case study to support the design of novel protein–protein interaction inhibitors**. *Drug Discov Today* 2019, **24**:551-559.
6. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, Fisher J, Jansen JM, Duca JS, Rush TS *et al.*: **Rethinking drug design in the artificial intelligence era**. *Nat Rev Drug Discov* 2020, **19**:353-364.
7. Caskey M, Klein F, Nussenzweig MC: **Broadly neutralizing anti-HIV-1 monoclonal antibodies in the clinic**. *Nat Med* 2019, **25**:547-553.
8. Navarro S, Ventura S: **Computational re-design of protein structures to improve solubility**. *Expert Opin Drug Discov* 2019, **14**:1077-1088.
9. Stephens AD, Kaminski Schierle GS: **The role of water in amyloid aggregation kinetics**. *Curr Opin Struct Biol* 2019, **58**:115-123.
10. Halle B, Helliwell JR, Kornyshev A, Engberts JBFN: **Protein hydration dynamics in solution: a critical survey**. *Philos Trans R Soc B Biol Sci* 2004, **359**:1207-1224.
11. Jeszenői N, Bálint M, Horváth I, Van Der Spoel D, Hetényi C: **Exploration of interfacial hydration networks of target-ligand complexes**. *J Chem Inf Model* 2016, **56**:148-158.
12. Susannah S, Ando N: **X-rays in the cryo-EM era: structural biology’s dynamic future**. *Biochemistry* 2018, **57**:277-285.
13. Zsidó BZ, Hetényi C: **Molecular structure, binding affinity, and biological activity in the epigenome**. *Int J Mol Sci* 2020, **21**:1-40.
14. Weichenberger CX, Afonine PV, Kantardjieff K, Rupp B: **The solvent component of macromolecular crystals**. *Acta Crystallogr Sect D Biol Crystallogr* 2015, **71**:1023-1038.

15. Koruza K, Mahon BP, Blakeley MP, Ostermann A, Schrader TE, McKenna R, Knecht W, Fisher SZ: **Using neutron crystallography to elucidate the basis of selective inhibition of carbonic anhydrase by saccharin and a derivative**. *J Struct Biol* 2019, **205**:147-154.

16. Malaspina LA, Wieduwilt EK, Bergmann J, Kleemiss F, Meyer B, Ruiz-López MF, Pal R, Hupf E, Beckmann J, Piltz RO *et al.*: **Fast and accurate quantum crystallography: from small to large, from light to heavy**. *J Phys Chem Lett* 2019, **10**:6973-6982.

17. Jeszenői N, Horváth I, Bálint M, Van Der Spoel D, Hetényi C: **Mobility-based prediction of hydration structures of protein surfaces**. *Bioinformatics* 2015, **31**:1959-1965.

18. Nittinger E, Schneider N, Lange G, Rarey M: **Evidence of water molecules - a statistical evaluation of water molecules based on electron density**. *J Chem Inf Model* 2015, **55**:771-783.

19. Nittinger E, Gibbons P, Eigenbrot C, Davies DR, Maurer B, Yu CL, Kiefer JR, Kuglstatter A, Murray J, Ortwine DF *et al.*: **Water molecules in protein–ligand interfaces. Evaluation of software tools and SAR comparison**. *J Comput Aided Mol Des* 2019, **33**:307-330.

20. Abel R, Young T, Farid R, Berne BJ, Friesner RA: **Role of the active-site solvent in the thermodynamics of factor Xa ligand binding**. *J Am Chem Soc* 2008, **130**:2817-2831.

21. •• Nittinger E, Flachsenberg F, Bietz S, Lange G, Klein R, Rarey M: **Placement of water molecules in protein structures: from large-scale evaluations to single-case examples**. *J Chem Inf Model* 2018, **58**:1625-1637.
A thoroughly validated, geometry-based prediction tool of interfacial waters with iterative shifting cycles for positional refinements.

22. Li Y, Gao YD, Holloway MK, Wang R: **Prediction of the favorable hydration sites in a protein binding pocket and its application to scoring function formulation**. *J Chem Inf Model* 2020 http://dx.doi.org/10.1021/acs.jcim.9b00619.

23. Wei W, Luo J, Waldispühl J, Moitessier N: **Predicting positions of bridging water molecules in nucleic acid-ligand complexes**. *J Chem Inf Model* 2019, **59**:2941-2951.

24. García-Sosa AT, Mancera RL, Dean PM: **WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes**. *J Mol Model* 2003, **9**:172-182.

25. Rudling A, Orro A, Carlsson J: **Prediction of ordered water molecules in protein binding sites from molecular dynamics simulations: the impact of ligand binding on hydration networks**. *J Chem Inf Model* 2018, **58**:350-361.

26. Truchon JF, Pettitt BM, Labute P: **A cavity corrected 3D-RISM functional for accurate solvation free energies**. *J Chem Theory Comput* 2014, **10**:934-941.

27. Hu B, Lill AM: **WATsite: hydration site prediction program with PyMOL interface**. *J Comput Chem* 2014, **35**:1255-1260.

28. • Masters MR, Mahmoud AH, Yang Y, Lill MA: **Efficient and accurate hydration site profiling for enclosed binding sites**. *J Chem Inf Model* 2018, **58**:2183-2188.
A hybrid method with a dynamic step for prediction of water positions in occluded binding sites.

29. Pradhan MR, Nguyen MN, Kannan S, Fox SJ, Kwoh CK, Lane DP, Verma CS: **Characterization of hydration properties in structural ensembles of biomolecules**. *J Chem Inf Model* 2019, **59**:3316-3329.

30. • Bucher D, Stouten P, Triballeau N: **Shedding light on important waters for drug design: simulations versus grid-based methods**. *J Chem Inf Model* 2018, **58**:692-699.
An expert account with realistic conclusions comparing solvent mapping tools.

31. Jeszenői N, Schilli G, Bálint M, Horváth I, Hetényi C: **Analysis of the influence of simulation parameters on biomolecule-linked water networks**. *J Mol Graph Model* 2018, **82**:117-128.

32. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: **GROMACS: fast, flexible, and free**. *J Comput Chem* 2005, **26**:1701-1718.

33. Zhong H, Wang Z, Wang X, Liu H, Li D, Liu H, Yao X, Hou T: **Importance of a crystalline water network in docking-based virtual screening: a case study of BRD4**. *Phys Chem Chem Phys* 2019, **21**:25276-25289.

34. Hu X, Maffucci I, Contini A: **Advances in the treatment of explicit water molecules in docking and binding free energy calculations**. *Curr Med Chem* 2019, **26**:7598-7622.

35. Maffucci I, Hu X, Fumagalli V, Contini A: **An efficient implementation of the Nwat-MMGBSA method to rescore docking results in medium-throughput virtual screenings**. *Front Chem* 2018, **6**:1-14.

36. Lu Q, Qi LW, Liu J: **Improving protein-ligand binding prediction by considering the bridging water molecules in autodock**. *J Theor Comput Chem* 2019, **18**:1-12.

37. •• Thomaston JL, Polizzi NF, Konstantinidi A, Wang J, Kolocouris A, Degrado WF: **Inhibitors of the M2 proton channel engage and disrupt transmembrane networks of hydrogen-bonded waters**. *J Am Chem Soc* 2018, **140**:15219-15226.
A textbook example on exploration of the role of water networks in inhibitor design.

38. Bhattarai S, Pippel J, Scaletti E, Idris R, Freundlieb M, Rolshoven G, Renn C, Lee SY, Abdelrahman A, Zimmermann H *et al.*: **2-substituted α,β-methylene-ADP derivatives: potent competitive ecto-5′-nucleotidase (CD73) inhibitors with variable binding modes**. *J Med Chem* 2020, **63**:2941-2957.

39. Geschwindner S, Ulander J: **The current impact of water thermodynamics for small-molecule drug discovery**. *Expert Opin Drug Discov* 2019, **14**:1221-1225.

40. Bodnarchuk MS: **Water, water, everywhere . . . it's time to stop and think**. *Drug Discov Today* 2016, **21**:1139-1146.

41. Maurer M, Oostenbrink C: **Water in protein hydration and ligand recognition**. *J Mol Recognit* 2019, **32**:1-19.

42. Ratkova EL, Dawidowski M, Napolitano V, Dubin G, Fino R, Ostertag MS, Sattler M, Popowicz G, Tetko IV: **Water envelope has a critical impact on the design of protein-protein interaction inhibitors**. *Chem Commun* 2020, **56**:4360-4363.

43. • Chen D, Li Y, Zhao M, Tan W, Li X, Savidge T, Guo W, Fan X: **Effective lead optimization targeting the displacement of bridging receptor-ligand water molecules**. *Phys Chem Chem Phys* 2018, **20**:24399-24407.
An in-depth thermodynamic study on the replacement of bridging water molecules for lead optimization.

44. • Kunstmann S, Gohlke U, Broeker NK, Roske Y, Heinemann U, Santer M, Barbirz S: **Solvent networks tune thermodynamics of oligosaccharide complex formation in an extended protein binding sit**. *J Am Chem Soc* 2018, **140**:10447-10455.
A mobility analysis and identification of static water networks influencing the thermodynamic signature of ligand binding.

45. Li A, Gilson MK: **Protein-ligand binding enthalpies from near-millisecond simulations: analysis of a preorganization paradox**. *J Chem Phys* 2018, **149**:1-15.

46. Krimmer SG, Betz M, Heine A, Klebe G: **Methyl, ethyl, propyl, butyl: futile but not for water, as the correlation of structure and thermodynamic signature shows in a congeneric series of thermolysin inhibitors**. *ChemMedChem* 2014, **9**:833-846.

47. • Wahl J, Smieško M: **Thermodynamic insight into the effects of water displacement and rearrangement upon ligand modifications using molecular dynamics simulations**. *ChemMedChem* 2018, **13**:1325-1335.
A systematic MD study and thermodynamic analysis on the role of the displacement of water molecules in ligand binding.

48. Wahl J, Smieško M: **Assessing the predictive power of relative binding free energy calculations for test cases involving displacement of binding site water molecules**. *J Chem Inf Model* 2019, **59**:754-765.

49. Kairys V, Baranauskiene L, Kazlauskiene M, Matulis D, Kazlauskas E: **Binding affinity in drug design: experimental and computational techniques**. *Expert Opin Drug Discov* 2019, **14**:755-768.

50. Lazaridis T: **Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory**. *J Phys Chem B* 1998, **102**:3531-3541.

51. Hüfner-Wulsdorf T, Klebe G: **Protein-ligand complex solvation thermodynamics: development, parameterization, and testing of GIST-based solvent functionals**. *J Chem Inf Model* 2020, **60**:1409-1423.

52. Balius TE, Fischer M, Stein RM, Adler TB, Nguyen CN, Cruz A, Gilson MK, Kurtzman T, Shoichet BK: **Testing inhomogeneous solvation theory in structure-based ligand discovery**. *Proc Natl Acad Sci U S A* 2017, **114**:6839-6846.

53. Sitkoff D, Sharp KA, Honig B: **Accurate calculation of hydration free energies using macroscopic solvent models**. *J Phys Chem* 1994, **98**:1978-1988.

54. Zhang H, Yin C, Yan H, Van Der Spoel D: **Evaluation of generalized born models for large scale affinity prediction of cyclodextrin host-guest complexes**. *J Chem Inf Model* 2016, **56**:2080-2092.

55. Genheden S, Ryde U: **The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities**. *Expert Opin Drug Discov* 2015, **10**:449-461.

56. Bayden AS, Moustakas DT, Joseph-McCarthy D, Lamb ML: **Evaluating free energies of binding and conservation of crystallographic waters using SZMAP**. *J Chem Inf Model* 2015, **55**:1552-1565.

57. Pastor M, Cruciani G, Watson KA: **A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure - activity relationship analysis**. *J Med Chem* 1997, **40**:4089-4102.

58. Horváth I, Jeszenői N, Bálint M, Paragi G, Hetényi C: **A fragmenting protocol with explicit hydration for calculation of binding enthalpies of target-ligand complexes at a quantum mechanical level**. *Int J Mol Sci* 2019, **20**:4384-4403.

59. Klamt A, Schüürmann G: **COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient**. *J Chem Soc Perkin Trans 2* 1993, **5**:799-805.

60. Cheng F, Kovács IA, Barabási AL: **Network-based prediction of drug combinations**. *Nat Commun* 2019, **10**:1197-1208.

61. • Darby JF, Hopkins AP, Shimizu S, Roberts SM, Brannigan JA, Turkenburg JP, Thomas GH, Hubbard RE, Fischer M: **Water networks can determine the affinity of ligand binding to proteins**. *J Am Chem Soc* 2019, **141**:15818-15826.

A structure-based approach linking mobility of water networks to ligand binding affinity.

62. Batra VK, Wilson SH: **Structure of a DNA polymerase abortive complex with the 8OG:dA base pair at the primer terminus**. *Commun Biol* 2020, **3**:8-11.

63. • Brysbaert G, Blossey R, Lensink MF: **The inclusion of water molecules in residue interaction networks identifies additional central residues**. *Front Mol Biosci* 2018, **5**:1-7.

A timely paper on graph theoretical analysis of residue interaction networks extended with water nodes.

64. Brysbaert G, Lorgouilloux K, Vranken WF, Lensink MF: **RINspector: a cytoscape app for centrality analyses and DynaMine flexibility prediction**. *Bioinformatics* 2018, **34**:294-296.

65. Majewski M, Ruiz-Carmona S, Barril X: **An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder**. *Commun Chem* 2019, **2**:110-118.

66. Schiebel J, Gaspari R, Wulsdorf T, Ngo K, Sohn C, Schrader TE, Cavalli A, Ostermann A, Heine A, Klebe G: **Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes**. *Nat Commun* 2018, **9**:3559-3574.

67. Venkatakrishnan AJ, Ma AK, Fonseca R, Latorraca NR, Kelly B, Betz RM, Asawa C, Kobilka BK, Dror RO: **Diverse GPCRs exhibit conserved water networks for stabilization and activation**. *Proc Natl Acad Sci U S A* 2019, **116**:3288-3293.

68. Elsässer SJ, Huang H, Lewis PW, Chin JW, Allis CD, Patel DJ: **DAXX envelops a histone H3.3-H4 dimer for H3.3-specific recognition**. *Nature* 2012, **491**:560-565.

69. Singh H, Vasa SK, Jangra H, Rovó P, Päslack C, Das CK, Zipse H, Schäfer LV, Linser R: **Fast microsecond dynamics of the protein-water network in the active site of human carbonic anhydrase II studied by solid-state NMR spectroscopy**. *J Am Chem Soc* 2019, **141**:19276-19288.

70. Gianti E, Carnevale V, DeGrado WF, Klein ML, Fiorin G: **Hydrogen-bonded water molecules in the M2 channel of the influenza A virus guide the binding preferences of ammonium-based inhibitors**. *J Phys Chem B* 2015, **119**:1173-1183.

71. DeLano WL: *The PyMOL Molecular Graphics System; Version 2.0*. New York, NY, USA: Schrödinger LLC; 2002.

72. Marvin 19.18., 2019, ChemAxon. (http://www.chemaxon.com).

73. Bastian M, Heymann S, Jacomy M: **Gephi: an open source software for exploring and manipulating networks**. *Int AAAI Conf Weblogs Soc Media* 2009.

D15

International Journal of
Molecular Sciences

MDPI

Review

The Advances and Limitations of the Determination and Applications of Water Structure in Molecular Engineering

Balázs Zoltán Zsidó, Bayartsetseg Bayarsaikhan, Rita Börzsei, Viktor Szél, Violetta Mohos and Csaba Hetényi *

Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary; zsido.balazs@pte.hu (B.Z.Z.); bayartsetseg704@yahoo.com (B.B.); rita.borzsei@aok.pte.hu (R.B.); szel.viktor@pte.hu (V.S.); mohos.violetta@aok.pte.hu (V.M.)
* Correspondence: hetenyi.csaba@pte.hu

Abstract: Water is a key actor of various processes of nature and, therefore, molecular engineering has to take the structural and energetic consequences of hydration into account. While the present review focuses on the target–ligand interactions in drug design, with a focus on biomolecules, these methods and applications can be easily adapted to other fields of the molecular engineering of molecular complexes, including solid hydrates. The review starts with the problems and solutions of the determination of water structures. The experimental approaches and theoretical calculations are summarized, including conceptual classifications. The implementations and applications of water models are featured for the calculation of the binding thermodynamics and computational ligand docking. It is concluded that theoretical approaches not only reproduce or complete experimental water structures, but also provide key information on the contribution of individual water molecules and are indispensable tools in molecular engineering.

Keywords: drug design; docking; crystallography; electron microscopy; solvation; free energy

check for updates

Citation: Zsidó, B.Z.; Bayarsaikhan, B.; Börzsei, R.; Szél, V.; Mohos, V.; Hetényi, C. The Advances and Limitations of the Determination and Applications of Water Structure in Molecular Engineering. *Int. J. Mol. Sci.* **2023**, *24*, 11784. https://doi.org/10.3390/ijms241411784

Academic Editor: Mario Milani

Received: 20 June 2023
Revised: 18 July 2023
Accepted: 20 July 2023
Published: 22 July 2023

1. Introduction

Water is a molecular jolly joker of a living nature. It is a main solvent in bulk solution and cellular interfaces and fills the void spaces in tissues (the mass of the human body is made up of ca. 60% water [1]). Water also acts as an active matrix component and is involved in the stabilization of the biomacromolecules mediating macromolecular interactions, e.g., in signaling pathways [2–5], and in the binding of small molecules to their target structures [6–12]. From a structural point of view, the role of water can be further classified. There are water molecules that form the bulk solvent accounting for 85% of the water content of a cell [13,14], and they might either be exchanged with bound waters or participate in the (de)stabilization of solute complexes. Buried water molecules stabilize solutes internally, giving 10% of the 'dry mass' of proteins. Water molecules also bridge between solute (macro)molecules and fill the void volumes of interaction interfaces [9,13,15]. They can form a hydration shell [16] that is either conserved or displaced upon ligand binding. If hydration shell water molecules are conserved upon ligand binding, they turn into bridges forming a static network of solute–water and water–water hydrogen bonds [13,15,17–24]. Such a static network is characterized by a low mobility and acts by stabilizing complexes. On the other hand, dynamic networks characterized by loosely bound water molecules with a high mobility participate in the complex destabilization or non-selective binding of various ligands.

In molecular engineering, the above structural roles of water can be translated into energetic contributions. For example, in a target–ligand interface (the main stage of drug design), conserved and leaving water molecules can be distinguished upon ligand binding [16,19,25–29]. Conserved waters tend to stay and form bridges in the target–ligand interface and are often referred to as 'happy' waters (Figure 1).

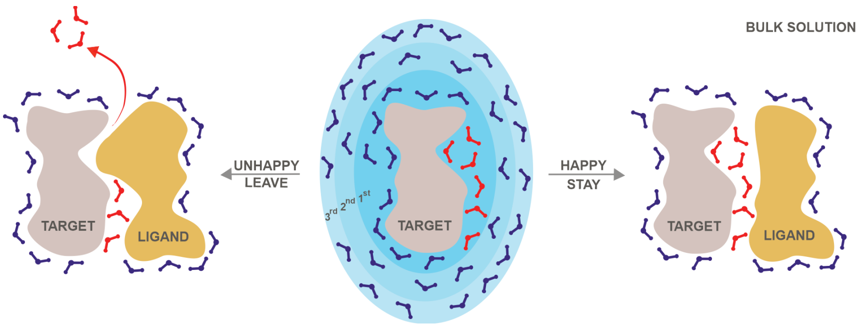


Figure 1. The hydration shells and the possible roles of interface water molecules during ligand binding. The target molecule (grey cloud, in the middle) is covered by hydration shells of surface water molecules (blue sticks), where the fading color of the shells represent the diminishing strength of interaction between the shell (also labelled by a serial number) and the target. Happy interface water molecules (red sticks, on the right) tend to stay, while unhappy water molecules (on the left) are displaced by the ligand (beige cloud) and leave (red arrow) to the bulk solution during the binding process.

There are also 'unhappy' water molecules displaced by the drug molecule during its binding to the target. An unhappy water molecule might offer a possibility for the enthalpic optimization of ligand binding. An indirect, 'unhappy', water-mediated interaction between a ligand and a target might be enthalpically less favorable compared to the direct binding of a ligand to a target amino acid residue after the displacement of an 'unhappy' water molecule [30]. During the displacement of an 'unhappy' water molecule, it moves to the bulk, and this process has a favorable entropic contribution to the free energy change of the binding reaction (ΔG_b) [28]. On the other hand, targeting 'happy' water molecules can be useful, as they often bridge between the target and ligand. A ligand can be optimized to participate in this bridging interaction by adding functional groups with a hydrogen bonding capacity to the drug molecule, providing a favorable enthalpic contribution to the ligand binding, because the additional hydrogen bonding capacity can form hydrogen bonds with 'happy' water molecules, as well as with hydrophilic target amino acids. In drug design, increasing the ligand interactions with happy water molecules and pushing unhappy water molecules away into the bulk (Figure 1) can therefore increase the ligand binding specificity [31–33] and affinity [33–35] to the target. Therefore, the importance of considering water molecules in the drug design process has been long recognized [10,36]. Besides ligand optimization, water molecules have been also utilized to improve docking results (See Section 5 for details [35,37,38]).

Despite the above importance of water in drug design, the determination of the structure and energy contribution of the water networks in intra- and intermolecular interactions is challenging for both experimental and theoretical approaches [13,39,40]. Water molecules are often too mobile, as they can change their positions in the meantime of picoseconds [6] and get lost in the large electron density maps of proteins [39]. On the other hand, theoretical approaches often have a considerable computational cost, calculating all the possible interactions with water molecules in a large simulation box. The present review gives a brief account on the above limitations and advances of the recent experimental and theoretical methodologies for water structure and their applications in the context of molecular engineering focused on biomolecules and drug design.

2. Experimental Determination of Water Structure

The experimental methods of X-ray/neutron crystallography [41], cryo-electron microscopy (Cryo-EM) [42–44], and Nuclear Magnetic Resonance (NMR) spectroscopy [45,46] can be considered as the primary techniques for the determination of molecular structures at the atomic level. While these methods provide a solid background for establishing the structure–activity relationships [47] of biomolecules and their complexes, they face several challenges in the determination of hydration structures. Most of these difficulties come from the complexity (hydration layers interconnected with hydrogen-bonding networks) and high mobility (dynamic exchange of water molecules between the layers) of hydration structures. Plausibly, water molecules and other non-amino acid moieties, such as ligands, ions, or metals, are not included in the amino acid sequence that is otherwise key information for protein structure determination, indicating the order and covalent links between amino acids.

More than half of the experimental structures published in the Protein Data Bank (PDB, [48,49]) contain at least one water molecule (Figure 2a). Crystallography is the most powerful technique for the exploration of the networks of several water molecules. Cryo-EM, NMR, and other methods can assign far fewer (often individual) water positions (Figure 2a).

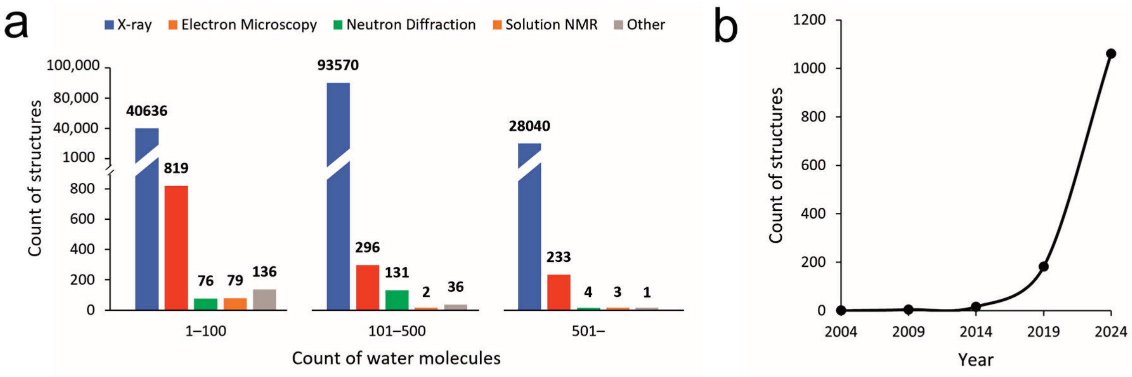


Figure 2. (**a**) Counts of structures containing water molecules resolved by different methods and deposited in the Protein Data Bank. Data were collected by the advanced search module of the PDB database: Entry features > Number of water molecules per deposited model. The number of structures was automatically separated by methods used for resolving. (**b**) The number of water-containing cryo-EM structures resolved in 5-year-long periods of time. The number on the x axis indicates the last year of the period.

However, crystallography provides only a static picture of the structure of solute molecules and their first surrounding water shell [50]. Moreover, the assignation of water positions in electron density maps gained via Fourier Transform from the crystal diffraction pattern is often complicated, even in the first shell. One of the methods is based on the low B factors, by which the waters bound to the protein surface or another water molecule located in the first hydration shell can be confidently identified [51]. In buried regions, such as binding pockets or active sites, even the third hydration layer can be resolved [50,52]. Problematic, partially ordered waters located mainly in the second hydration shell [53] can be assigned using D_2O-H_2O “neutron difference maps” [54]. This method uses the large difference in neutron scattering by deuterated and light waters, resulting in peaks of only water locations, while the scattering of the solute remains the same [51,53,54]. While neutron diffraction is capable of detecting not only oxygen but also hydrogen/deuterium atoms, and has been continuously developed [55], it is still less widespread due to the technical complexity of the method [56] and the limited accessibility of the neutron sources based on only four nuclear reactors worldwide [57], also reflected by the small number of structures [49] resolved by this method (Figure 2a).

There are various computational methods that help with the assignation of water positions in electron density maps, sometimes equipped with quantum- and/or molecular mechanics refinements [58]. For example, PHENIX [59–61] is a frequently used system for macromolecular crystallographic structure solutions, in which a bulk-solvent determination protocol is based on both maximum-likelihood and least-squares target functions [62]. Coot performs a cluster analysis on a residual map to find water places [63]. The assigned waters are then checked based on their distance from the hydrogen-bond donors and acceptors, temperature factor, or electron-density level [63]. ARP/wARP [64] includes a fully automated placement for finding ordered water molecules using least-square refinement, in combination with the Fo–Fc difference of the electron density maps [65] (where Fo and F_C are the observed and calculated structure factor amplitudes), and geometric assumptions such as interatomic distances, angles, and van der Waals radii, etc. [66]. Despite the development of new assignation tools, the determination of correct water positions remains problematic, especially if the water structure is disordered surrounding non-polar atoms [67] or has fewer tetrahedral hydrogen bonds [68] at partially occupied solvent sites of low density.

The number of water molecules determined by crystallography mainly depends on the size and form of the system [52], as well as its resolution. By increasing the size of the system, the number of water molecules [41] also increases and the solvent becomes considerably disordered. At the size of the proteins (MW > 30.000), the resolution is usually between 1.5 and 2.5 Å, having a high background noise level in Fourier maps due to the high incoherent scattering cross-section of the numerous hydrogen atoms [41], which makes the solute assignation more difficult [41,69]. Furthermore, in the case of large biomolecules, the number of hydration layers increases, resulting in a weaker and more diffuse solvent density [41]. The other problem is that the electron density of water molecules is similar to that of small iso-electronic ions (e.g., sodium and ammonium), leading to inaccurate assignation. Moreover, the experimental conditions also affect the successful, accurate, and valid water assignation. The limitations of crystallography include the difficulty of the crystallization of biomolecules, especially in the case of large, non-globular, or disordered systems [50]. Furthermore, it is also questionable how the crystallization procedure, such as packing and cryogenic temperature, modifies the native structure of the biomolecule [50] and its hydration shells. It has been proven that the hydration structure of a biomolecule highly depends on temperature [70].

In the case of cryo-EM, high-resolution structural information is gained from thousands of images produced by transmitting an electron beam through the protein sample embedded into a special vitreous environment instead of crystals. Thus, the biomolecules can be studied in a more “native” environment, with different conformational and/or functional states, and this allows for the resolution of structures in a higher molecular weight range than that of X-ray crystallography [71]. Due to the progressive improvement in technological and refining processes, the resolution of cryo-EM maps has been entered into the atomic dimension, where the resolvability of individual atoms, including solvent water atoms, is accessible [72]. The first cryo-EM structure with water molecules was published in 2003 (PDB code: 1uon) [73] at a resolution of 7.6 Å, which is too low for the identification of individual atoms. Due to the ‘Resolution Revolution’ [74,75], which started in 2013, less than a decade ago, when the first near-atomic resolution cryo-EM structures were published [76–78], the number of water-containing cryo-EM structures has exponentially increased (Figure 2b). This tendency might forecast that cryo-EM structures will catch up to the number of X-ray structures in the next decades, especially in the case of large protein complexes, cellular machines, and viruses [79]. The above computational methods used for the assignation of waters in X-ray crystallography could also be applied to cryo-EM maps. The development of new assignation tools has emerged in this field as well. The assignation of individual atomic positions in cryo-EM can be performed using methods such as SWIM (segmentation-guided water and ion modelling) [80] and UnDowser in MolProbity [81,82].

Unlike crystallography and cryo-EM, NMR spectroscopy is suitable for examining small proteins or oligopeptides in solutions adopting various conformations [50]. Water–protein interactions can be identified by using the nuclear Overhauser effect and/or rotating-frame Overhauser effect between the water protons and protein atom nuclei [83]. Here, individual water molecules can be determined that are located in the first hydration shell and bound to the protein instead of a complex 3D hydration structure [84]. It is notable that this method is limited by the short-term period of protein–water interactions, the hydrogen exchange with unstable protein moieties, and long-range dipole coupling, and identifies only 1–100 water molecules at best (Figure 2a). Additionally, while crystallography and cryo-EM provide direct information on the positions of water oxygen atoms, solution NMR is based on different principles.

3. Calculation of Water Structure

While there is an impressive, continuous development of experimental structure determination methods, the previous Section also highlighted the limitations of their assignment of the positions of water molecules [9,13,40]. To fill the gap of missing experimental hydration structures, extensive theoretical research has been conducted and resulted in various methods for the calculation of water positions (Table 1).

Table 1. The categorization and performance of theoretical methods of prediction of hydration structure.

Method	Concept	Type [a]	#System/#Water [b]	Match Tolerance (Å)	SR (%)
3D-RISM [c] [85–87]	Knowledge	IF	18/113 [d]	2.5	91
		IF	13/113 [e]	1.5	65
		SF	8/101 [e]	1.5	60
AcquaAlta [c] [88]	Geometry	IF	20/77	1.4	76
Auto-SOL [c] [89]	Geometry	SF	5/1337	1.5	64
AQUARIUS [f] [90]	Knowledge	SF	7/1376	1.4	59
Fold-X [c] [91]	Energy	SF	74/2687	1.0	76
Forli et al., 2012 [c] [92]	Geometry [g]	IF	27/51	2.0	96
HADDOCK [c] [93]	Geometry [g]	IF	27/50	2.0	90
Huggins and Tidor, 2011 [94]	Geometry	IF	5/19	2.0	68
HydraMap [c] [95]	Dynamic	IF	13/113 [e]	1.5	72
		SF	8/101 [e]	1.5	69
HyPred [f] [96]	Dynamic	SF	3/233	1.0	12
MobyWat [c] [13,40]	Dynamic	SF	20/1500	1.5	80
		IF	31/344	1.5	90
Particle concept [h] [97]	Geometry	IF	200/232	1.5	35
Splash'Em [c] [98]	Knowledge	IF	91/230	1.0	62
SZMAP [h] [99]	Knowledge	IF	18/113 [d]	2.5	96
WaterDock [c] [100]	Energy	SF	7/92	2.0	88
WaterFLAP [h] [87,101,102]	Knowledge	IF	18/113 [d]	2.5	98
WaterMap [h] [37,87]	Dynamic	SF	1/11	1.5	82
		IF	18/113 [d]	2.5	96
WarPP [c] [103]	Geometry	IF	1500/20,000	1.0	80
WATGEN [f] [104]	Geometry	IF	126/1264	2.0	88
WATsite [c] [95,105]	Dynamic	IF	13/113 [e]	1.5	75
		SF	8/101 [e]	1.5	77

[a] Water molecules in the target–ligand interface (IF), and on unbound target surface (SF) are considered, respectively. [b] The count of systems/the count of experimental water oxygen positions used in the cited study for method validation. [c] Freeware or free trial for academic use. [d] These data are taken from the comparative paper [87]. [e] These data are taken from the paper [95]. [f] Website no longer available. [g] With docking search. [h] Commercially available.

Despite the flaws of these experimental methods, the validation of theoretical methods still relies on the experimental water oxygen positions. The positions of the predicted water oxygen and experimental water oxygen are compared, and if the distance is within a tolerance threshold, then it is accepted as a successfully predicted water oxygen position. The ratio of the count of the successfully predicted water oxygen positions and all the available experimental water oxygen positions can be considered as a success rate (SR, this number is expressed in percentage after multiplication with 100 in Table 1). The validation and comparison of different theoretical methods can be easily performed based on SR values (Table 1).

Out of the four types (bulk, buried, interface, and surface) of water molecules mentioned in the Introduction (Figure 1), mostly surface and interface water molecules are investigated by theoretical methods (Table 1). Hydrated targets have surface water molecules in their first hydration shell [40] that have a stabilizing function on the macromolecular structure. Target–ligand complexes also have interface water molecules bridging between the target and ligand [13,15,19,25,106]. The prediction of interface water molecules can be accomplished very precisely with SR values even above 90% (Table 1) [9], as these molecules are captured between the target and the ligand, and there is enough space to fit only the water molecules participating in the interaction. Surface water molecules tend to have a higher mobility (B-factors) and can be predicted with SRs of ca. 80% (Table 1) [24,29,40]. That is, the natural uncertainty of surface water positions tends to result in a lower success of their prediction [39,40].

Either static or dynamic methods are used for the prediction of interface or surface water molecules. Static methods assume a static hydration shell and predict the binding sites of the water molecules on the surface of a dry experimental solute structure [40]. Finding a binding site can rely on energy calculations, scoring, prior knowledge, and information on H-bonds, and neural networks have also been applied [107]. Knowledge-based methods rely on the information found mainly in X-ray crystallographic structures (see previous Section). The main limitation of knowledge-based approaches is the assembly of an appropriate test set. The quality of X-ray crystallographic water oxygen positions varies greatly (see previous Section) and the methods perform better on similar structures that are involved in their test sets. Some methods assign a score to the experimental water molecules. Energy calculations may also apply popular docking tools to predict water molecule positions. Energy- or grid-based methods try to locate the energetically favorable positions of water molecules using probes that mimic them. Static methods can accurately identify the water molecules at the interfaces of the proteins and ligands, as these waters are usually static; however, a dynamic exchange of water molecules between the bulk solvent and the protein surface is disregarded by these methods. Generally, static methods do not consider an explicit water model and provide fast results. However, the quickness of these methods often involves a compromise in their precision.

Dynamic methods rely on extensive molecular dynamics (MD) simulations or other global search techniques using an explicit water model and allowing for the mobility of individual water molecules. All atomic movements are recorded into trajectories and the protein–water, ligand–water, and water–water interactions can be followed. This includes a dynamic exchange of water molecules with the bulk solvent and the displacement of water molecules from the binding site by ligands. However, the analysis of each trajectory in a large-scale study using various systems would be time-consuming. To tackle this, the distribution density averages of the water molecules or their occupancy at binding sites might be investigated. Dynamic approaches offer information not only on the location of water molecules, but the displacement of water molecules can be also studied. MD-based thermodynamic analyses or a comparison of the hydration structures of the apo and holo targets can follow the application of these dynamic approaches.

The counts of the systems and water molecules involved in the validation differ in different methods (Table 1). In future studies, the involvement of at least 1000 and 100 experimental (reference) water positions can be recommended for surface and interface predictions, respectively. Preferably, at least 10 different (protein or complex) systems should be used to have a diverse set of water positions. Notably, the SR depends on the choice of match tolerance, where the highest value is 2.5 Å, but more commonly 1.4–2.0 Å is used, which seems to be the consensus for method validation (Table 1). Naturally, when the match tolerance is set to a higher value (2.0–2.5 Å), the methods achieve better SRs. Notably, the calculated water positions and SR values correspond to a certain biomacromolecular structure (or PDB ID), and the use of high-resolution structures can be recommended for the calculation of the SR. While SRs provide a fair comparison of methods, the number of experimental water oxygen positions used for the method validation and testing is similarly important. Notably, MobyWat, WATGEN, and WarPP use more than 300 experimental water positions for the validation of interface hydration. Auto-SOL, AQUARIUS, Fold-X, and MobyWat use more than 1300 waters to test their surface predictions.

The theoretical approaches of the above sections complement well the experimental methods for the atomic-level determination of water structures. In some cases, these methods also offer a complete hydration structure [13,40] of the protein surfaces and interfaces, which is often not achieved using experimental methods due to assignation problems (Section 2). Knowledge of the complete water structure is especially important for the calculation of (single molecular) the energy contribution of the (de)solvation process of drug–target binding (next Sections).

4. Water in the Structure-Based Calculation of Binding Thermodynamics

Water influences the thermodynamics of various biochemical interactions [108,109] important in molecular engineering. For example, ligand binding is described by binding free energy (ΔG_b), a net measure of the strength of target–ligand interactions. During the formation of target–ligand complexes, hydration shells undergo considerable changes (Figure 1) and, therefore, the mediation of the interactions between the drug and target partners is fairly dependent on the water molecules. There are implicit and explicit water models for the calculation of the energetics of the (de)solvation during ligand binding. Implicit solvation models consider the solvent as a continuous medium around solutes and manifest in the formulae, e.g., in electrostatic terms [110]. Explicit water models place numerous water molecules in the simulation box and set various molecular properties for the water prototype used in copies [111,112]. Both types of models have been implemented at the molecular mechanics (MM) and quantum mechanics (QM) levels of calculations.

At the molecular mechanics level, implicit water models such as MM-PB(GB)/SA [113] are widely used and based on the solution of the theoretically accurate, but computationally expensive Poisson–Boltzmann (PB) equation, or a simplified but scalable Generalized Born (GB) equation, to obtain the polar contribution of the solvation free energy change on an ensemble of MD snapshots. The solute cavity formation within the solvent and the van der Waals interactions between the solute and the solvent are represented by a nonpolar term often based on solvent-accessible surface areas (SA) [114]. Docking programs have also implemented implicit water models in their scoring functions due to their simple formulation and low computational costs. Notably, the scoring functions of docking methods require the fastest possible approaches to maintain their high-throughput nature. For example, the popular docking program AutoDock [115] applies the method of Stouten et al. [116], which calculates the solvation free energy as a sum of the atomic contributions with a linear relationship between the percentage of free volume around the atom and its contribution. At the same time, a PB-based distance-dependent dielectric function was also implemented in the Coulomb potential of AutoDock, which dampens the water permittivity value and corrects the screening effects near the solute surfaces [117]. In this way, a continuous transition of the relative permittivity of the medium is considered as we go from the bulk water to the protein surface. Similar implicit solvation terms have also been implemented in other popular docking software such as DOCK [118], MOE [119], and FITTED [120].

While implicit water models are useful for the approximation of long-range electrostatic forces considering the above-mentioned screening effect of solvent dielectric [110], they cannot handle hydration shells (Figure 1) and specific water-mediated linkages. However, the absence or presence of a certain water molecule at the binding site can drastically modify the overall affinity of ligands [121,122]. Thus, an accurate calculation of the binding thermodynamics is a rather impossible undertaking without the representation of individual water molecules.

Explicit water models have been introduced to overcome the above-mentioned limitations of implicit approaches. At the MM level, there are various explicit models, such as SPC [123], TIP3P [124], and TIP4P [124], where the abbreviated names refer to the charge systems and sites (parameters) of the water molecule prototype. Explicit approaches allow for the calculation of the energy contributions of individual waters, e.g., using the statistical mechanical inhomogeneous fluid solvation (IFST, [125]) or grid inhomogeneous solvation (GIST, [126]) theory. In this way, the enthalpic and entropic terms of bound waters can be also considered, like in Wscore [127], DOCK-GIST [35], and AutoDock-GIST [38]. In some cases, the incorporation of explicit waters with the above methods has improved the correlation between the experimentally determined binding affinity and the docking score [38,127], while other works have not observed such improvements [87,128].

In the realm of quantum mechanics, theory permits a more accurate calculation of the charge distribution of molecules compared to MM. The assessment of the electrostatic interaction between the solute and water, in theory, can be included in the self-consistent field (SCF) calculation using dielectric continuum models [110]. However, for realistic solute cavities, it requires a numerical iterative process for every SCF cycle, which is extremely computationally demanding [129]. The Conductor-like Screening Model (COSMO) [130] solves this problem using a Green function description with analytical gradients, making the method practically applicable. COSMO can be considered as an advanced version of the polarizable continuum model (PCM, [131,132]), and is the most accurate implicit solvation model for semi-empirical QM. There is also a universal solvation model based on solute electron density (SMD, [133]), which is usually implemented for more computationally demanding levels of QM. At the semi-empirical level, the combination of PM6s [134–137] and PM7 [137–139] parametrizations, combined with the implicit model of COSMO, is a popular choice for estimating the binding affinities of ligands to targets.

Advances in computational speed and linear scaling methods [140] have allowed for the combination of implicit (COSMO) and explicit models handling long-range electrostatics and individual water contributions, respectively. Such hybrid approaches present a fast QM alternative of MM scoring functions for drug design. For example, Nikitina et al. inserted possible interface waters into hydrogen bond donor–acceptor sites and used the PM3 method [141,142]. Horváth et al. predicted interface waters using a molecular-dynamics-based method, MobyWat [40], and utilized the hybrid water model with PM7 parametrization for the estimation of the binding enthalpies (ΔH_b) of ligands [143]. Here, the inclusion of explicit waters in the hybrid model yielded, e.g., a 3-fold smaller relative error when compared with vacuum calculations (Figure 3). Cavasotto et al. used a single-point PM7 calculation, keeping crystallographic waters in the binding interface in their QM docking scoring methodology to show encouraging enrichment factors on 10 protein targets [144]. The latter studies also extract the binding site environment from the target (similarly to Figure 3) to further reduce the computational time. The ΔG_b was also calculated by Hyslova et al., using a DFT-D3 and PM6-D3X4 combined method with crystallographic waters in the binding pocket, achieving a better fit with the combined implicit/explicit procedure (R^2 = 0.68) compared to the implicit alone (R^2 = 0.49) [145].

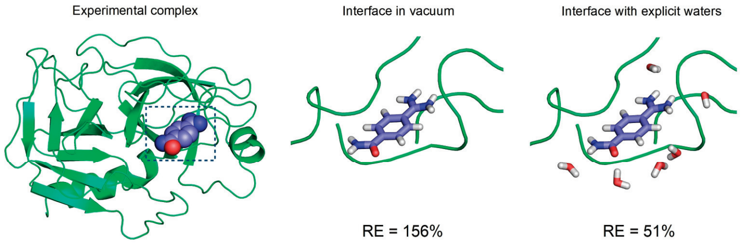


Figure 3. The complex structure of beta-trypsin (on the left, target in cartoon representation) and p-amidinobenzamidine (ligand marked with spheres). The target–ligand interface extracted for ΔH_b calculations is marked with a box. The close-up of the dry (middle) and explicitly hydrated (right, used for hybrid calculations) interface with a ligand in sticks representation. Relative errors (RE) of the calculated binding enthalpy (ΔH_b) values of the dry and hybrid models are indicated below the corresponding interface structure. The RE values were calculated as RE(%) = 100 ($\Delta H_{b,calculated} - \Delta H_{b,experimental}$)/$\Delta H_{b,experimental}$. The coordinates and ΔH_b values were re-used from a previous study [143] (Table 1 and Supporting Supplementary Table S5, β = 0). The incorporation of explicit water molecules in the ΔH_b calculation considerably reduced the RE in this case.

5. Water in Target-Ligand Docking

Target–ligand complex structures (Figure 1) are key to the engineering of new drugs. Computational docking can supply such atomic-resolution complex structures rapidly and, therefore, it is a widely used [146–148] alternative of experimental structure determination techniques (Section 2) in ligand screening projects [149–151]. Water molecules are active participants in real docking situations, as described in the Introduction (Figure 1). However, the proper use of these water molecules during computational docking is not trivial [36]. The inclusion of happy waters (Figure 1) bridging in the target–ligand interface may help to increase the precision of the docking calculation. On the other hand, if unhappy waters (Figure 1) are included in the interface, they would erroneously block the docking to the target sites used by the ligand in reality. Thus, the misuse of unhappy waters in an interface obviously leads to the mis-docking of the ligand. However, without knowledge of the true hydrated complex structure, it is rather difficult to distinguish between happy and unhappy water molecules in advance. Docking with waters is therefore a true “chicken and egg situation”, where let us say water is the chicken and the docked ligand is the egg. Docking is expected to produce a proper target–ligand complex for the decision on the inclusion of happy water molecules in (or the exclusion of unhappy ones from) the docking itself. This awkward situation is reflected in the corresponding literature. Several studies have reported that the incorporation of specific water molecules in the docking process significantly improved the docking performance [127,152–154], while others have found that including water molecules had little effect on this performance [155,156]. Several fast docking tools and strategies [38,92,97,127,157–166] have been developed to incorporate waters in the binding site during docking simulations. Many of these tools work with experimentally determined (known) water positions [160,163–165].

A simple way of incorporating water molecules into docking simulations is to include them as a static part of the target [167], where the positions and orientations of these water molecules are kept restrained during the simulation [162,167]. This strategy is used most in molecular docking studies and has been shown to be effective [168–170]. An improvement to the restrained water model is the displaceable water model, where the included water molecules can be switched on/off automatically during the simulation so that a ligand can keep the favorable water molecules and displace the non-favorable water molecules (GOLD, FlexX, FITTED, and DOCK). These included waters have mostly fixed positions or a limited mobility during the docking process, while some methods allow for the waters to change their positions and orientations through the search algorithm (in FITTED).

Other methods solvate the ligand and then dock the solvated ligand with full flexibility, as the waters are kept or displaced depending on the entropy and/or energy contributions during the simulation (AutoDock4, MVD, and Glide). Such ligand-centric methods treat water molecules as a flexible part of the ligand, so they present the same flexibility as the ligand itself. RosettaLigand includes water movement both independently and dependently from the ligand during the initial stage, while considering the full flexibility of the target and the ligand through MC search. However, when the method was evaluated on a dataset of 341 diverse protein/ligand complexes from CSAR, no significant improvement was observed in the docking success rate [165]. This could be caused by there being no solvent-specific scoring adjustments in RosettaLigand other than the desolvation energy calculated using an implicit solvent model. Such a desolvation term in a force-field-based scoring function is often calibrated for protein–ligand complexes with no explicit solvent [165].

In many cases, the experimental water positions are not available or the hydration structure is not complete. In such situations, theoretical methods (Section 3) can supply the water positions for the docking calculations. Solvation before docking and a short molecular dynamics (MD) simulation were performed to improve these water positions, and after the encounter of the interacting partners, the water was removed based on a Monte-Carlo approach in HADDOCK [93,171] and a semi-explicit water model implemented in Rosetta [172]. This method improved the docking results for HADDOCK when compared to docking without explicit water molecules. FITTED [120] treats water molecules as a part of the target, and conserved water molecules are considered by an entropic penalty in the final score. This improves the docking accuracy for HIV-1 inhibitors. GOLD relies on two programs, FlexX [97,173] to pre-calculate the energetically favorable water sites and insert spherical water molecules (particles), and Consolv [174] to predict the water molecules that are likely to be displaced [166]. There is an upper limit of water molecules that is handled by this approach, so as not to increase the complexity and therefore the computational costs above a rational limit [36]. Although GOLD predicted the conservation or displacement of water molecules with a high efficiency, the effect on the ligand binding pose prediction was moderate [166]. Slide [160] enables the virtual screening of a relatively large set of ligands using Consolv [174] for the water prediction.

HydroDock [175], a new approach, separates the chicken and egg situation and solves the hydrated docking using a parallel approach. The ligand is docked into the target without waters (dry docking). Simultaneously, the whole target is filled up with the explicit water molecules predicted by MobyWat [13,40]. Then, the dry docked complex and the hydrated target are merged and the water molecules that clash with the dry docked ligand are removed. The resultant complex is then energy minimized to set the proper orientation of the hydrogen bonds, and a short MD simulation is performed to yield the representative binding mode. HydroDock achieved a high accuracy in the case of ion-channel-bound ligand docking (Figure 4), one of the hardest cases of including water molecules in molecular docking [9,175]. As a specific example, the HydroDock method was validated on the ion channels of the influenza A and SARS-CoV-2 viruses.

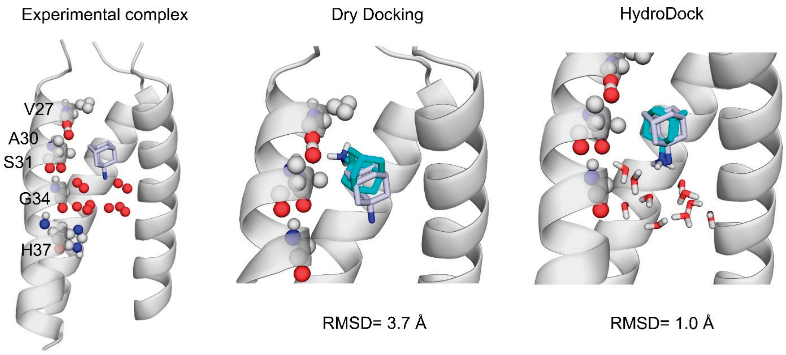


Figure 4. The role of water in ligand binding and the incorporation of explicit water molecules into docking using HydroDock. The influenza virus A ion channel is shown as grey cartoon, target amino acids are shown as spheres and labelled according to the 6bkk [176] PDB structure. Experimental water

oxygen positions are shown as red spheres, and water molecules after HydroDock are shown as red and white sticks [175]. The experimental amantadine structure is shown as grey sticks, and the calculated structures as teal sticks. Dry docking (in the middle) fails to reproduce the experimental ligand binding mode, however, docking with water molecules (on the right) improved similarity with experimental results greatly. Root mean squared deviation (RMSD) is calculated after superimposition of the calculated to the experimental structure, between ligand heavy atoms.

As the inclusion of explicit water molecules increases computational costs [177], the scoring functions of many fast docking methods do not treat explicit water molecules with the proper partial charges and terms for their enthalpic or entropic contributions [177,178]. However, the way that water molecules are treated in the binding site and how their energetic contributions are evaluated is considered to be a key factor greatly affecting the docking performance [127]. To improve this docking performance, the contribution of water-mediated interactions and entropic effects may be considered for individual water molecules. A common modification to scoring functions is to add an entropy penalty, using a positive constant for each included water molecule to model the loss of rigid-body entropy favoring the displacement of the water molecules. However, in the case of large ligands, this approach leads to extremely positive energy contributions, necessitating a modification of the scoring function of AutoDock [179]. Moreover, Friesner and co-workers showed that the contributions of some water molecules to the free energy of binding can be much larger than others [180]. Some attempts have been made to calculate the target surface water sites and thermodynamics prior to the docking process, using third-party tools such as GIST and WaterMap, and to incorporate the solvation information in the scoring function (WScore, AutoDock-GIST, and DOCK-GIST). Although evaluation studies have reported only minor improvements in the success rate of docking for WScore and AutoDock-GIST, such an incorporation of explicit waters into the energy calculation definitely improved the correlation between the experimentally determined binding affinity and the docking score [38,127]. On the other hand, knowledge-based methods such as Consolv [174] (a k-nearest-neighbor-based classifier trained on 5542 molecules taken from 30 independently solved protein structures) can be used to determine the probability of the water molecules in the binding site to be conserved or displaced, as well as their corresponding desolvation penalty values (implemented in Slide) [160]. Instead of on-the-fly energy evaluations, scoring functions with more accurate desolvation functions can be implemented as re-scoring tools after the docking. For example, Wang et al. used molecular-mechanics–Poisson–Boltzmann-surface-area (MM–PBSA) re-scoring to find HIV-1 reverse transcriptase inhibitors [181], and several studies have reported that rescoring using a molecular-mechanics–generalized-Born-surface-area (MM–GBSA) method improved the enrichment of the known ligands for several enzymes and even the identification of substrates [182–184].

In the last decade, targeting protein–protein/DNA/RNA interactions has been considered to be a promising strategy for drug discovery [185–188], and a growing number of docking methods have been specifically developed for this [189]. Their shallow and relatively large interface (more than 1500 Å2 compared to the 300−1000 Å2 range for binding sites) [190] makes it readily accessible to the solvent or water-permeable in the case of nucleic acids. For nucleic acids, unlike proteins, their phosphate groups and corresponding counter ions (such as Mg^{2+} or K^{+}) cause polarization upon the water molecules and functional groups of drugs. Thus, water molecules often play an important role in the ligand recognition and complex stabilization for nucleic acids, as well as proteins. There are several publications that have reported improvement in the success rate of the docking results when water molecules were included for RNA [191], DNA, and protein–protein complexes [192,193]. However, due to their large, solvent-accessible interface, it is extremely challenging to incorporate water molecules into the process of docking macromolecules within a reasonable computation time. Thus, the effect of these water molecules is often ignored in protein–protein/DNA/RNA docking, in which the desolvation penalty is estimated as proportional to the solvent-accessible surface area. There have been attempts to incorporate solvation effects in macromolecule docking. For example, HADDOCK explicitly treats water molecules by performing rigid-body docking on solvated macromolecules, followed by a Monte-Carlo (MC) simulation that displaces the waters based on their probabilities to form water-mediated contact, predicted using the Kyte-Doolittle scale [194]. Pavlovicz et al. developed a semi-explicit water model (implemented in Rosetta), in which a modified MC simulation displaces or adds explicit solvent molecules from bulk, followed by an energy evaluation with an implicit solvation energy term. Both attempts have improved the docking and ranking results [171,172].

6. Conclusions

Molecular engineering and drug design have been continuously fueled by the development of experimental structure determination techniques. However, the determination of the position of individual water molecules is often limited by the low resolution of their measurements. Theoretical calculations can supply the atomic-resolution hydration structure of target–ligand interfaces with a high precision, and often complement experimental techniques. The energetic contribution of individual water molecules to the full thermodynamics of target–ligand binding can be also calculated. There has been an improvement in the application of water structures in computational docking, a technique often used in the high throughput virtual screening of ligands in the drug industry. While “docking with waters” is still a problematic “chicken and egg situation”, a number of methods have been featured that answer this challenge as well.

Author Contributions: Conceptualization, C.H.; writing—original draft preparation, B.Z.Z., B.B., R.B., V.S., V.M., C.H.; writing—review and editing, B.Z.Z., B.B., R.B., V.S., V.M., C.H.; visualization, B.Z.Z., B.B., R.B., V.S., V.M., C.H.; supervision, C.H.; project administration, B.Z.Z., C.H.; funding acquisition, B.Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Supported by the PTE ÁOK KA-2022-26 grant. The project has been supported by the European Union, co-financed by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. We acknowledge that the results of this research have been achieved using the DECI resource Archer2 based in the UK at the National Supercomputing Service with support from the PRACE aisbl.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chumlea, W.C.; Guo, S.S.; Zeller, C.M.; Reo, N.V.; Siervogel, R.M. Total Body Water Data for White Adults 18 to 64 Years of Age: The Fels Longitudinal Study. *Kidney Int.* **1999**, *56*, 244–252. [CrossRef] [PubMed]
2. Elsässer, S.J.; Huang, H.; Lewis, P.W.; Chin, J.W.; Allis, C.D.; Patel, D.J. DAXX Envelops a Histone H3.3–H4 Dimer for H3.3-Specific Recognition. *Nature* **2012**, *491*, 560–565. [CrossRef]
3. Manolaridis, I.; Kulkarni, K.; Dodd, R.B.; Ogasawara, S.; Zhang, Z.; Bineva, G.; O’Reilly, N.; Hanrahan, S.J.; Thompson, A.J.; Cronin, N.; et al. Mechanism of Farnesylated CAAX Protein Processing by the Intramembrane Protease Rce1. *Nature* **2013**, *504*, 301–305. [CrossRef] [PubMed]
4. Musset, B.; Smith, S.M.E.; Rajan, S.; Morgan, D.; Cherny, V.V.; DeCoursey, T.E. Aspartate 112 Is the Selectivity Filter of the Human Voltage-Gated Proton Channel. *Nature* **2011**, *480*, 273–277. [CrossRef] [PubMed]
5. Ostmeyer, J.; Chakrapani, S.; Pan, A.C.; Perozo, E.; Roux, B. Recovery from Slow Inactivation in K+ Channels Is Controlled by Water Molecules. *Nature* **2013**, *501*, 121–124. [CrossRef] [PubMed]
6. Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **2008**, *108*, 74–108. [CrossRef]

7. Ball, P. Water Is an Activematrix of Life for Cell and Molecular Biology. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 13327–13335. [CrossRef]
8. Bellissent-Funel, M.C.; Hassanali, A.; Havenith, M.; Henchman, R.; Pohl, P.; Sterpone, F.; van der Spoel, D.; Xu, Y.; Garcia, A.E. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.* **2016**, *116*, 7673–7697. [CrossRef]
9. Zsidó, B.Z.; Hetényi, C. The Role of Water in Ligand Binding. *Curr. Opin. Struct. Biol.* **2021**, *67*, 1–8. [CrossRef]
10. Bodnarchuk, M.S. Water, Water, Everywhere... It's Time to Stop and Think. *Drug Discov. Today* **2016**, *21*, 1139–1146. [CrossRef]
11. de Simone, A.; Dodson, G.G.; Verma, C.S.; Zagari, A.; Fraternali, F. Prion and Water: Tight and Dynamical Hydration Sites Have a Key Role in Structural Stability. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7535–7540. [CrossRef]
12. Miyano, M.; Ago, H.; Saino, H.; Hori, T.; Ida, K. Internally Bridging Water Molecule in Transmembrane α-Helical Kink. *Curr. Opin. Struct. Biol.* **2010**, *20*, 456–463. [CrossRef]
13. Jeszenői, N.; Bálint, M.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Exploration of Interfacial Hydration Networks of Target-Ligand Complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148–158. [CrossRef]
14. Pradhan, M.R.; Nguyen, M.N.; Kannan, S.; Fox, S.J.; Kwoh, C.K.; Lane, D.P.; Verma, C.S. Characterization of Hydration Properties in Structural Ensembles of Biomolecules. *J. Chem. Inf. Model.* **2019**, *59*, 3316–3329. [CrossRef] [PubMed]
15. Ahmad, M.; Gu, W.; Geyer, T.; Helms, V. Adhesive Water Networks Facilitate Binding of Protein Interfaces. *Nat. Commun.* **2011**, *2*, 261. [CrossRef]
16. Laage, D.; Elsaesser, T.; Hynes, J.T. Water Dynamics in the Hydration Shells of Biomolecules. *Chem. Rev.* **2017**, *117*, 10694–10725. [CrossRef] [PubMed]
17. Bruce Macdonald, H.E.; Cave-Ayland, C.; Ross, G.A.; Essex, J.W. Ligand Binding Free Energies with Adaptive Water Networks: Two-Dimensional Grand Canonical Alchemical Perturbations. *J. Chem. Theory Comput.* **2018**, *14*, 6586–6597. [CrossRef]
18. Zhong, H.; Wang, Z.; Wang, X.; Liu, H.; Li, D.; Liu, H.; Yao, X.; Hou, T. Importance of a Crystalline Water Network in Docking-Based Virtual Screening: A Case Study of BRD4. *Phys. Chem. Chem. Phys.* **2019**, *21*, 25276–25289. [CrossRef] [PubMed]
19. Venkatakrishnan, A.J.; Ma, A.K.; Fonseca, R.; Latorraca, N.R.; Kelly, B.; Betz, R.M.; Asawa, C.; Kobilka, B.K.; Dror, R.O. Diverse GPCRs Exhibit Conserved Water Networks for Stabilization and Activation. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 3288–3293. [CrossRef]
20. Breiten, B.; Lockett, M.R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C.M.; Heroux, A.; Krilov, G.; Whitesides, G.M. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein-Ligand Binding. *J. Am. Chem. Soc.* **2013**, *135*, 15579–15584. [CrossRef]
21. Brysbaert, G.; Blossey, R.; Lensink, M.F. The Inclusion of Water Molecules in Residue Interaction Networks Identifies Additional Central Residues. *Front. Mol. Biosci.* **2018**, *5*, 88. [CrossRef]
22. Jeszenői, N.; Schilli, G.; Bálint, M.; Horváth, I.; Hetényi, C. Analysis of the Influence of Simulation Parameters on Biomolecule-Linked Water Networks. *J. Mol. Graph. Model.* **2018**, *82*, 117–128. [CrossRef]
23. Kunstmann, S.; Gohlke, U.; Broeker, N.K.; Roske, Y.; Heinemann, U.; Santer, M.; Barbirz, S. Solvent Networks Tune Thermodynamics of Oligosaccharide Complex Formation in an Extended Protein Binding Site. *J. Am. Chem. Soc.* **2018**, *140*, 10447–10455. [CrossRef]
24. Rudling, A.; Orro, A.; Carlsson, J. Prediction of Ordered Water Molecules in Protein Binding Sites from Molecular Dynamics Simulations: The Impact of Ligand Binding on Hydration Networks. *J. Chem. Inf. Model.* **2018**, *58*, 350–361. [CrossRef] [PubMed]
25. Jukič, M.; Konc, J.; Gobec, S.; Janežič, D. Identification of Conserved Water Sites in Protein Structures for Drug Design. *J. Chem. Inf. Model.* **2017**, *57*, 3094–3103. [CrossRef] [PubMed]
26. Wahl, J.; Smieško, M. Thermodynamic Insight into the Effects of Water Displacement and Rearrangement upon Ligand Modifications Using Molecular Dynamics Simulations. *ChemMedChem* **2018**, *13*, 1325–1335. [CrossRef]
27. Hüfner-Wulsdorf, T.; Klebe, G. Protein–Ligand Complex Solvation Thermodynamics: Development, Parameterization, and Testing of GIST-Based Solvent Functionals. *J. Chem. Inf. Model.* **2020**, *60*, 1409–1423. [CrossRef]
28. Krimmer, S.G.; Betz, M.; Heine, A.; Klebe, G. Methyl, Ethyl, Propyl, Butyl: Futile but Not for Water, as the Correlation of Structure and Thermodynamic Signature Shows in a Congeneric Series of Thermolysin Inhibitors. *ChemMedChem* **2014**, *9*, 833–846. [CrossRef]
29. García-Sosa, A.T.; Mancera, R.L.; Dean, P.M. WaterScore: A Novel Method for Distinguishing between Bound and Displaceable Water Molecules in the Crystal Structure of the Binding Site of Protein-Ligand Complexes. *J. Mol. Model.* **2003**, *9*, 172–182. [CrossRef] [PubMed]
30. Chen, D.; Li, Y.; Zhao, M.; Tan, W.; Li, X.; Savidge, T.; Guo, W.; Fan, X. Effective Lead Optimization Targeting the Displacement of Bridging Receptor–Ligand Water Molecules. *Phys. Chem. Chem. Phys.* **2018**, *20*, 24399–24407. [CrossRef]
31. Harriman, G.; Greenwood, J.; Bhat, S.; Huang, X.; Wang, R.; Paul, D.; Tong, L.; Saha, A.K.; Westlin, W.F.; Kapeller, R.; et al. Acetyl-CoA Carboxylase Inhibition by ND-630 Reduces Hepatic Steatosis, Improves Insulin Sensitivity, and Modulates Dyslipidemia in Rats. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E1796–E1805. [CrossRef]
32. Collin, M.-P.; Lobell, M.; Hübsch, W.; Brohm, D.; Schirok, H.; Jautelat, R.; Lustig, K.; Bömer, U.; Vöhringer, V.; Héroult, M.; et al. Discovery of Rogaratinib (BAY 1163877): A Pan-FGFR Inhibitor. *ChemMedChem* **2018**, *13*, 437–445. [CrossRef] [PubMed]
33. Beuming, T.; Farid, R.; Sherman, W. High-Energy Water Sites Determine Peptide Binding Affinity and Specificity of PDZ Domains. *Protein Sci.* **2009**, *18*, 1609–1619. [CrossRef] [PubMed]

34. Jung, S.W.; Kim, M.; Ramsey, S.; Kurtzman, T.; Cho, A.E. Water Pharmacophore: Designing Ligands Using Molecular Dynamics Simulations with Water. *Sci. Rep.* **2018**, *8*, 10400. [CrossRef]
35. Balius, T.E.; Fischer, M.; Stein, R.M.; Adler, T.B.; Nguyen, C.N.; Cruz, A.; Gilson, M.K.; Kurtzman, T.; Shoichet, B.K. Testing Inhomogeneous Solvation Theory in Structure-Based Ligand Discovery. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E6839–E6846. [CrossRef] [PubMed]
36. de Beer, S.; Vermeulen, N.; Oostenbrink, C. The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem.* **2010**, *10*, 55–66. [CrossRef]
37. Abel, R.; Young, T.; Farid, R.; Berne, B.J.; Friesner, R.A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831. [CrossRef]
38. Uehara, S.; Tanaka, S. AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking. *Molecules* **2016**, *21*, 1604. [CrossRef]
39. Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **2015**, *55*, 771–783. [CrossRef]
40. Jeszenői, N.; Horváth, I.; Bálint, M.; Van Der Spoel, D.; Hetényi, C. Mobility-Based Prediction of Hydration Structures of Protein Surfaces. *Bioinformatics* **2015**, *31*, 1959–1965. [CrossRef]
41. Savage, H.; Wlodawer, A. Determination of Water Structure around Biomolecules Using X-Ray and Neutron Diffraction Methods. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 1986; pp. 162–183.
42. Halle, B. Protein Hydration Dynamics in Solution: A Critical Survey. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2004**, *359*, 1207–1224. [CrossRef]
43. Frank, J. Averaging of Low Exposure Electron Micrographs of Non-Periodic Objects. *Ultramicroscopy* **1975**, *1*, 159–162. [CrossRef]
44. Henderson, R.; Unwin, P.N.T. Three-Dimensional Model of Purple Membrane Obtained by Electron Microscopy. *Nature* **1975**, *257*, 28–32. [CrossRef]
45. Wüthrich, K. The Way to NMR Structures of Proteins. *Nat. Struct. Biol.* **2001**, *8*, 923–925. [CrossRef]
46. Wüthrich, K. Brownian Motion, Spin Diffusion and Protein Structure Determination in Solution. *J. Magn. Reson.* **2021**, *331*, 107031. [CrossRef] [PubMed]
47. Zsidó, B.Z.; Hetényi, C. Molecular Structure, Binding Affinity, and Biological Activity in the Epigenome. *Int. J. Mol. Sci.* **2020**, *21*, 4134. [CrossRef] [PubMed]
48. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef] [PubMed]
49. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
50. Biedermannová, L.; Schneider, B. Hydration of Proteins and Nucleic Acids: Advances in Experiment and Theory. A Review. *Biochim. et Biophys. Acta (BBA)—Gen. Subj.* **2016**, *1860*, 1821–1835. [CrossRef]
51. Mattos, C.; Ringe, D. Solvent Structure. In *International Tables for Crystallography*; International Union of Crystallography: Chester, UK, 2006; pp. 623–647.
52. Lounnas, V.; Pettitt, B.M. A Connected-Cluster of Hydration around Myoglobin: Correlation between Molecular Dynamics Simulations and Experiment. *Proteins: Struct. Funct. Genet.* **1994**, *18*, 133–147. [CrossRef]
53. Kossiakoff, A.A.; Sintchak, M.D.; Shpungin, J.; Presta, L.G. Analysis of Solvent Structure in Proteins Using Neutron D2O-H2O Solvent Maps: Pattern of Primary and Secondary Hydration of Trypsin. *Proteins: Struct. Funct. Genet.* **1992**, *12*, 223–236. [CrossRef] [PubMed]
54. Shpungin, J.; Kossiakoff, A.A. [24] A Method of Solvent Structure Analysis for Proteins Using D_2O-H_2O Neutron Difference Maps. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 1986; pp. 329–342.
55. Chatake, T.; Fujiwara, S. A Technique for Determining the Deuterium/Hydrogen Contrast Map in Neutron Macromolecular Crystallography. *Acta Crystallogr. D Struct. Biol.* **2016**, *72*, 71–82. [CrossRef] [PubMed]
56. Tanaka, I.; Chatake, T.; Fujiwara, S.; Hosoya, T.; Kusaka, K.; Niimura, N.; Yamada, T.; Yano, N. Current Status and near Future Plan of Neutron Protein Crystallography at J-PARC. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 2020; pp. 101–123.
57. Kono, F.; Kurihara, K.; Tamada, T. Current Status of Neutron Crystallography in Structural Biology. *Biophys. Physicobiol* **2022**, *19*, e190009. [CrossRef] [PubMed]
58. Schiffer, C.; Hermans, J. Promise of Advances in Simulation Methods for Protein Crystallography: Implicit Solvent Models, Time-Averaging Refinement, and Quantum Mechanical Modeling. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 2003; pp. 412–461.
59. Adams, P.D.; Afonine, P.V.; Bunkóczi, G.; Chen, V.B.; Davis, I.W.; Echols, N.; Headd, J.J.; Hung, L.-W.; Kapral, G.J.; Grosse-Kunstleve, R.W.; et al. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 213–221. [CrossRef]
60. Adams, P.D.; Grosse-Kunstleve, R.W.; Hung, L.-W.; Ioerger, T.R.; McCoy, A.J.; Moriarty, N.W.; Read, R.J.; Sacchettini, J.C.; Sauter, N.K.; Terwilliger, T.C. PHENIX: Building New Software for Automated Crystallographic Structure Determination. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 1948–1954. [CrossRef]

61. Echols, N.; Morshed, N.; Afonine, P.V.; McCoy, A.J.; Miller, M.D.; Read, R.J.; Richardson, J.S.; Terwilliger, T.C.; Adams, P.D. Automated Identification of Elemental Ions in Macromolecular Crystal Structures. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70*, 1104–1114. [CrossRef]
62. Afonine, P.V.; Grosse-Kunstleve, R.W.; Adams, P.D. A Robust Bulk-Solvent Correction and Anisotropic Scaling Procedure. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61*, 850–855. [CrossRef]
63. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. Features and Development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 486–501. [CrossRef]
64. Langer, G.; Cohen, S.X.; Lamzin, V.S.; Perrakis, A. Automated Macromolecular Model Building for X-ray Crystallography Using ARP/WARP Version 7. *Nat. Protoc.* **2008**, *3*, 1171–1179. [CrossRef]
65. Lamb, A.L.; Kappock, T.J.; Silvaggi, N.R. You Are Lost without a Map: Navigating the Sea of Protein Structures. *Biochim. et Biophys. Acta (BBA)—Proteins Proteom.* **2015**, *1854*, 258–268. [CrossRef]
66. Lamzin, V.S.; Wilson, K.S. Automated Refinement for Protein Crystallography. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 1997; pp. 269–305.
67. Levitt, M.; Park, B.H. Water: Now You See It, Now You Don't. *Structure* **1993**, *1*, 223–226. [CrossRef]
68. Deng, G.-H.; Shen, Y.; Chen, H.; Chen, Y.; Jiang, B.; Wu, G.; Yang, X.; Yuan, K.; Zheng, J. Ordered-to-Disordered Transformation of Enhanced Water Structure on Hydrophobic Surfaces in Concentrated Alcohol–Water Solutions. *J. Phys. Chem. Lett.* **2019**, *10*, 7922–7928. [CrossRef]
69. Carugo, O. Correlation between Occupancy and B Factor of Water Molecules in Protein Crystal Structures. *Protein Eng. Des. Sel.* **1999**, *12*, 1021–1024. [CrossRef]
70. Reuhl, M.; Vogel, M. Temperature-Dependent Dynamics at Protein–Solvent Interfaces. *J. Chem. Phys.* **2022**, *157*, 074705. [CrossRef] [PubMed]
71. Cheng, Y.; Glaeser, R.M.; Nogales, E. How Cryo-EM Became so Hot. *Cell* **2017**, *171*, 1229–1231. [CrossRef] [PubMed]
72. Pintilie, G.; Zhang, K.; Su, Z.; Li, S.; Schmid, M.F.; Chiu, W. Measurement of Atom Resolvability in Cryo-EM Maps with Q-Scores. *Nat. Methods* **2020**, *17*, 328–334. [CrossRef]
73. Zhang, X.; Walker, S.B.; Chipman, P.R.; Nibert, M.L.; Baker, T.S. Reovirus Polymerase Λ3 Localized by Cryo-Electron Microscopy of Virions at a Resolution of 7.6 Å. *Nat. Struct. Mol. Biol.* **2003**, *10*, 1011–1018. [CrossRef] [PubMed]
74. Kühlbrandt, W. The Resolution Revolution. *Science (1979)* **2014**, *343*, 1443–1444. [CrossRef]
75. Li, X.; Mooney, P.; Zheng, S.; Booth, C.R.; Braunfeld, M.B.; Gubbens, S.; Agard, D.A.; Cheng, Y. Electron Counting and Beam-Induced Motion Correction Enable Near-Atomic-Resolution Single-Particle Cryo-EM. *Nat. Methods* **2013**, *10*, 584–590. [CrossRef]
76. Allegretti, M.; Mills, D.J.; McMullan, G.; Kühlbrandt, W.; Vonck, J. Atomic Model of the F420-Reducing [NiFe] Hydrogenase by Electron Cryo-Microscopy Using a Direct Electron Detector. *eLife* **2014**, *3*, e01963. [CrossRef]
77. Amunts, A.; Brown, A.; Bai, X.; Llácer, J.L.; Hussain, T.; Emsley, P.; Long, F.; Murshudov, G.; Scheres, S.H.W.; Ramakrishnan, V. Structure of the Yeast Mitochondrial Large Ribosomal Subunit. *Science (1979)* **2014**, *343*, 1485–1489. [CrossRef]
78. Liao, M.; Cao, E.; Julius, D.; Cheng, Y. Structure of the TRPV1 Ion Channel Determined by Electron Cryo-Microscopy. *Nature* **2013**, *504*, 107–112. [CrossRef]
79. Renaud, J.-P.; Chari, A.; Ciferri, C.; Liu, W.; Rémigy, H.-W.; Stark, H.; Wiesmann, C. Cryo-EM in Drug Discovery: Achievements, Limitations and Prospects. *Nat. Rev. Drug Discov.* **2018**, *17*, 471–492. [CrossRef]
80. Pintilie, G.; Chiu, W. Validation, Analysis and Annotation of Cryo-EM Structures. *Acta Crystallogr. D Struct. Biol.* **2021**, *77*, 1142–1152. [CrossRef] [PubMed]
81. Prisant, M.G.; Williams, C.J.; Chen, V.B.; Richardson, J.S.; Richardson, D.C. New Tools in MolProbity Validation: CaBLAM for CryoEM Backbone, UnDowser to Rethink "Waters," and NGL Viewer to Recapture Online 3D Graphics. *Protein Sci.* **2020**, *29*, 315–329. [CrossRef]
82. Hryc, C.F.; Baker, M.L. Beyond the Backbone: The Next Generation of Pathwalking Utilities for Model Building in CryoEM Density Maps. *Biomolecules* **2022**, *12*, 773. [CrossRef] [PubMed]
83. Armstrong, B.D.; Han, S. Overhauser Dynamic Nuclear Polarization To Study Local Water Dynamics. *J. Am. Chem. Soc.* **2009**, *131*, 4641–4647. [CrossRef]
84. Otting, G. NMR Studies of Water Bound to Biological Molecules. *Prog. Nucl. Magn. Reson. Spectrosc.* **1997**, *31*, 259–285. [CrossRef]
85. Kovalenko, A.; Hirata, F. Three-Dimensional Density Profiles of Water in Contact with a Solute of Arbitrary Shape: A RISM Approach. *Chem. Phys. Lett.* **1998**, *290*, 237–244. [CrossRef]
86. Kovalenko, A.; Hirata, F. Self-Consistent Description of a Metal–Water Interface by the Kohn–Sham Density Functional Theory and the Three-Dimensional Reference Interaction Site Model. *J. Chem. Phys.* **1999**, *110*, 10095–10112. [CrossRef]
87. Nittinger, E.; Gibbons, P.; Eigenbrot, C.; Davies, D.R.; Maurer, B.; Yu, C.L.; Kiefer, J.R.; Kuglstatter, A.; Murray, J.; Ortwine, D.F.; et al. Water Molecules in Protein–Ligand Interfaces. Evaluation of Software Tools and SAR Comparison. *J. Comput. Aided Mol. Des.* **2019**, *33*, 307–330. [CrossRef] [PubMed]
88. Rossato, G.; Ernst, B.; Vedani, A.; Smieško, M. AcquaAlta: A Directional Approach to the Solvation of Ligand–Protein Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 1867–1881. [CrossRef]
89. Vedani, A.; Huhta, D.W. Algorithm for the Systematic Solvation of Proteins Based on the Directionality of Hydrogen Bonds. *J. Am. Chem. Soc.* **1991**, *113*, 5860–5862. [CrossRef]

90. Pitt, W.R.; Goodfellow, J.M. Modelling of Solvent Positions around Polar Groups in Proteins. *Protein Eng. Des. Sel.* **1991**, *4*, 531–537. [CrossRef] [PubMed]
91. Schymkowitz, J.W.H.; Rousseau, F.; Martins, I.C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. Prediction of Water and Metal Binding Sites and Their Affinities by Using the Fold-X Force Field. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10147–10152. [CrossRef] [PubMed]
92. Forli, S.; Olson, A.J. A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking. *J. Med. Chem.* **2012**, *55*, 623–638. [CrossRef]
93. van Dijk, A.D.J.; Bonvin, A.M.J.J. Solvated Docking: Introducing Water into the Modelling of Biomolecular Complexes. *Bioinformatics* **2006**, *22*, 2340–2347. [CrossRef]
94. Huggins, D.J.; Tidor, B. Systematic Placement of Structural Water Molecules for Improved Scoring of Protein-Ligand Interactions. *Protein Eng. Des. Sel.* **2011**, *24*, 777–789. [CrossRef]
95. Li, Y.; Gao, Y.; Holloway, M.K.; Wang, R. Prediction of the Favorable Hydration Sites in a Protein Binding Pocket and Its Application to Scoring Function Formulation. *J. Chem. Inf. Model.* **2020**, *60*, 4359–4375. [CrossRef]
96. Virtanen, J.J.; Makowski, L.; Sosnick, T.R.; Freed, K.F. Modeling the Hydration Layer around Proteins: HyPred. *Biophys. J.* **2010**, *99*, 1611–1619. [CrossRef]
97. Rarey, M.; Kramer, B.; Lengauer, T. The Particle Concept: Placing Discrete Water Molecules during Protein-Ligand Docking Predictions. *Proteins: Struct. Funct. Genet.* **1999**, *34*, 17–28. [CrossRef]
98. Wei, W.; Luo, J.; Waldispühl, J.; Moitessier, N. Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes. *J. Chem. Inf. Model.* **2019**, *59*, 2941–2951. [CrossRef]
99. Bayden, A.S.; Moustakas, D.T.; Joseph-McCarthy, D.; Lamb, M.L. Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model.* **2015**, *55*, 1552–1565. [CrossRef]
100. Ross, G.A.; Morris, G.M.; Biggin, P.C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS ONE* **2012**, *7*, e32036. [CrossRef]
101. Mason, J.S.; Bortolato, A.; Weiss, D.R.; Deflorian, F.; Tehan, B.; Marshall, F.H. High End GPCR Design: Crafted Ligand Design and Druggability Analysis Using Protein Structure, Lipophilic Hotspots and Explicit Water Networks. *In Silico Pharmacol.* **2013**, *1*, 23. [CrossRef]
102. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J.S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294. [CrossRef]
103. Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**, *58*, 1625–1637. [CrossRef] [PubMed]
104. Bui, H.-H.; Schiewe, A.J.; Haworth, I.S. WATGEN: An Algorithm for Modeling Water Networks at Protein-Protein Interfaces. *J. Comput. Chem.* **2007**, *28*, 2241–2251. [CrossRef] [PubMed]
105. Hu, B.; Lill, M.A. WATsite: Hydration Site Prediction Program with PyMOL Interface. *J. Comput. Chem.* **2014**, *35*, 1255–1260. [CrossRef] [PubMed]
106. Barillari, C.; Taylor, J.; Viner, R.; Essex, J.W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587. [CrossRef]
107. Huang, P.; Xing, H.; Zou, X.; Han, Q.; Liu, K.; Sun, X.; Wu, J.; Fan, J. Accurate Prediction of Hydration Sites of Proteins Using Energy Model with Atom Embedding. *Front. Mol. Biosci.* **2021**, *8*, 756075. [CrossRef]
108. Lazaridis, T.; Karplus, M. Thermodynamics of Protein Folding: A Microscopic View. *Biophys. Chem.* **2002**, *100*, 367–395. [CrossRef]
109. Warshel, A. Energetics of Enzyme Catalysis. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 5250–5254. [CrossRef]
110. Cramer, C.J.; Truhlar, D.G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161–2200. [CrossRef] [PubMed]
111. Spoel, D.; Zhang, J.; Zhang, H. Quantitative Predictions from Molecular Simulations Using Explicit or Implicit Interactions. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1560. [CrossRef]
112. Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *J. Chem. Theory Comput.* **2017**, *13*, 1034–1043. [CrossRef] [PubMed]
113. Kuhn, B.; Kollman, P.A. A Ligand That Is Predicted to Bind Better to Avidin than Biotin: Insights from Computational Fluorine Scanning. *J. Am. Chem. Soc.* **2000**, *122*, 3909–3916. [CrossRef]
114. Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J.Z.H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119*, 9478–9508. [CrossRef]
115. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef] [PubMed]
116. Stouten, P.F.W.; Frömmel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simul.* **1993**, *10*, 97–120. [CrossRef]
117. Mehler, E.L.; Solmajer, T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Eng. Des. Sel.* **1991**, *4*, 903–910. [CrossRef]
118. Allen, W.J.; Balius, T.E.; Mukherjee, S.; Brozell, S.R.; Moustakas, D.T.; Lang, P.T.; Case, D.A.; Kuntz, I.D.; Rizzo, R.C. DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132–1156. [CrossRef]

119. *Molecular Operating Environment (MOE)*; 2022.02 Chemical Computing Group ULC: Montreal, QC, Canada, 2023; Available online: https://www.chemcomp.com/Research-Citing_MOE.htm (accessed on 18 July 2023).
120. Corbeil, C.R.; Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449. [CrossRef]
121. Liu, C.; Wrobleski, S.T.; Lin, J.; Ahmed, G.; Metzger, A.; Wityak, J.; Gillooly, K.M.; Shuster, D.J.; McIntyre, K.W.; Pitt, S.; et al. 5-Cyanopyrimidine Derivatives as a Novel Class of Potent, Selective, and Orally Active Inhibitors of P38α MAP Kinase. *J. Med. Chem.* **2005**, *48*, 6261–6270. [CrossRef]
122. Nasief, N.N.; Tan, H.; Kong, J.; Hangauer, D. Water Mediated Ligand Functional Group Cooperativity: The Contribution of a Methyl Group to Binding Affinity Is Enhanced by a COO—Group Through Changes in the Structure and Thermodynamics of the Hydration Waters of Ligand–Thermolysin Complexes. *J. Med. Chem.* **2012**, *55*, 8283–8302. [CrossRef] [PubMed]
123. Berendsen, H.J.C.; Postma, J.P.M.; van Gunsteren, W.F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Springer: Berlin/Heidelberg, Germany, 1981; pp. 331–342.
124. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
125. Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102*, 3531–3541. [CrossRef]
126. Nguyen, C.N.; Kurtzman Young, T.; Gilson, M.K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril. *J. Chem. Phys.* **2012**, *137*, 044101. [CrossRef] [PubMed]
127. Murphy, R.B.; Repasky, M.P.; Greenwood, J.R.; Tubert-Brohman, I.; Jerome, S.; Annabhimoju, R.; Boyles, N.A.; Schmitz, C.D.; Abel, R.; Farid, R.; et al. WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand–Receptor Docking. *J. Med. Chem.* **2016**, *59*, 4364–4384. [CrossRef]
128. Bucher, D.; Stouten, P.; Triballeau, N. Shedding Light on Important Waters for Drug Design: Simulations versus Grid-Based Methods. *J. Chem. Inf. Model.* **2018**, *58*, 692–699. [CrossRef]
129. Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235. [CrossRef]
130. Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc. Perkin Trans. 2* **1993**, 799–805. [CrossRef]
131. Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094. [CrossRef] [PubMed]
132. Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-PCM Solvation Model. *J. Comput. Chem.* **2003**, *24*, 669–681. [CrossRef]
133. Marenich, A.V.; Cramer, C.J.; Truhlar, D.G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396. [CrossRef]
134. Dobeš, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P. Semiempirical Quantum Mechanical Method PM6-DH2X Describes the Geometry and Energetics of CK2-Inhibitor Complexes Involving Halogen Bonds Well, While the Empirical Potential Fails. *J. Phys. Chem. B* **2011**, *115*, 8581–8589. [CrossRef]
135. Pecina, A.; Meier, R.; Fanfrlík, J.; Lepšík, M.; Řezáč, J.; Hobza, P.; Baldauf, C. The SQM/COSMO Filter: Reliable Native Pose Identification Based on the Quantum-Mechanical Description of Protein–Ligand Interactions and Implicit COSMO Solvation. *Chem. Commun.* **2016**, *52*, 3312–3315. [CrossRef]
136. Fanfrlík, J.; Bronowska, A.K.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. A Reliable Docking/Scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands. *J. Phys. Chem. B* **2010**, *114*, 12666–12678. [CrossRef]
137. Urquiza-Carvalho, G.A.; Fragoso, W.D.; Rocha, G.B. Assessment of Semiempirical Enthalpy of Formation in Solution as an Effective Energy Function to Discriminate Native-like Structures in Protein Decoy Sets. *J. Comput. Chem.* **2016**, *37*, 1962–1972. [CrossRef]
138. Sulimov, A.V.; Kutov, D.C.; Katkova, E.V.; Ilin, I.S.; Sulimov, V.B. New Generation of Docking Programs: Supercomputer Validation of Force Fields and Quantum-Chemical Methods for Docking. *J. Mol. Graph. Model.* **2017**, *78*, 139–147. [CrossRef]
139. Sulimov, A.V.; Kutov, D.C.; Taschilova, A.S.; Ilin, I.S.; Stolpovskaya, N.V.; Shikhaliev, K.S.; Sulimov, V.B. In Search of Non-Covalent Inhibitors of SARS-CoV-2 Main Protease: Computer Aided Drug Design Using Docking and Quantum Chemistry. *Supercomput. Front. Innov.* **2020**, *7*. [CrossRef]
140. Stewart, J.J.P. Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations. *Int. J. Quantum Chem.* **1996**, *58*, 133–146. [CrossRef]
141. Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. Semiempirical Calculations of Binding Enthalpy for Protein-Ligand Complexes. *Int. J. Quantum Chem.* **2004**, *97*, 747–763. [CrossRef]
142. Nikitina, E.; Sulimov, V.; Grigoriev, F.; Kondakova, O.; Luschenka, S. Mixed Implicit/Explicit Solvation Modelsin Quantum Mechanical Calculations OfBinding Enthalpy for Protein–LigandComplexes. *Int. J. Quantum Chem.* **2006**, *106*, 1943–1963. [CrossRef]

143. Horváth, I.; Jeszenői, N.; Bálint, M.; Paragi, G.; Hetényi, C. A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level. *Int. J. Mol. Sci.* **2019**, *20*, 4384. [CrossRef] [PubMed]
144. Cavasotto, C.N.; Aucar, M.G. High-Throughput Docking Using Quantum Mechanical Scoring. *Front. Chem.* **2020**, *8*, 246. [CrossRef]
145. Hylsová, M.; Carbain, B.; Fanfrlík, J.; Musilová, L.; Haldar, S.; Köprülüoğlu, C.; Ajani, H.; Brahmkshatriya, P.S.; Jorda, R.; Kryštof, V.; et al. Explicit Treatment of Active-Site Waters Enhances Quantum Mechanical/Implicit Solvent Scoring: Inhibition of CDK2 by New Pyrazolo[1,5-a]Pyrimidines. *Eur. J. Med. Chem.* **2017**, *126*, 1118–1128. [CrossRef]
146. Pinzi, L.; Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* **2019**, *20*, 4331. [CrossRef] [PubMed]
147. Śledź, P.; Caflisch, A. Protein Structure-Based Drug Design: From Docking to Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102. [CrossRef] [PubMed]
148. Dong, D.; Xu, Z.; Zhong, W.; Peng, S. Parallelization of Molecular Docking: A Review. *Curr. Top. Med. Chem.* **2018**, *18*, 1015–1028. [CrossRef]
149. Ballante, F.; Kooistra, A.J.; Kampen, S.; de Graaf, C.; Carlsson, J. Structure-Based Virtual Screening for Ligands of G Protein–Coupled Receptors: What Can Molecular Docking Do for You? *Pharmacol. Rev.* **2021**, *73*, 1698–1736. [CrossRef]
150. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949. [CrossRef]
151. Potlitz, F.; Link, A.; Schulig, L. Advances in the Discovery of New Chemotypes through Ultra-Large Library Docking. *Expert. Opin. Drug Discov.* **2023**, *18*, 303–313. [CrossRef]
152. Lu, S.-Y.; Jiang, Y.-J.; Lv, J.; Zou, J.-W.; Wu, T.-X. Role of Bridging Water Molecules in GSK3β-Inhibitor Complexes: Insights from QM/MM, MD, and Molecular Docking Studies. *J. Comput. Chem.* **2011**, *32*, 1907–1918. [CrossRef] [PubMed]
153. Santos, R.; Hritz, J.; Oostenbrink, C. Role of Water in Molecular Docking Simulations of Cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, *50*, 146–154. [CrossRef] [PubMed]
154. Kumar, A.; Zhang, K.Y.J. Investigation on the Effect of Key Water Molecules on Docking Performance in CSARdock Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1880–1892. [CrossRef] [PubMed]
155. de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N.P.E. Binding Mode Prediction of Cytochrome P450 and Thymidine Kinase Protein−Ligand Complexes by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **2005**, *48*, 2308–2318. [CrossRef] [PubMed]
156. Birch, L.; Murray, C.; Hartshorn, M.; Tickle, I.; Verdonk, M. Sensitivity of Molecular Docking to Induced Fit Effects in Influenza Virus Neuraminidase. *J. Comput. Aided Mol. Des.* **2002**, *16*, 855–869. [CrossRef] [PubMed]
157. Lu, J.; Hou, X.; Wang, C.; Zhang, Y. Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. *J. Chem. Inf. Model.* **2019**, *59*, 4540–4549. [CrossRef]
158. Sun, H.; Zhao, L.; Peng, S.; Huang, N. Incorporating Replacement Free Energy of Binding-Site Waters in Molecular Docking. *Proteins: Struct. Funct. Bioinform.* **2014**, *82*, 1765–1776. [CrossRef]
159. Mahmoud, A.H.; Masters, M.R.; Yang, Y.; Lill, M.A. Elucidating the Multiple Roles of Hydration for Accurate Protein-Ligand Binding Prediction via Deep Learning. *Commun. Chem.* **2020**, *3*, 19. [CrossRef]
160. Schnecke, V.; Kuhn, L.A. Virtual Screening with Solvation and Ligand-Induced Complementarity. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?* Kluwer Academic Publishers: Dordrecht, The Netherlands, 2010; pp. 171–190.
161. Therrien, E.; Weill, N.; Tomberg, A.; Corbeil, C.R.; Lee, D.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 7. Impact of Protein Flexibility and Water Molecules on Docking-Based Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2014**, *54*, 3198–3210. [CrossRef]
162. Lie, M.A.; Thomsen, R.; Pedersen, C.N.S.; Schiøtt, B.; Christensen, M.H. Molecular Docking with Ligand Attached Water Molecules. *J. Chem. Inf. Model.* **2011**, *51*, 909–917. [CrossRef] [PubMed]
163. Huang, N.; Shoichet, B.K. Exploiting Ordered Waters in Molecular Docking. *J. Med. Chem.* **2008**, *51*, 4862–4865. [CrossRef] [PubMed]
164. Davis, I.W.; Baker, D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **2009**, *385*, 381–392. [CrossRef]
165. Lemmon, G.; Meiler, J. Towards Ligand Docking Including Explicit Interface Water Molecules. *PLoS ONE* **2013**, *8*, e67536. [CrossRef]
166. Verdonk, M.L.; Chessari, G.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Nissink, J.W.M.; Taylor, R.D.; Taylor, R. Modeling Water Molecules in Protein−Ligand Docking Using GOLD. *J. Med. Chem.* **2005**, *48*, 6504–6515. [CrossRef]
167. Stanzione, F.; Giangreco, I.; Cole, J.C. Use of Molecular Docking Computational Tools in Drug Discovery. *Prog. Med. Chem.* **2021**, *60*, 273–343.
168. Roberts, B.C.; Mancera, R.L. Ligand−Protein Docking with Water Molecules. *J. Chem. Inf. Model.* **2008**, *48*, 397–408. [CrossRef]
169. Hartshorn, M.J.; Verdonk, M.L.; Chessari, G.; Brewerton, S.C.; Mooij, W.T.M.; Mortenson, P.N.; Murray, C.W. Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726–741. [CrossRef] [PubMed]

170. Thilagavathi, R.; Mancera, R.L. Ligand−Protein Cross-Docking with Water Molecules. *J. Chem. Inf. Model.* **2010**, *50*, 415–421. [CrossRef] [PubMed]
171. Kastritis, P.L.; Visscher, K.M.; van Dijk, A.D.J.; Bonvin, A.M.J.J. Solvated Protein-Protein Docking Using Kyte-Doolittle-Based Water Preferences. *Proteins: Struct. Funct. Bioinform.* **2013**, *81*, 510–518. [CrossRef]
172. Pavlovicz, R.E.; Park, H.; DiMaio, F. Efficient Consideration of Coordinated Water Molecules Improves Computational Protein-Protein and Protein-Ligand Docking Discrimination. *PLoS Comput. Biol.* **2020**, *16*, e1008103. [CrossRef] [PubMed]
173. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489. [CrossRef]
174. Raymer, M.L.; Sanschagrin, P.C.; Punch, W.F.; Venkataraman, S.; Goodman, E.D.; Kuhn, L.A. Predicting Conserved Water-Mediated and Polar Ligand Interactions in Proteins Using a K-Nearest-Neighbors Genetic Algorithm. *J. Mol. Biol.* **1997**, *265*, 445–464. [CrossRef]
175. Zsidó, B.Z.; Börzsei, R.; Szél, V.; Hetényi, C. Determination of Ligand Binding Modes in Hydrated Viral Ion Channels to Foster Drug Design and Repositioning. *J. Chem. Inf. Model.* **2021**, *61*, 4011–4022. [CrossRef]
176. Thomaston, J.L.; Polizzi, N.F.; Konstantinidi, A.; Wang, J.; Kolocouris, A.; DeGrado, W.F. Inhibitors of the M2 Proton Channel Engage and Disrupt Transmembrane Networks of Hydrogen-Bonded Waters. *J. Am. Chem. Soc.* **2018**, *140*, 15219–15226. [CrossRef]
177. Bello, M.; Martínez-Archundia, M.; Correa-Basurto, J. Automated Docking for Novel Drug Discovery. *Expert. Opin. Drug Discov.* **2013**, *8*, 821–834. [CrossRef]
178. Yuriev, E.; Agostino, M.; Ramsland, P.A. Challenges and Advances in Computational Docking: 2009 in Review. *J. Mol. Recognit.* **2011**, *24*, 149–164. [CrossRef]
179. Hetényi, C.; Paragi, G.; Maran, U.; Timár, Z.; Karelson, M.; Penke, B. Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins. *J. Am. Chem. Soc.* **2006**, *128*, 1233–1239. [CrossRef] [PubMed]
180. Young, T.; Abel, R.; Kim, B.; Berne, B.; Friesner, R. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein–Ligand Binding. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 808–813. [CrossRef] [PubMed]
181. Wang, J.; Kang, X.; Kuntz, I.D.; Kollman, P.A. Hierarchical Database Screenings for HIV-1 Reverse Transcriptase Using a Pharmacophore Model, Rigid Docking, Solvation Docking, and MM−PB/SA. *J. Med. Chem.* **2005**, *48*, 2432–2444. [CrossRef]
182. Huang, N.; Kalyanaraman, C.; Irwin, J.J.; Jacobson, M.P. Physics-Based Scoring of Protein−Ligand Complexes: Enrichment of Known Inhibitors in Large-Scale Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 243–253. [CrossRef] [PubMed]
183. Kalyanaraman, C.; Bernacki, K.; Jacobson, M.P. Virtual Screening against Highly Charged Active Sites: Identifying Substrates of Alpha−Beta Barrel Enzymes. *Biochemistry* **2005**, *44*, 2059–2071. [CrossRef]
184. Perola, E. Minimizing False Positives in Kinase Virtual Screens. *Proteins Struct. Funct. Bioinform.* **2006**, *64*, 422–435. [CrossRef]
185. Collie, G.W.; Parkinson, G.N. The Application of DNA and RNA G-Quadruplexes to Therapeutic Medicines. *Chem. Soc. Rev.* **2011**, *40*, 5867. [CrossRef] [PubMed]
186. Dasari, S.; Bernard Tchounwou, P. Cisplatin in Cancer Therapy: Molecular Mechanisms of Action. *Eur. J. Pharmacol.* **2014**, *740*, 364–378. [CrossRef]
187. Howe, J.A.; Wang, H.; Fischmann, T.O.; Balibar, C.J.; Xiao, L.; Galgoci, A.M.; Malinverni, J.C.; Mayhood, T.; Villafania, A.; Nahvi, A.; et al. Selective Small-Molecule Inhibition of an RNA Structural Element. *Nature* **2015**, *526*, 672–677. [CrossRef]
188. Wang, M.; Yu, Y.; Liang, C.; Lu, A.; Zhang, G. Recent Advances in Developing Small Molecules Targeting Nucleic Acid. *Int. J. Mol. Sci.* **2016**, *17*, 779. [CrossRef]
189. Feng, Y.; Yan, Y.; He, J.; Tao, H.; Wu, Q.; Huang, S.-Y. Docking and Scoring for Nucleic Acid–Ligand Interactions: Principles and Current Status. *Drug Discov. Today* **2022**, *27*, 838–847. [CrossRef]
190. Ran, X.; Gestwicki, J.E. Inhibitors of Protein–Protein Interactions (PPIs): An Analysis of Scaffold Choices and Buried Surface Area. *Curr. Opin. Chem. Biol.* **2018**, *44*, 75–86. [CrossRef]
191. Li, Y.; Shen, J.; Sun, X.; Li, W.; Liu, G.; Tang, Y. Accuracy Assessment of Protein-Based Docking Programs against RNA Targets. *J. Chem. Inf. Model.* **2010**, *50*, 1134–1146. [CrossRef] [PubMed]
192. Mayol, G.F.; Defelipe, L.A.; Arcon, J.P.; Turjanski, A.G.; Marti, M.A. Solvent Sites Improve Docking Performance of Protein–Protein Complexes and Protein–Protein Interface-Targeted Drugs. *J. Chem. Inf. Model.* **2022**, *62*, 3577–3588. [CrossRef] [PubMed]
193. Parikh, H.I.; Kellogg, G.E. Intuitive, but Not Simple: Including Explicit Water Molecules in Protein-Protein Docking Simulations Improves Model Quality. *Proteins: Struct. Funct. Bioinform.* **2014**, *82*, 916–932. [CrossRef]
194. Kyte, J.; Doolittle, R.F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

D16

hetenyi.csaba_83_23

Bioinformatics, 31(12), 2015, 1959–1965
doi: 10.1093/bioinformatics/btv093
Advance Access Publication Date: 13 February 2015
Original Paper

Structural bioinformatics

Mobility-based prediction of hydration structures of protein surfaces

Norbert Jeszenői[1], István Horváth[2], Mónika Bálint[3], David van der Spoel[4] and Csaba Hetényi[5,*]

[1]Department of Genetics, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, [2]Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720 Szeged, [3]Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary, [4]Uppsala Center for Computational Chemistry, Science for Life Laboratory, Department of Cell and Molecular Biology, University of Uppsala, Box 596, SE-75124 Uppsala, Sweden, and [5]MTA-ELTE Molecular Biophysics Research Group, Hungarian Academy of Sciences, Pázmány sétány 1/C, 1117 Budapest, Hungary

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on November 4, 2014; revised on January 9, 2015; accepted on February 10, 2015

Abstract

Motivation: Hydration largely determines solubility, aggregation of proteins and influences interactions between proteins and drug molecules. Despite the importance of hydration, structural determination of hydration structure of protein surfaces is still challenging from both experimental and theoretical viewpoints. The precision of experimental measurements is often affected by fluctuations and mobility of water molecules resulting in uncertain assignment of water positions.

Results: Our method can utilize mobility as an information source for the prediction of hydration structure. The necessary information can be produced by molecular dynamics simulations accounting for all atomic interactions including water–water contacts. The predictions were validated and tested by comparison to more than 1500 crystallographic water positions in 20 hydrated protein molecules including enzymes of biomedical importance such as cyclin-dependent kinase 2. The agreement with experimental water positions was larger than 80% on average. The predictions can be particularly useful in situations where no or limited experimental knowledge is available on hydration structures of molecular surfaces.

Availability and implementation: The method is implemented in a standalone C program MobyWat released under the GNU General Public License, freely accessible with full documentation at http://www.mobywat.com.

Contact: csabahete@yahoo.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Water molecules located on protein surfaces play fundamental structural and functional roles in biology. For example, hydrogen bonds formed by waters stabilize protein structure (Nisius and Grzesiek, 2012) and affect folding (Cheung *et al.*, 2002; Levy and Onuchic, 2006). Surface water molecules are mediators of the assembly of β-amyloid protofilaments of Alzheimer's disease (Thirumalai *et al.*, 2011) and there is evidence that structurally conserved waters are parts of electron transfer networks (Antonyuk *et al.*, 2013) such as respiratory chain (de la Lande *et al.*, 2010). Structures of many G-protein-coupled receptors are also stabilized by hydration (Angel *et al.*, 2009). A recent study (Xu and Leitner, 2014) suggests that structural water molecules are also involved in thermal conductance of proteins, in photochemistry, as well as playing a fundamental role in charge transfer, allostery and energy flow (Fang *et al.*, 2009).

Water molecules are often considered essential parts of the protein structure (Petsko and Ringe, 2009) and the first hydration shell is a key determinant of the solubility and aggregation of solute molecules (Israelachvili and Wennerström, 1996). Protein–protein and protein–ligand interactions are influenced by surface-bound water molecules, and therefore, knowledge of their location is of great importance during structure-based drug design (Baron *et al.*, 2012; García-Sosa, 2013). Tightly bound water molecules can affect the chemical diversity of designed ligands (García-Sosa and Mancera, 2006) leading to simple rules for the use of water molecules in drug design (García-Sosa *et al.*, 2005) and also in interpretation of ligand-based pharmacophore models (Lloyd *et al.*, 2004). Inclusion of explicit water molecules in drug design (Mancera, 2007) have been thoroughly studied and was found to be of central importance in ligand–protein docking (Roberts and Mancera, 2008; Thilagavathi and Mancera, 2010).

Although hydration structure is important, it has hitherto proven to be very difficult to determine at the atomic level by experimental means largely due to mobility and complexity of interactions of water molecules located on a protein surface. The residence of a water molecule on the surface and its exchange with bulk are affected not primarily by the strength of protein–water interactions, but it is 'rather a topography that prevents the water molecule from exchanging by a cooperative mechanism' (Halle, 2004 a). Importantly, such a cooperative mechanism of exchange also governs several water–water interactions that can often be detected between surface water molecules (Finney, 1977). It is problematic to handle (and to predict) the residence of water molecules in the hydration layer of a protein using merely thermodynamic or kinetic approaches (Halle, 2004a).

Crystallography is the prime experimental method for detection of water positions, via electron density maps and used as the *de facto* standard (Savage and Wlodawer, 1986). However, there are still numerous limitations of this method coming from low resolution of large structural assemblies (Finney, 1977), assignment problems (Afonine *et al.*, 2013; Badger *et al.*, 1997), and artifacts due to cryogenic temperatures used (Halle, 2004b).

A number of computational methods have been proposed for prediction of hydration structure on protein surfaces. Such methods generally require the 'dry' protein structure as an input and provide predictions for hydration structure using a variety of algorithms. A large group of the methods uses fast and simplified approaches disregarding exchange (mobility) between surface and bulk water molecules and dynamics of the hydration structure. They assume a static picture of hydration shells and focus on finding appropriate binding sites of water molecules on the protein surface using scoring schemes, energy calculations (Schymkowitz *et al.*, 2005), prior knowledge (Pitt and Goodfellow, 1993), H-bonding information (Vedani and Huhta, 1991) or artificial neural networks (Ehrlich *et al.*, 1998). Several studies (Makarov *et al.*, 1998; Truchon *et al.*, 2014; Virtanen *et al.*, 2010) have dealt with construction and use of density distribution functions of hydration shells for different atom types occurring in proteins. Limitations of generalized, density-based approaches were discussed in detail (Henchman and McCammon, 2002). These methods ignore dynamics and cooperativity governing hydration.

With advancement of computational infrastructure and force fields, the efficiency and chemical accuracy of atomic level Monte-Carlo and molecular dynamics (MD) simulations has increased enormously in the past decades (Michel *et al.*, 2009; Pettitt and Karplus, 1987) enabling their applications in cutting edge drug design projects (Dror *et al.*, 2012). It has become a routine task to generate MD trajectories with explicit water molecules for virtually any protein of interest. Atomistic simulations of MD hold a conceptual advantage over the static or density-based (trained) methods as the mobility, a key determinant of hydration structure is described directly at atomic level. Whereas such benefits of atomic MD calculations have been extensively used in analyses (Schoenborn *et al.*, 1995), there are not many MD-based methods for prediction of the hydration structure (Abel *et al.*, 2008; Cui *et al.*, 2013; Henchman and McCammon, 2002). These approaches focus on all individual positions of hydrating water molecules and apply various evaluation schemes such as the definition of time averaged positions (Henchman and McCammon, 2002) for calculation of the hydration structure. In this study, we introduce a mobility-based atomic-level method for prediction of hydration structure of molecular surfaces using only 'dry' protein structures as input. Our method was tested on 20 proteins, and the corresponding computational procedures are provided in a program MobyWat, which can be used in conjunction with any MD software that can produce all-atom MD trajectories.

2 Algorithm

2.1 Prediction

Logging molecular movements of all water molecules during a time period provides mobility information required by the prediction process used here. Such a log-book (a trajectory) is preferably generated by MD calculations with an explicit water model. Generation of molecular trajectories was performed by the GROMACS (Hess *et al.*, 2008; Pronk *et al.*, 2013) MD package in this study. During additional post-MD and preparatory steps a standard protocol was followed (Supplementary Methods S1.2).

Mobility information of the trajectory is transformed into the hydration structure of the protein surface during the prediction process outlined in Figure 1. All predictions can be performed with the program MobyWat designed and written in C implementing the prediction protocols of this study. Detailed descriptions of the algorithms can be found in Supplementary Algorithm S2.1 and also in the User's Manual of the program.

Briefly, during the prediction procedure, MobyWat performs clustering of water molecules in candidate pools filtered from the corresponding MD frames. Besides the usual spatial position-based (POS) clustering, an identity (ID)-based algorithm was also introduced with ranking variants named all-inclusive (IDa) and elitist (IDe, Supplementary Algorithm S2.1.5). The procedure ends up in prediction lists including the coordinates and mobility values of water molecules in Protein Databank (PDB) format. A merged (MER) prediction list can be also produced combining the results of the above IDa, IDe and POS predictions.

2.2 Validation

The identification of matches between experimental and predicted water positions is used for validating algorithms of MobyWat. From the matches, a success rate (SR_X) value is calculated for a prediction list ($X =$ IDa, IDe, POS or MER, Eq. 1). The higher the SR_X value, the more successful a prediction is in comparison with crystallographic water positions. For comparison and estimation of the effect of clustering, per frame SR_n values

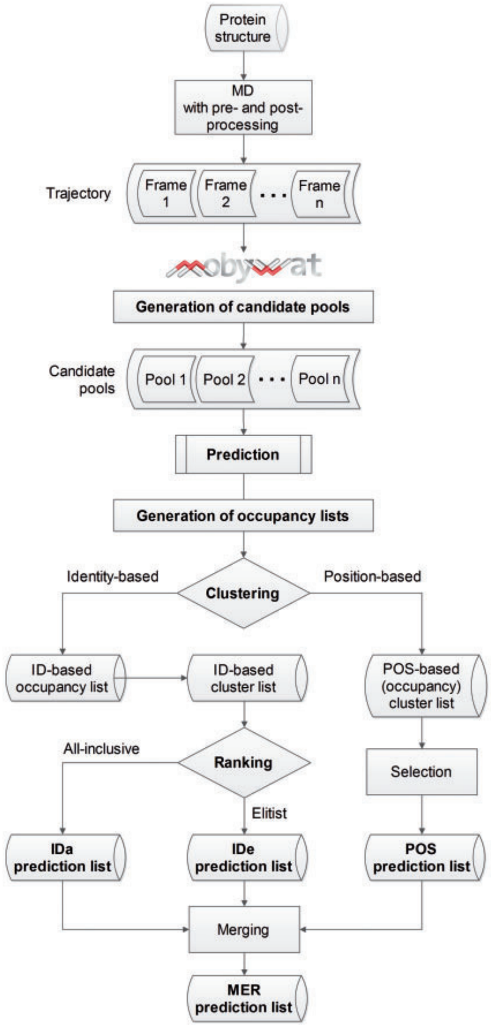


Fig. 1. The prediction process

are also calculated for each candidate pool using the analysis mode of MobyWat ($X = n$, Eq. 1).

$$\mathrm{SR}_X = 100\frac{\text{Number of matches in } X}{\text{Number of water molecules in the reference pool}}\%,$$
$$\text{where } X = \begin{cases} \text{IDa/IDe/POS/MER (prediction list in validation)}, \\ n \text{ (denotes the nth candidate pool in analysis)}. \end{cases} \quad (1)$$

Further details on validation including selection and calibration of tolerance values are described in Supplementary Algorithm S2.2, and Figure S1. Twenty reference protein systems used for validation and external tests are listed in Tables S1 and S8.

3 Results and Discussion

3.1 Sampling versus predictions

MobyWat predictions are based on atomic mobility data of all water molecules obtained from MD simulations. In this study, mobility of a predicted water molecule is defined by its occupancy value (Supplementary Eq. S2). Occupancy can be counted using a collection (sample) of hydrated protein structures. Such a sample can be collected as a series of hydrated experimental structures of the same protein (Carugo, 1999; Patel *et al.*, 2014), or generated by computational methods. Sample collection from experimental structures is not an option for this purpose as the number of hydrated structures is limited to available entries available in the PDB. In addition, if there are hydrated PDB structures available, then comparative analysis can be performed by other tools (García-Sosa *et al.*, 2003; Patel *et al.*, 2014) which proved to be useful for selection of consensus or conserved water molecules.

However, in most of the cases, only a single structure of the same protein is available. Thus, computational generation of hydration states of a protein is presently the only tractable approach to produce an appropriate sample even if only a 'dry' protein surface is available lacking experimentally determined positions of water molecules. Among computational techniques atomic level MD simulation with an explicit water model is the obvious choice of sampling method. The user needs to supply only a 'dry' protein structure and a series of hydrated protein structures are resulted as an MD trajectory. MD-generated, raw hydration structures are sometimes used even as references in comparison with other methods (Ross *et al.*, 2012). However, important parameters such as the minimal length of an MD simulation necessary for a predictive sampling have not been determined. To address this question, 1-µs-long MD simulations were performed for the protein systems of the validation set producing a sample of 1000 frames spaced at 1 ns. SR_n values were calculated for each pool according to Eq. 1 and plotted in Figure 2A for Alzheimer's amyloid precursor protein (system 2FMA). Descriptive statistics of SR_n values are provided for all validation systems in Supplementary Table S4. The descriptive statistics show a good performance of raw MD sampling with mean SR_n values ranging between 44.6 and 72.7. The SR_n values fluctuate randomly during the 1 µs time-scale of the trajectory (Fig. 2A). This finding can be explained by the short residence time of water molecules in the hydration shell of protein surface (Halle, 2004a). During 1 µs water molecules can change their positions many times, and occurrence of frames with large SR_n values (with a lot of matching water positions) is unpredictable and non-deterministic.

In summary, MD provides an appropriate sampling with good SR_n values. However, the performance of a 'prediction' based on a single frame (randomly) picked from a trajectory is non-deterministic. Thus, a valid prediction cannot be guaranteed if using only one frame. Processing several frames of a trajectory may be a better way to maximize SR and arrive at valid predictions. Accordingly, validation, calibration and measurement of the performance of prediction algorithms are described in the forthcoming sections.

3.2 Validation, performance and robustness

The prediction parameters dmax, ctol and ptol (Supplementary Table S3) were calibrated for all four types of prediction algorithms implemented in MobyWat. The calibration process is documented in Supplementary Results S3.2. Optimal sampling conditions were also determined, as the final step of the validation process. Using calibrated values of parameters, MobyWat predictions were performed for all proteins by processing 1000 coordinate frames from 1-µs-long trajectories. The results are shown for system 2FMA (Fig. 2A), and for all systems of the Validation set (Supplementary Table S4). The SR values yielded by the predictions were significantly higher than the mean SR_n from raw MD, and in many cases they were close to the maximal SR_n values. Thus, all four algorithms

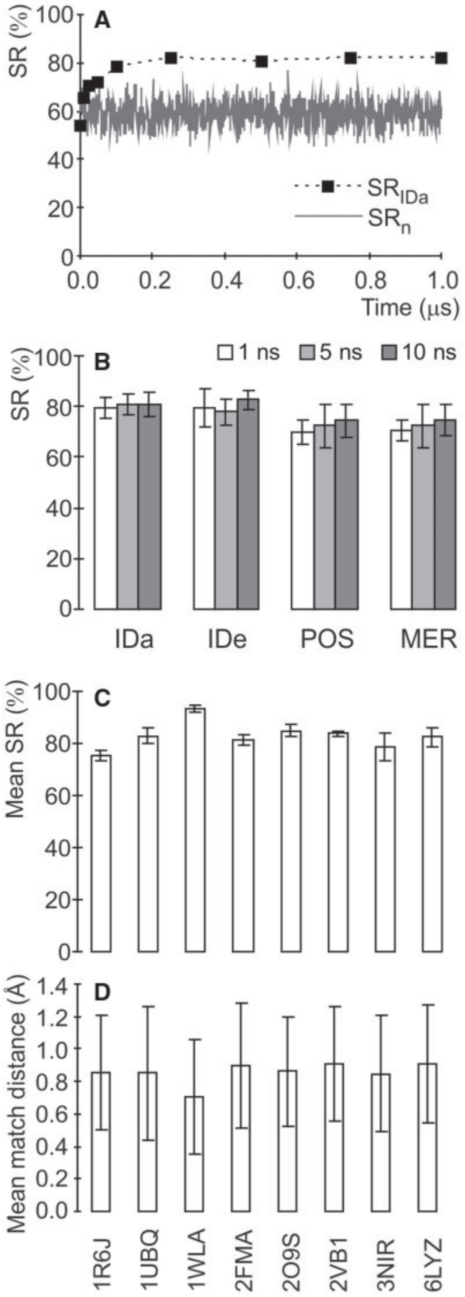


Fig. 2. (**A**) Success rates of Alzheimer's amyloid precursor protein (system 2FMA) calculated for the pools of the raw MD trajectory frames (SR$_n$) and resulting from IDa prediction of MobyWat (SR$_{IDa}$). MD trajectory of 1 μs with 1000 frames was used as a sample. (**B**) Effect of sampling time on the performance of prediction algorithms. Ten thousand frames were used for prediction with sampling times 1, 5 and 10 ns. Mean values are calculated from SRs obtained for the Validation set. Standard deviations are shown as error bars. (**C**) Reproducibility of MD sampling in terms of mean SR values calculated from three independent MD runs for each protein system. (**D**) Mean distances in matched pairs of predicted and reference water oxygen atoms plotted for all systems. Error bars denote standard deviations

Table 1. Success rates (%): statistics calculated for raw MD sampling and prediction results achieved by MobyWat

PDB ID[a]	Raw MD[b] (SR$_n$ in Eq. 1)			MobyWat[b,c]			
	Min.	Mean	Max.	SR$_{IDa}$	SR$_{IDe}$	SR$_{POS}$	SR$_{MER}$
Validation set							
1R6J	41.4	52.4	64.1	71.8	76.2	64.6	65.8
2FMA	39.4	61.5	80.3	80.3	83.6	77.1	77.1
2O9S	46.2	62.2	77.9	87.5	85.6	78.9	78.9
2VB1	44.9	59.9	71.7	82.6	84.1	79.7	80.4
3NIR	33.9	59.0	80.4	80.4	83.9	71.4	71.4
Mean	41.2	59.0	74.9	80.5	82.7	74.3	74.7
SD^V	4.9	3.9	7.0	5.7	3.7	6.3	6.1
Test set 1							
1UBQ	28.6	53.9	82.4	85.7	80.0	68.6	74.3
1WLA	31.4	68.5	94.3	94.3	88.6	82.9	82.9
6LYZ	32.2	54.2	72.9	78.0	81.5	71.2	71.2
Mean	30.7	58.8	83.2	86.0	83.4	74.2	76.1
SD^E	1.9	8.3	10.7	8.2	4.6	7.6	6.0

[a] Mean and standard deviation (SD) values of success rates were calculated for systems of external test and validation separately. [b] Sampling conditions: 10 ns MD run time, 1.0001×10^4 frames. [c] Success rates of MobyWat predictions were calculated with default mtol = 1.5 Å, $b_{max} = 30.0$ Å^2, $d_{max} = 3.5$ Å, $p_{tol} = 2.5$ Å and c_{tol} according to Supplementary Table S5.

resulted in valid predictions. Whereas sampling of 1-μs-long trajectories provided good predictions, such simulations with explicit waters can be computationally demanding. Figure 2A shows that SR$_{IDa}$ values exceeded the SR$_n$ curve and reached a plateau relatively early, after 100–200 ns sampling time. This finding suggested that shortening the sampling time should be possible without a large drop in SR of the prediction. Increasing the sampling frequency (frame count) is also a logical step to achieve reliable predictions with shortened sampling time. Indeed, results in Table 1 reveal that 10-ns-long trajectories with increased frame count yielded mean SR values of >80% for the Validation set, similarly to the 1-μs-long runs (Supplementary Table S4).

Figure 2B shows that the good performance of ID-based prediction algorithms was preserved at 1, 5 and 10 ns sampling times averaged for all systems used in Validation set. In the cases of MER and POS, there is a 5% increase in average SR values if comparing trajectories of 1 and 10 ns length. In summary, the ID-based algorithms outperformed POS and MER predictions, and they provide good predictions even at 1 ns sampling time (Table 1, Fig. 2B).

To evaluate system-independence of our method, a test of the predictions was performed. Systems of Test set 1 (1UBQ, 1WLA and 6LYZ) have relatively moderate resolution and a low number of assigned water positions per protein surface area (Supplementary Table S1). The same set had been used earlier in a study (Virtanen *et al.*, 2010) applying a solvent density-based approach. Detailed comparison of our results using the standards of the earlier study (Supplementary Results S3.3) indicates that overall performance of MobyWat is good if compared with solvent density-based results. For comparability with the above validation results performance of MobyWat on Test set 1 was also evaluated using the standards of this study and the results are listed separately in Table 1. All four algorithms provide valid predictions with SR significantly higher than average values of SR$_n$. Moreover, the mean SR values of Test set 1 are comparable to or slightly higher than mean SR values obtained for Validation set (Table 1) indicating system-independence of the method.

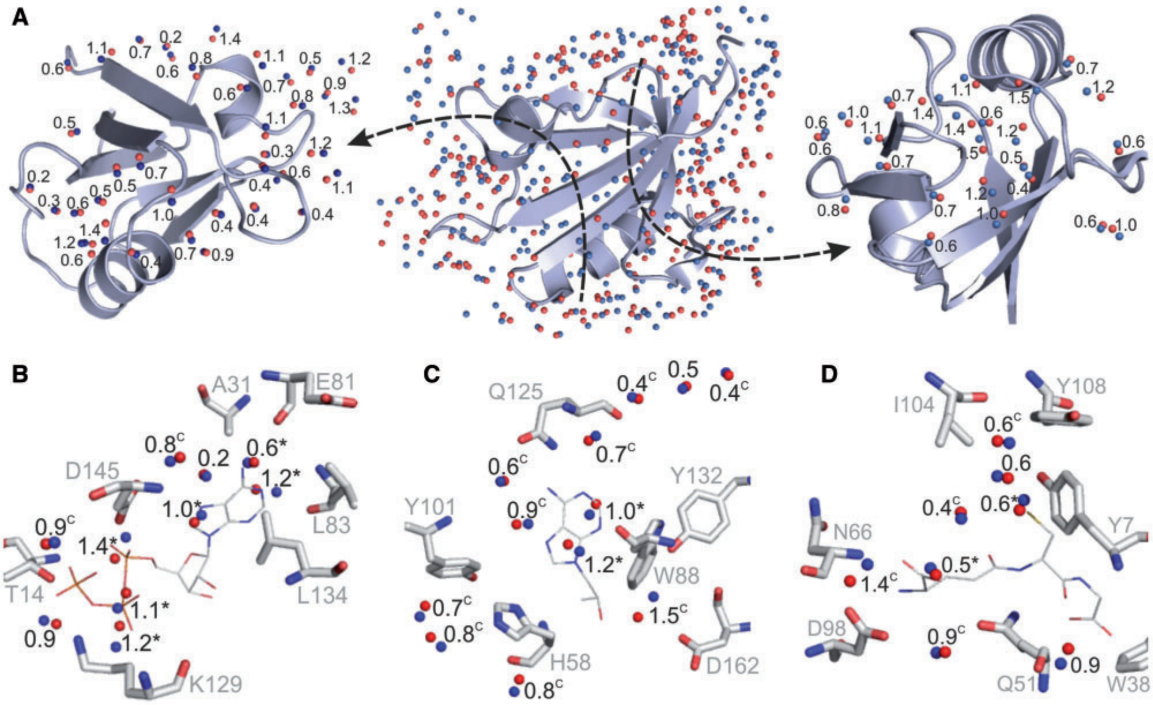


Fig. 3. (**A**) Prediction results for system 1R6J. (**B–D**) Featured binding sites of apo enzymes cyclin-dependent kinase 2 (system 1HCL, B), thymidine kinase (system 1E2H, C) and glutathione S-transferase (system 16GS, D). Ligands were inserted from superimposed ligand-bound enzyme structures (PDB codes 1HCK, 1E2I and 5GSS) for comparison with water positions. Match distances between crystallographic (red spheres) and predicted (blue spheres) water oxygen atoms are given in Å. Conserved and replaceable water molecules are marked with C and asterisk at the distance values, respectively

Reproducibility is also a key issue of robustness. As MobyWat operations are reproducible by their algorithmic definition, reproducibility tests can be performed for the MD sampling process. MD trajectories are inherently chaotic in practical applications due to hardware-dependent rounding of floating point calculations, the use of dynamic load balancing in parallel execution and so on. Therefore, it is common to repeat MD calculations with different starting atomic velocity values to test the convergence of trajectories. Practically, this can be done by selecting different seed numbers of the velocity generator routine. During the tests, three MD trajectories of all systems were produced using three different sets of initial velocities. For these trajectories, predictions were made using the top performer algorithms of Table 1.

The corresponding three SR values were averaged for all systems and plotted in Figure 2C. Their standard deviations are found to be small compared with mean values for all systems, and MD sampling is therefore shown to be reproducible in terms of SR. Improvements in the quality of force fields, in particular the introduction of polarization, may improve the reproducibility of water prediction further (Lopes *et al.*, 2013).

During validations and tests, MobyWat automatically calculated SR values using a match tolerance of 1.5 Å which is the upper limit for the detection of matches between predicted and reference water molecule pairs (Section 2.2). To further quantify the precision of matches, statistics of distances of all matched pairs of the top performer algorithms were calculated (Fig. 2D). It can be seen that mean match distances are below 1 Å for all systems. Matching water positions of one of the systems is shown in Figure 3A, and three other systems are depicted in Supplementary Figure S4.

3.3 Featured test examples

Test set 2 containing 12 proteins was assembled to further check the performance of MobyWat predictions. Using prediction algorithm IDa, a mean SR of 87% was achieved for this set. The members of Test set 2 and the resulted SR values are listed in Supplementary Table S8. Below the prediction results obtained for three enzymatic systems of Test set 2 are discussed focusing on their active sites.

Cyclin-dependent kinase 2 (Cdk2) is a key enzyme in cell cycle control and a promising drug target in oncology (Akli *et al.*, 2011) that also affects senescence (Chenette, 2010). A change of the hydration structure of the active site of Cdk2 due to ligand binding has been reported with obvious implications for drug design (Schulze-Gahmen *et al.*, 1996). A good agreement was obtained between predicted (blue spheres, Fig. 3B) and experimental reference (red spheres) water positions verifying that MobyWat accurately predicted the hydration structure of the active site of apo Cdk2 (Fig. 3B). Notably, experimental water positions were used in comparisons of Figures. 3B–D without any restrictions on their B-factors. Insertion of the ligand (ATP, thin lines in Fig. 3B) from the superimposed ATP-bound Cdk2 structure reveals that six waters (marked with asterisks in Fig. 3) are displaced by the ligand during binding. Release of such water molecules has a favorable contribution to binding entropy of the ligand, and therefore, their identification is important for thermodynamics-driven engineering of new ligands. The results were not affected by the chemical nature of ligand binding as waters replaced by both the charged phosphate moieties and the non-charged adenine ring were found correctly. This finding is in agreement with our general results showing that prediction quality is independent on the type of interacting amino acids

(Supplementary Results S3.7). The second example (Fig. 3C) features the nucleoside binding pocket of thymidine kinase from Herpes simplex type 1. This enzyme has been involved in enzyme-prodrug gene therapy of cancer (Vogt *et al.* 2000). Besides two replaceable water molecules, MobyWat precisely predicted several conserved water positions (marked with C in Fig. 3) existing in both the apo and the ligand-bound enzyme structures. Similar to the cases of replaceable water molecules, locating conserved water sites precisely is also important during the design of new ligands. A complete chain of waters leading to the active site was also predicted correctly (top-right corner of Fig. 3C). The third binding pocket in Figure 3D belongs to glutathione S-transferase, an important detoxifying enzyme (Wu and Dong, 2012). Binding chemistry of glutathione, the peptidic ligand of this enzyme is remarkably different from the previous two ligands with heteroaromatic cores (Fig. 3B and C). However, the quality of MobyWat prediction of the surrounding water positions is similarly good as it was in the other two examples.

MobyWat produces a prediction list including water positions in increasing order of mobility scores (Supplementary Algorithm S2.1.5) where experimentally verified (positive) predictions are mostly located at the top of the prediction list. It was found (Supplementary Results S3.4) that 88% of positive predictions for whole protein surfaces are located in the top 50% of the prediction list. As active sites are the most important spots on enzymes, it was also checked how mobility scores work for these specific segments of the surface. 20 of 24 (85%) of the correctly predicted water positions shown in Figures. 3B–D are located in the top 15% of the prediction lists. Thus, in the cases of active sites investigated, the mobility scores short-list the positive candidates very efficiently at the top of the prediction list. This indicates that water molecules in the active sites of enzymes are predicted with higher fidelity than other water molecules residing on the surface. This result can in part be explained by the presence of conserved water molecules surrounding the ligands, most of which are located at the top 5% of prediction lists. Notably, half of replaceable water molecules occupying active sub-sites in the apo structures were also ranked at top 10%.

4 Conclusions

MD has become an indispensable tool of prediction of structure of proteins and protein–ligand complexes (Shan *et al.*, 2011; Söderhjelm *et al.*, 2012). However, there are only a few MD-based methods for the prediction of hydration structure using explicit simulation of water contacts. Here, we presented MobyWat, a freely available program validated and tested on more than 1500 experimental water positions in 20 different protein surfaces. The prediction process of MobyWat aims at finding the least mobile (most occupied) points of the hydration structure. It was shown that MD simulation is an appropriate sampling technique for such predictions. MobyWat performs predictions using mobility information cumulated in MD trajectories. Two predictive approaches were implemented and tested. The first approach uses only spatial information (coordinates) for a candidate water position. This can be done for example by averaging trajectory frames and producing solvent densities (Virtanen *et al.*, 2010) or by clustering water molecules along the trajectory and counting frequencies of their occurrence in candidate positions. In this study, a second approach was introduced based on identification records of water molecules rather than spatial positions. On average, the identity-based predictions provided higher success rate values than positional and merged algorithms. This is probably a consequence of the position-independent philosophy of the identity-based algorithms.

Valid predictions do not require trajectories from long MD runs. The typical lifetime of a hydrogen bond is a few pico seconds only, virtually independent of the environment (van der Spoel *et al.*, 2006). Consequently, due to rapid exchange and equilibration of water positions relatively short simulations (e.g. 1–10 ns) with regular saving of coordinates suffice. Thus, with a moderate computational effort valid predictions can be achieved.

Limitations of mobility-based predictions were also investigated via an analysis of non-matched water positions of eight systems (Supplementary Results S3.7 and Appendix 2). The analysis identified location of waters above shallow protein sites and/or far from the surface to be a limiting factor in a few cases. Further work is on the way to overcome such limitations using a relative coordinate definition and testing combined MD sampling schemes.

MobyWat algorithms were coded in the portable C language. As the program has to perform calculations on numerous atoms in numerous frames (e.g. $10^4 \times 10^4$) special attention was paid to the efficient use of memory. MobyWat can be used in conjunction with any MD program as it reads frames from PDB files. However, for efficient use of memory and disk space MobyWat also reads and writes xdr-type portable binary trajectory files called xtc in GROMACS.

Mobility is often considered as a disturbing property hampering experimental determination of positions of water molecules on protein surfaces. In this study, it was shown that mobility can be utilized as an information source for prediction of hydration structure. If experimental determination of water structure is not available or incomplete, MobyWat can offer an alternative solution.

Acknowledgements

Part of the simulations were carried out on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the 'Abisko' supercomputer of the High Performance Computing Center North (HPC2N, Sweden, grant SNIC2013-26-6). We acknowledge PRACE for awarding us access to resources Monte Rosa based in Switzerland at CSCS Swiss National Supercomputing Centre, and NIIFI SC based in Hungary at NIIF National Information Infrastructure Development Institute.

Funding

The work was supported by the Hungarian Scientific Research Fund (OTKA K112807) and the MedinProt project of the Hungarian Academy of Sciences. We are thankful to the Gedeon Richter Pharmaceutical Plc. for a pre-doctoral scholarship (to N.J.).

Conflict of Interest: none declared.

References

Abel,R. *et al.* (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.*, **130**, 2817–2831.

Afonine,P.V. *et al.* (2013) Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Cryst.*, **D69**, 625–634.

Akli,S. *et al.* (2011) Cdk2 is required for breast cancer mediated by the low-molecular-weight isoform of Cyclin E. *Cancer Res.*, **71**, 3377–3386.

Angel,T.E. *et al.* (2009) Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proc. Natl Acad. Sci. USA*, **106**, 147367−14372.

Antonyuk,S.V. *et al.* (2013) Structures of protein–protein complexes involved in electron transfer. *Nature*, **496**, 123−126.

Badger,J. (1997) Modeling and refinement of water molecules and disordered solvent. *Methods Enzymol.*, **277**, 344−352.

Baron,R. *et al.* (2012) Hydrophobic association and volume-confined water molecules. In:Gohlke,H. (ed.) *Protein–Ligand Interactions*. Wiley-VCH Verlag GmbH & Co. KGaA, Wennheim.

Carugo,O. (1999) Correlation between occupancy and B-factor of water molecules in protein crystal structures. *Protein Eng.*, **12**, 1021−1024.

Chenette,E.J. (2010) Senescence: a key role for CDK2. *Nat. Rev. Cancer.*, **10**, 84.

Cheung,M.S. *et al.* (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl Acad. Sci. USA*, **99**, 685−690.

Cui,G. *et al.* (2013) SPAM: a simple approach for profiling bound water molecules. *J. Chem. Theor. Comput.*, **9**, 5539–5549.

de la Lande,A. *et al.* (2010) Surface residues dynamically organize water bridges to enhance electron transfer between proteins. *Proc. Natl Acad. Sci. USA*, **107**, 11799–11804.

Dror,R.O. *et al.* (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.*, **41**, 429−452.

Ehrlich,L. *et al.* (1998) Prediction of protein hydration sites from sequence by modular neural networks. *Protein Eng. Des. Sel.*, **11**, 11−19.

Fang,C. *et al.* (2009) Mapping GFP structure evolution during proton transfer with femtosecond Raman spectroscopy. *Nature*, **462**, 200−204.

Finney,J.L. (1977) The organization and function of water in protein crystals. *Philos. Trans. R. Soc. Lond. B*, **278**, 3−32.

García-Sosa,A.T. (2013) Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. *J. Chem. Inf. Model.*, **53**, 1388−1405.

García-Sosa,A.T., *et al.* (2003) WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein–ligand complexes. *J. Mol. Mod.*, **9**, 172−182.

Garcia-Sosa,A.T., *et al.* (2005) Including tightly-bound water molecules in de novo drug design. Exemplification through the in silico generation of poly(ADP-ribose)polymerase ligands. *J. Chem. Inf. Model.*, **45**, 624–633.

Garcia-Sosa,A.T., and Mancera,R.L. (2006) The effect of tightly-bound water molecules on scaffold diversity in the computer-aided de novo ligand design of CDK2 inhibitors. *J. Mol. Mod.* **12**, 422–431.

Halle,B. (2004a) Protein hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc. Lond. B*, **359**, 1207−1224.

Halle,B. (2004b) Biomolecular cryocrystallography: structural changes during flash-cooling. *Proc. Natl Acad. Sci. USA*, **101**, 4793−4798.

Henchman,R.H. and McCammon,J.A. (2002) Extracting hydration sites around proteins from explicit water simulations. *J. Comput. Chem.*, **23**, 861–869.

Hess,B. *et al.* (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435−447.

Israelachvili,J. and Wennerström,H. (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature*, **379**, 219−225.

Levy,Y. and Onuchic,J.N. (2006) Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 389−415.

Lloyd,D.G. *et al.* (2004) The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comput. Aided Mol. Des.*, **18**, 89–100.

Lopes,P.E.M. *et al.* (2013) Polarizable force field for peptides and proteins based on the classical drude oscillator. *J. Chem. Theor. Comput.*, **9**, 5430−5449.

Makarov,V.A. *et al.* (1998) Reconstructing the protein-water interface. *Biopolymers*, **45**, 469–478.

Mancera,R.L. (2007) Molecular modeling of hydration in drug design. *Curr. Opin. Drug Discov. Dev.*, **10**, 275–280.

Michel,J. *et al.* (2009) Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.*, **131**, 15403–15411.

Nisius,L. and Grzesiek,S. (2012) Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network. *Nat. Chem.*, **4**, 711−717.

Patel,H. *et al.* (2014) PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics*, **30**, 2978−2980.

Petsko,G.A. and Ringe,D. (2009) *Protein Structure and Function.* Oxford University Press Inc., New York.

Pettitt,B.M. and Karplus,M. (1987) The structure of water surrounding a peptide: a theoretical approach. *Chem. Phys. Lett.*, **136**, 383–386.

Pitt,W.R. and Goodfellow,J.M. (1991) Modelling of solvent positions around polar groups in proteins. *Protein Eng.*, **4**, 531−537.

Pronk,S. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics,* **29**, 845−854.

Roberts,B.C. and Mancera,R.L. (2008) Ligand-protein docking with water molecules. *J. Chem. Inf. Model.*, **48**, 397–408.

Ross,G.A. *et al.* (2012) Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS ONE*, **7**, e32036.

Savage,H. and Wlodawer,A. (1986) Determination of water structure around biomolecules using x-ray and neutron diffraction methods. *Methods Enzymol.*, **127**, 162–183.

Schoenborn,B.P. *et al.* (1995) Hydration in protein crystallography. *Prog. Biophys. Mol. Biol.*, **64**, 105−119.

Schulze-Gahmen,U. *et al.* (1996) High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J. Med. Chem.* **39**, 4540–4546.

Schymkowitz,J.W.H. *et al.* (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.

Shan,Y. *et al.* (2011) How does a drug molecule find its target binding site? *J. Am. Chem. Soc.*, **133**, 9181−9183.

Söderhjelm,P. *et al.* (2012) Locating binding poses in protein–ligand systems using reconnaissance metadynamics. *Proc. Natl Acad. Sci. USA*, **109**, 5170−5175.

Thilagavathi,R. and Mancera,R.L. (2010) Ligand–protein cross docking with water molecules. *J. Chem. Inf. Model.*, **50**, 415–421.

Thirumalai,D. *et al.* (2011) Role of water in protein aggregation and amyloid polymorphism. *Acc. Chem. Res.*, **45**, 83−92.

Truchon,J-F. *et al.* (2014) A cavity corrected 3D-RISM functional for accurate solvation free energies. *J. Chem. Theor. Comput.*, **10**, 934–941.

van der Spoel,D. *et al.* (2006) Thermodynamics of hydrogen bonding in hydrophilic and hydrophobic media. *J. Phys. Chem. B*, **110**, 4393−4398.

Vedani,A. and Huhta,D.W. (1991) An algorithm for the systematic solvation of proteins based on the directionality of hydrogen bonds. *J. Am. Chem. Soc.*, **113**, 5860−5862.

Virtanen,J.J. *et al.* (2010) Modeling the hydration layer around proteins: HyPred. *Biophys. J.*, **99**, 1611–1619.

Vogt,J. *et al.* (2000) Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. *Proteins*, **41**, 545–553.

Wu,B. and Dong,D. (2012) Human cytosolic glutathione transferases: structure, function, and drug discovery. *Trends Pharmacol. Sci.*, **33**, 656–668.

Xu,Y. and Leitner,M.D. (2014) Vibrational energy flow through the green fluorescent protein−water interface: communication maps and thermal boundary conductance. *J. Phys. Chem. B.*, **118**, 7818−7826.

hetenyi.csaba_83_23

D17

hetenyi.csaba_83_23

Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

Exploration of Interfacial Hydration Networks of Target–Ligand Complexes

Norbert Jeszenői,[†,‡] Mónika Bálint,[§] István Horváth,[‖] David van der Spoel,[⊥] and Csaba Hetényi*,[#]

[†]Department of Genetics, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary

[‡]MTA NAP-B Molecular Neuroendocrinology Group, Institute of Physiology, Szentágothai Research Center, Center for Neuroscience, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary

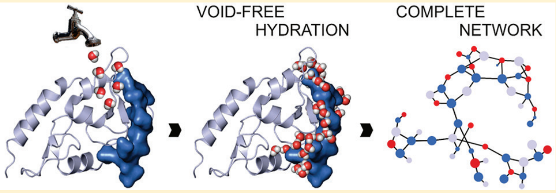
[§]Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary

[‖]Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720 Szeged, Hungary

[⊥]Uppsala Center for Computational Chemistry, Science for Life Laboratory, Department of Cell and Molecular Biology, University of Uppsala, Box 596, SE-75124 Uppsala, Sweden

[#]MTA-ELTE Molecular Biophysics Research Group, Hungarian Academy of Sciences, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary

S Supporting Information



ABSTRACT: Interfacial hydration strongly influences interactions between biomolecules. For example, drug–target complexes are often stabilized by hydration networks formed between hydrophilic residues and water molecules at the interface. Exhaustive exploration of hydration networks is challenging for experimental as well as theoretical methods due to high mobility of participating water molecules. In the present study, we introduced a tool for determination of the complete, void-free hydration structures of molecular interfaces. The tool was applied to 31 complexes including histone proteins, a HIV-1 protease, a G-protein-signaling modulator, and peptide ligands of various lengths. The complexes contained 344 experimentally determined water positions used for validation, and excellent agreement with these was obtained. High-level cooperation between interfacial water molecules was detected by a new approach based on the decomposition of hydration networks into static and dynamic network regions (subnets). Besides providing hydration structures at the atomic level, our results uncovered hitherto hidden networking fundaments of integrity and stability of complex biomolecular interfaces filling an important gap in the toolkit of drug design and structural biochemistry. The presence of continuous, static regions of the interfacial hydration network was found necessary also for stable complexes of histone proteins participating in chromatin assembly and epigenetic regulation.

INTRODUCTION

Complex systems of nature may often be described as networks of interdependent nodes.[1−3] Mapping and description of complex, dynamic networks is challenging.[4−6] Hydration networks of water molecules of complex interfaces play a central role in establishing and mediation of molecular interactions, and their exploration is of general importance in materials sciences and biomedicine.[7−11] Interfacial hydration networks have to be considered in drug development[12−19] when studying amyloidogenesis in Alzheimer's disease,[20] targeting protein−protein interactions,[21,22] etc.

Exploration of hydration networking would require precise knowledge of the complete hydration structure of the interface. However, experimental assignment of all water positions is problematic[23−26] due to well-documented limitations of X-ray crystallography[27] often caused by the inherent mobility of water. Similarly, many of the available theoretical methods also suffer from improper or lacking modeling of water−water contacts and mobility. It was shown that molecular dynamics (MD) is able to calculate the latter factors accurately, suggesting a solution for the hydration problem.[28−30] The number of MD-based approaches of mapping hydration is still moderate, and the experimental validation using reference water positions is very limited (Table S1).

Functional characterization of hydration networks may include either thermodynamic analyses or descriptions of the topography of intermolecular interactions.[31] Hydration thermodynamics has been thoroughly studied.[32−35] Graph topology-based methods have been developed[36,37] for structure−activity relationships of small molecules and H-bonding patterns in proteins.[38] Here, we provide an approach for the topological description of co-operation of water molecules validated by characterization of hydration networks of 31 large biomolecular complexes (Table 1) of protein targets and peptide/protein ligands. The complexes have hydrophilic backbone amide groups and/or side-chains at the interfaces. The X-ray crystallographic structures of these complexes

Received: October 21, 2015
Published: December 24, 2015

DOI: 10.1021/acs.jcim.5b00638

Table 1. Target–Ligand Complexes Investigated

PDB ID	res. (Å)	target	ligand	buriedness (%)[a]	waters[b]
1B32	1.75	oligopeptide bindig protein A	KMK	70	8
1B3F	1.80	oligopeptide bindig protein A	KHK	70	7
1B46	1.80	oligopeptide bindig protein A	KPK	69	8
1B4Z	1.75	oligopeptide bindig protein A	KDK	69	10
1B51	1.80	oligopeptide bindig protein A	KSK	70	9
1B58	1.80	oligopeptide bindig protein A	KYK	71	7
1B5I	1.90	oligopeptide bindig protein A	KNK	70	8
1B5J	1.80	oligopeptide bindig protein A	KQK	70	10
1B9J	1.80	oligopeptide bindig protein A	KLK	71	8
1BBZ	1.65	abl tyrosine kinase SH3 domain	APSPYPPPP	57	6
1JET	1.20	oligopeptide bindig protein A	KAK	68	8
1JEU	1.25	oligopeptide bindig protein A	KEK	71	9
1JEV	1.30	oligopeptide bindig protein A	KWK	69	6
1JYR	1.55	growth factor receptor-bound protein 2	Ace-S-Ptr[c]-VNVQ-NH2	34	3
1QKA	1.80	oligopeptide bindig protein A	KRK	68	7
1QKB	1.80	oligopeptide bindig protein A	KVK	70	8
1TP5	1.54	presynaptic density protein 95	KKETWV	42	4
2BBA	1.65	ephrin type-B receptor 4	NYLFSPDGPIARAW	49	10
2H2D	1.70	sirtuin	KKGQSTSRHK-Aly[c]-LMFKTEG	32	5
2H2G	1.63	sirtuin	histone H3 tail HA-Aly[c]-RVTIQKKD	47	7
2H2H	2.20	sirtuin	histone H4 tail HA-Aly[c]-TVTSLD	45	2
2O4K	1.60	HIV-1 protease	atazanavir[c]	71	6
2OLB	1.40	oligopeptide bindig protein A	KKK	68	8
2X6M	1.62	camelid antibody fragment	GYQDYEPEA	33	5
3QGJ	1.30	alpha-lytic protease	Ace-AAP-2a1[c]	63	3
3QL9	0.93	transcriptional regulator ATRX	histone H3 tail ARTKQTAR-M3l[c]-STGGKA	43	16
3RO3	1.10	G-protein-signaling modulator 2	QVDSVQRWMEDLKLMTE	45	12
3U43[d]	1.72	colicin-E2 immunity protein	colicin-E2	12	22
4H9N[d]	1.95	death domain-associated protein 6	histone H3.3 wild type/H4	38	49
4H9O[d]	2.05	death domain-associated protein 6	histone H3.3 G90 M mutant/H4	39	35
4H9Q[d]	1.95	death domain-associated protein 6 mutant	histone H3.3/H4	38	38

[a] Buriedness was calculated as $50(SA_{Target} + SA_{Ligand} - SA_{Complex})/SA_{Ligand}$ using surface areas (SA) from PyMol.[39] [b] Counts of crystallographic water positions located in the interface defined by a dmax = 3.5 Å (Methods). [c] Nonamino-acid residues are defined in Table S8. [d] Complexes with protein ligands of more than 100 residues.

contain 344 interfacial water positions, forming the experimental references for this work. The size of the ligands varies from tripeptides up to proteins. Among the targets there are proteins with buried binding pockets, such as HIV-1 protease (PDB code 2O4K) and also other systems with shallow, extended binding interfaces.

The latter group includes complexes of histone proteins (2H2G, 2H2H, 3QL9, 4H9N, 4H9O, 4H9Q) that are important constituents of the nucleosome and key molecules involved in epigenetic regulation.[40−42]

This paper has a two-fold aim: (i) elaboration and validation of a tool for calculation of complete interfacial hydration structures of the above complexes by extending MobyWat[28] a recent MD-approach for the study of biomolecular hydration to interfaces and (ii) exploration of the hydration networks for characterization of high-level cooperation between interfacial water molecules at target–ligand interfaces. Thus, after determining hydration structures at the atomic level, we aim at uncovering static and dynamic hydration networks and their role in stabilizing even large biomolecular interfaces.

■ RESULTS AND DISCUSSION

Complete Hydration of Complex Interfaces. In order to explore the complete hydration networks, our approach provides void-free hydration of the complex interfaces via generation of MD trajectories and their processing with MobyWat, a tool for prediction of hydration structures. In the present study, three methods of increasing complexity (Figure 1) were investigated and validated on the systems of Table 1.

Matches between experimental and calculated water positions were quantified as success rates (SR, eq 1) for all systems and are summarized in Table 2. As there may be various distance criteria to define matches with experimental positions (Table S1), the SR evaluations were performed at two different match tolerances. The SR was determined for the 31 individual systems as well as for all 344 reference water positions (Table 2).

Performance. The simplest Method 1 (M1, Figure 1, Figure S7) works on the free surface of the target without using the structure of the ligand, similarly to other methods (Table S1) and our previous results (Table S2). The ligand is involved in M1 to exclude conflicting water molecules but not for the predictive calculations of the method. Although M1 produced a complete match (SR = 100%) for four systems, there were also 12 systems where SR ≤ 50%. M1 assumes that the target surface is the main determinant of interfacial hydration structure. However, the results are highly system-dependent (Table 2), and therefore, M1 works well for systems with

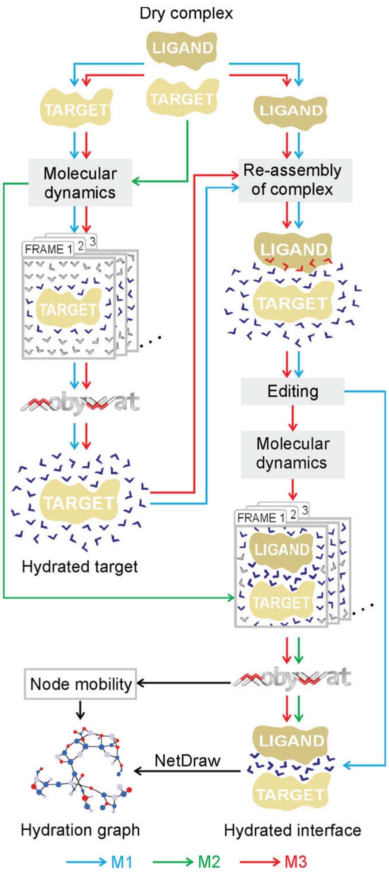


Figure 1. Calculation of interfacial hydration structure and graph. Method 1 (M1) uses only the target structure in molecular dynamics (MD) and MobyWat steps. Methods 2 and 3 (M2 and M3) involve both target and ligand structures in the predictive calculations. In Method 3, additional MD and MobyWat steps were introduced for complete filling up of void volumes of the interface. In the editing step, a minimal ligand−water distance threshold was used to standardize the removal of water molecules conflicting with the ligand (marked with red).

relatively small ligands, where the pocket is deep and the ligand is well buried (Table 1).

As the success of target-based M1 depends on the actual topography of the system, two additional strategies were investigated involving the ligand structure in the prediction. Method 2 (M2, Figure 1, Figure S8) works with the entire target−ligand complex, i.e., water molecules can interact with both faces during MD and MobyWat steps. Similar to M1, a crude hydration procedure is performed at first, where the entire complex is hydrated using pre-equilibrated water positions as provided by the default hydration algorithm of the MD software package (Figure S1). Intermolecular interactions are optimized by a single MD simulation.

Thus, in M2, water−ligand interactions are calculated, as well. In most of the cases, SR values improved considerably compared to M1 (Table 2). However, despite the involvement of the ligand structure and the optimization of interactions, the first crude hydration step of M2 can easily result in void spaces and non-optimal arrangement of the water molecules at the interface due to the restricted access of hidden interface regions to bulk water and their limited translational and rotational freedom (Figure S1).

To overcome such limitations of M2, an additional hydration step was introduced in Method 3 (M3; Figure 1, Figure S9). In this step, the surface of the free target molecule is hydrated without the ligand using short MD and a MobyWat steps as in M1. Thus, an optimal hydration of the cavities can be achieved, as the migration of water molecules is not limited to/from the bulk (Figure S1). Having an optimally loaded target surface, the ligand is positioned back so as to form the interface with the target. This interface can be considered soaked, leaving as many water molecules as possible for the next MD step (Figure 1) to reduce the volume of unwanted void spaces. Finally, additional MD and MobyWat steps are performed to re-equilibrate all interactions of water molecules in the presence of the ligand (Figure 1). The effect of clustering tolerance of the MobyWat step was also investigated (Table S3), and 1.0 Å was found to be the optimal value similarly to the prediction of surface hydration.[28] In terms of SR (Table 2), the overall efficiency of M3 was the best among the three strategies investigated. Although M3 involves additional steps, they are not very

Table 2. Efficiency of Hydration Methods Expressed as Success Rates (SR)

	method[a]/match tolerance (Å)					
	M1		M2		M3	
PDB ID	/1.50	/1.75	/1.50	/1.75	/1.50	/1.75
1B32	50	75	100	100	100	100
1B3F	71	71	71	71	71	71
1B46	38	38	88	88	100	100
1B4Z	60	60	90	90	80	90
1B51	56	67	89	89	89	89
1B58	71	71	100	100	71	100
1B5I	38	50	88	88	88	100
1B5J	60	60	80	80	80	90
1B9J	38	63	75	75	75	88
1BBZ	50	50	83	100	100	100
1JET	50	50	75	75	63	75
1JEU	67	67	78	78	89	100
1JEV	100	100	100	100	100	100
1JYR	50	50	100	100	100	100
1QKA	57	57	86	100	100	100
1QKB	100	100	63	75	75	88
1TP5	50	75	75	100	100	100
2BBA	50	50	80	80	90	90
2H2D	80	80	60	80	100	100
2H2G	71	71	86	100	86	100
2H2H	100	100	100	100	100	100
2O4K	100	100	83	83	100	100
2OLB	50	50	100	100	88	88
2X6M	60	60	100	100	100	100
3QGJ	33	67	100	100	100	100
3QL9	69	69	81	88	94	100
3RO3	50	50	92	100	100	100
3U43	73	82	82	89	95	100
4H9N	57	59	94	96	88	94
4H9O	60	60	94	94	94	100
4H9Q	66	66	79	79	89	92
overall[b]	62	65	86	90	90	95

[a] A value of 100 refers to full match with experimental positions of water oxygen atoms. [b] Overall SR is calculated from total counts of calculated and reference water positions of all systems (eq 1).

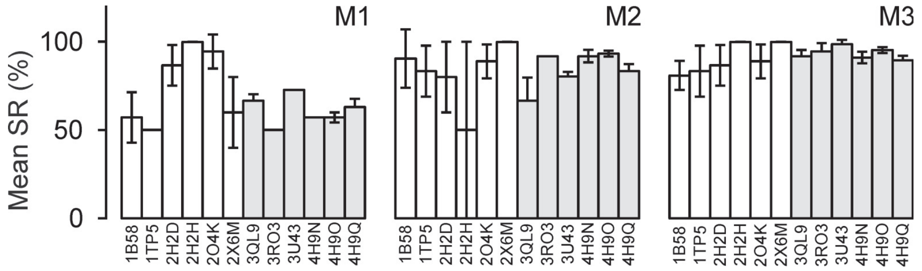


Figure 2. Reproducibility and efficiency of hydration methods. Interface hydration was performed in triplicate for six randomly selected (empty columns) and the six largest (gray columns) systems by all three methods. Efficiencies are plotted as mean success rate (SR) values. Reproducibility is shown as the corresponding standard deviations (error bars). Method 3 (M3) has the best overall efficiency and good reproducibility.

demanding computationally. M3 is therefore recommended if complete (void-free, Figure S1) exploration of the hydration structure of the complex interfaces is needed.

Although current explicit water models cannot reproduce all properties of water,[43] they have great advantages in simulation of protein−water and water−water interactions, peptide and protein folding,[44−48] calculation of hydration,[49] and binding[50] thermodynamics. Selection and appropriate combination of a water model and a protein force field is not trivial for any tasks. For example, advanced four-site models such as TIP4P/2005 have certain advantages[48] in simulating temperature-dependent protein folding with the Amber 03 force field, and the SPC/E three-site model had good performance for calculation of hydration thermodynamics with three different force fields.[49] In the present study, we use the popular[43] TIP3P water model[51] combined with the Amber99SB-ILDN protein force field.[52] This combination performed well in previous studies.[28,52] We also tested the TIP4P-OPLS/AA combination (Supporting Table S7 in ref 28) for prediction of hydration structure and found no significant increase in SR values over the TIP3P model. Furthermore, simulations on hydration structure can benefit from the high mobility of TIP3P allowing increased sampling,[46] which is important for accessing buried interfacial regions. With advances in the precision of biomolecular force fields and water models, we expect that the performance of the methods of Figure 1 will further increase.

Reproducibility and External Comparisons. All three methods of the present study are based on the generation of MD trajectories of water molecules residing in the simulation box. However, MD trajectories produced for the same system may differ substantially from each other.[28] For some reason, this issue is often not considered in MD-based studies on prediction of hydration structure, and tests of reproducibility by evaluation of multiple MD trajectories are often missing. To test the reproducibility of the methods, the predictions were repeated three times for 12 systems (Figure 2) using three different initial velocity distributions.

Six of the 12 systems were picked randomly (empty columns in Figure 2), whereas the other six systems contain the largest number of reference interface water molecules (gray columns in Figure 2). Similar to the results of Table 2, the average SRs of the repeated predictions show that M1 had the lowest overall prediction efficiency (lowest columns in Figure 2) and M2 had problems with reproducibility for certain systems probably due to nonoptimal arrangements as described above (large error bars in Figure 2). M3 proved to be the best method in the reproducibility tests in terms of overall mean and standard deviation of SRs (92 ± 5%).

The performance of M3 (expressed as SR) was compared to the results of two independent, external studies using the same complex interfaces. There are 14 complexes in Table 1 with oligopeptide binding proteins which were used as test cases for a geometrical method AcquaAlta.[53] Details of our results obtained with M3 and AcquaAlta are summarized in Table S4. We found that M3 yielded an overall 10% increase in SR, and for certain systems (1B32, 1QKB), the increase was more than 40%. The improved performance of our M3 method can be attributed to the use of molecular dynamics simulations and the calculation of water networks which is not done in AcquaAlta.[53] In another recent Critical Assessment of Predicted Interactions (CAPRI) project,[29] predictions were performed for the hydration structure of a protein−protein interface (3U43 in Table 1) using various methods. In terms of SR, CAPRI predictions ranged between 52% and 91% (Table S5) using a relatively high match tolerance of 2.0 Å. The best performing method with an SR of 91% was based on MD simulations, whereas others were driven by previous knowledge on water positions or homologies. For the same interface, M3 found all crystallographic reference water positions (SR = 100%, Table S5).

The SR values summarized in Table 2 and the above comparisons with other studies show that the use of faces of both partners of a complex structure is necessary for exhaustive hydration of an interface. Neglecting the ligand structure may result in loss of important ligand−water contacts and incomplete determination of the interfacial hydration structure. The use of an MD-based approach and filling steps of M3 are crucial to eliminate cavities while accounting for water−water interactions and allowing water mobility is essential for complete exploration of the hydration structure.

Characterization of Hydration Networks. Along with the full hydration structure (Table 2, Figure S2, Table S6), M3 also supplies the mobility of each interfacial water molecule using a simple formula (eq 2). The M3-calculated water positions are listed and numbered in an increasing order of mobility scaled uniformly between 0 and 100. This M3 list of interfacial water molecules and their mobility values form the foundation of our new approach for the characterization of hydration networks. The characterization protocol was implemented as the NetDraw mode of program MobyWat (see Methods and Figure S10). NetDraw determines the interactions of interfacial water molecules with each other, and the bulk water and solute (ligand + target) molecules result in network graphs with mobility assigned to all water nodes.

The interfacial hydration networks of two systems, the oligopeptide binding protein A (1QKA), and the transcriptional

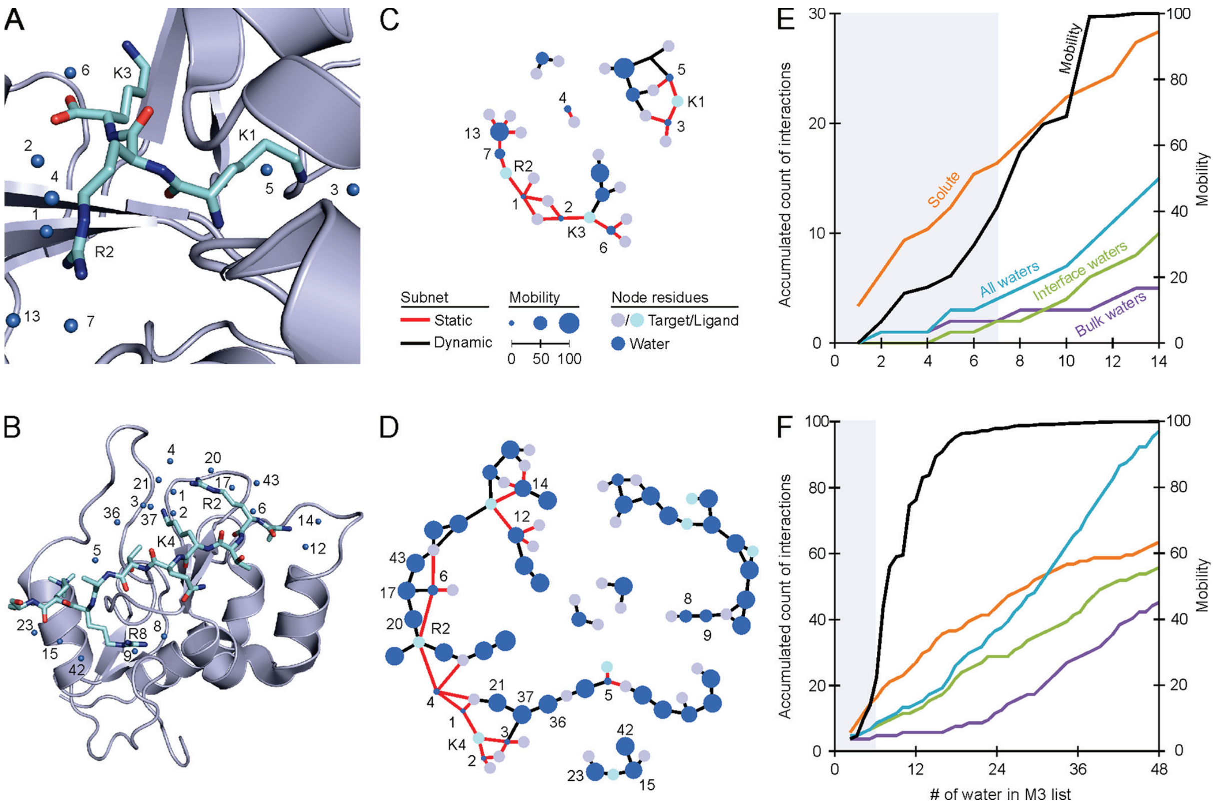


Figure 3. Characterization of hydration networks of protein−peptide complex interfaces of systems 1QKA (A, C, E) and 3QL9 (B, D, F). Positions of water molecules (A, B) were calculated by M3 and used for generation of hydration network graphs (C, D) of the interfaces. Static subnets of red edges are represented in both systems (C, D). Accumulated count of interactions of interface water molecules (E, F) with solute (protein + peptide) molecules, each other, bulk waters, and all waters (interface + bulk) are shown as functions of serial numbers of interface water molecules as listed by M3. The corresponding mobility values are also plotted as a separate curve. In the low (≤50) mobility region, the curves have gray background. Numbering of water positions (nodes) follows the numbering of M3 lists throughout the figure. For clarity, labels were attached only to water molecules and residues discussed in the text. See Figure S2 and Table S6 for distances of M3 water positions (A, B) to matching references.

regulator ATRX (3QL9) with radically different topographies and ligands are shown in Figure 3C and D. System 1QKA has a tripeptide ligand deeply buried (Figure 3A, Table 1) in the target protein, with a compact interface involving a small number of waters and limited communication with the bulk. The relatively large ligand of system 3QL9 is an N-terminal peptide tail of a histone H3 protein trimethylated on Lys9. Coding of epigenetic regulation is attributed[40,42] to such methylation and other post-translational modifications of Lys and Arg side-chains in H3. The histone peptide tail in 3QL9 binds at a shallow surface of the target transcriptional regulator protein (Figure 3B) leaving the Arg and Lys side-chains open to interactions with water molecules of the bulk.

Static Subnets. In the hydration networks of both systems, subnets of low (≤50) mobility water nodes can be observed, where 50 is the approximate inflection point of the mobility curve of 3QL9 (Figure 3F). The M3-listed top 7 of 14 (50%) and 6 of 48 (12.5%) water positions belong to this category for systems 1QKA and 3QL9, respectively. Such low mobility water molecules usually have more than two different contacts with the solute (red edges in Figure 3 C, D), and they are essential parts of its structure.[54] In extreme cases of water #1 in 1QKA (Figure 3C) and water #4 in 3QL9 (Figure 3D), all four possible hydrogen bonds are formed mostly with solute partners. The network graph (Figure 3D) is very densely connected around the low mobility nodes compared to other regions of the network. At the corresponding low mobility domains (gray background in Figures 3 E, F) of accumulative curves of interaction counts, the dominancy of interactions with solute faces can be observed. Here, the curves representing interactions of interface water molecules with the solute run above the other curves accounting for interactions with water.

The above findings suggested the introduction of an upper mobility limit of 50 as an identification criterion of static nodes in the characterization protocol. On the basis of such simple criteria (Methods), the small interfacial hydration network of system 1QKA was characterized as almost completely static (Figure 3C), whereas in 3QL9, the static edges were gathered in a well-defined core (Figure 3D). In both cases, static subnets are centered at charged ligand side-chains of the coding Arg and Lys residues (waters #1−4, Figure 3 A, B) or strong H-bonding backbone amide group (water #5, Figure 3B). Thus, the number of contacts with bulk waters is marginal in static subnets (bulk curve in the gray region of Figure 3 E, F). Such H-bonding networks of buried regions isolated from bulk water are of central importance of kinetic stability of complexes due to shielding of target-ligand H-bonds by the solute.[30] Besides target−ligand H-bonds, the above static water−water subnets are also often shielded from the bulk, and therefore, their presence and detection may be important for ligand design. In

DOI: 10.1021/acs.jcim.5b00638

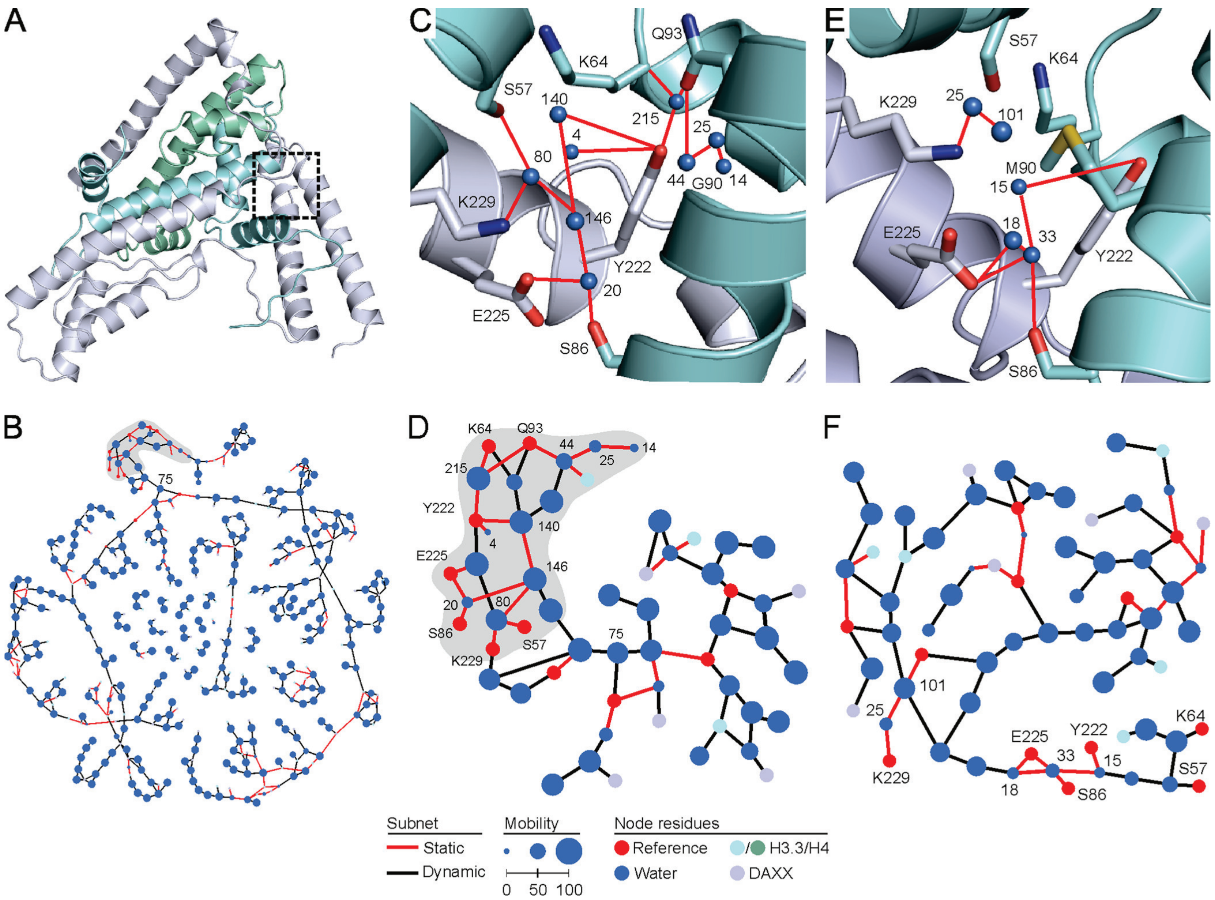


Figure 4. Stability of a histone−chaperon complex requires the presence of a continuous, static hydration subnet in the interface. (A) Ternary complex of DAXX-H3.3-H4 proteins is stabilized by hydration network connecting key residues[41] of the interface region (box) at the mutated G90 residue of the three proteins. Positions of interfacial water molecules were calculated by M3 and used for generation of the complete hydration network of the wild type complex 4H9N (B) and the G90 M mutant 4H9O (Figure S6). Interfacial hydration graphs at G90 were separated and further characterized, and a continuous static subnet (marked with gray background in B, D) could be identified in 4H9N (C,D). It was demolished in 4H9O (E, F).

the case of system 3QL9, the static hydration subnet (Figure 3D) anchors the N-terminal part of the histone peptide tail to the protein target (Figure 3B). The rest of the 3QL9 hydration network contains water molecules of increased mobility and is discussed in the next section.

Dynamic Subnets. Whereas static subnets bridge between the solute faces, the role of dynamic segments of the interfacial hydration structure is fairly complex. The accumulated interaction plot of system 3QL9 (Figure 3F) shows that the static region of low mobility waters (gray background) is followed by a steep ascend for interactions with (all) water molecules and a saturation phase for the count of solute contacts. Indeed, the corresponding hydration subnets (black edges in Figure 3D) of the 3QL9 interface contain water molecules of high mobility (>50) hydrogen-bonding to each other and an increased count of contacts to the bulk. The high rate of exchange with each other and the bulk is an important feature of the water molecules of these subnets, and therefore, they can be considered as “dynamic”. Such dynamic networks with changeable nodes and edges often occur in real life problems, whereas traditional network analyses work on static networks.[55] Thus, a distinction between dynamic and static networks is common[55,56] in network science, and this classification was adopted for the hydration subnets in the present study. Due to a small number of contacts to solute partners and increased topological distances, it is shown (Figure 3D) that the density of edges in dynamic subnets is small compared to the static ones. Whereas the static hydration subnet anchors the histone peptide at its N-terminus to the target, dynamic water nodes #8 and #9 stabilize the binding of the mobile C-terminus (Figure 3B, D). Sometimes dynamic nodes form separate graphs (waters #15, #23, and #42, Figure 3D). Water molecules of the highest mobility values can be considered as “bulk-like” nodes starting from about #30, where the curve of interaction counts with “all waters” exceeds that with solutes (Figure 3F) due to a sharp increase in bulk contacts. The network graph of 3QL9 (Figure 3D) also shows that subnets of dynamic (bulk-like) water molecules can affect topologically distant static ones. For example, waters #21 and #36 affect #3 via #37, waters #20 and #43 affect #6 via #17, and so on. Such high-level connections of the dynamic hydration subnet can provide an “access channel” from the static regions toward the bulk, which is an important factor of (de)stabilizing of the complexes[30,57] via, e.g., destroying the above-mentioned protecting shields[30] above the H-bonds of the buried, static regions.

Beyond the Hydration Structure. Static and moderately dynamic interface water positions were found by the M3 method and X-ray crystallography (Figure S2) equally well. M3 detects most of these positions with low mobility as top candidates (see Figure S3 for a performance analysis and validation details). Dynamic positions with high mobility (bulk-like behavior) were identified only by M3, and those are missing from the crystallographic interface structure. However, the existence of the corresponding continuous void spaces in the interface volume is improbable (Figure S4). Therefore, the void-free hydration of the interface is an important feature of M3. Determination of the full hydration structures with all atoms and contacts led to exploration of the complete interfacial hydration networks including the dynamic regions. Using the network graphs equipped with mobility values (Figure 3 C, D) of water nodes allowed the above characterization of hydration networks by their decomposition into static and dynamic subnets. A further example on the use of subnet characterization for system 3U43 is shown in Figure S13.

Case Study of a Histone−Chaperone Interface. Like the histone H3 involved in system 3QL9 (Figure 3), H4 is an important constituent of the nucleosome. The hydration network of a DAXX-H3.3-H4 ternary complex (4H9N, Figure 4A) was investigated here. DAXX (also known as death-associated protein 6) is a chaperone protein, involved in the pathophysiology of several tumorous diseases.[58−60] It was shown[41] that the integrity of the interfacial hydration structure (Figure 4C) between reference residues Y222, E225, and K229 of DAXX and G90 and K64 of H3.3 is crucial for the stability of the wild type ternary complex 4H9N. A single mutant G90 M of H3.3 (4H9O) resulted in an approximate 50% reduction in its binding to DAXX.[41] It was suggested that the replacement of water molecules by the Met side-chain alone does not explain reduced binding, and it is rather a deintegration of the entire interfacial hydrogen bond network, which is responsible for the effect.[41]

For investigation of the deintegration of the interfacial hydration network, void-free hydration structures of the wild type (4H9N) and mutant (4H9O) ternary complexes were determined by M3. This was particularly challenging, as the protein partners are enveloped into each other, and large, nonburied (Table 1) interfaces are formed between their interacting chains (Figure 4A). The M3-calculated hydration structures of both 4H9N and 4H9O were validated using the available crystallographic water positions, and SRs of about 90% were achieved (Table 2, Figure S2). Similar to the systems of Figure 3, the complete interfacial hydration networks were produced from the hydration structures (Figure 4B, Figures S5 and S6).

The above-mentioned key protein residues[41] were used as references for separation of graphs (Figure 4D, F) relevant for network analysis of the interface region at the mutated G90 residue (Figure 4A, C, E, Figures S5 and S6, Methods, Figure S12). In the separate graph of 4H9N (Figure 4D), central water node #75 can be observed connecting three branches. The branch containing seven reference nodes on the left side of Figure 4D is a continuous static subnet marked with gray background and red lines. This static subnet is a compact core within the entire network (Figure 4B) and has high density of connections (red edges) similar to the static subnet of system 3QL9 described in Figure 3D. Indeed, the vast majority of the edges (81%) in this branch link to protein nodes, whereas this ratio is less than a half (40%) in the entire graph (Table S7). A total of 48% of the protein nodes of the entire graph are linked in the static branch. In the case of 4H9O, this value is decreased to 26% (Table S7) reflecting a remarkable reduction of the static subnet. The differences between 4H9N and 4H9O are also striking if comparing the morphology of the corresponding hydration graphs. In the graph of 4H9O (Figure 4F), the continuous static branch was demolished, and the distribution of the reference protein nodes and static edges became diffuse if compared to 4H9N (Figure 4D). This difference is also shown between Figure 4C and E, which show the spatial plot of the static subnet and the surrounding protein parts in the cases of 4H9N and 4H9O, respectively. In addition, certain protein nodes (K229, S57) have changed their positions in the 4H9O graph (Figure 4F) or disappeared like Q93.

The above exploratory work revealed the presence of a continuous static core in the interfacial hydration network of the stable, wild type ternary protein complex (4H9N). The diffusion of the static core and rearrangement of its links resulted in a dynamic, deintegrated hydration network leading to the reduced binding[41] of the protein partners in the mutant system 4H9O.

CONCLUSIONS

Hydration structures of complex interfaces were explored using a new tool M3 based on all-atom MD calculations and accounting for explicit water−water contacts and mobility. M3 provided hydration structures of interfaces between proteins and peptides with high accuracy and reproducibility allowing construction of their complete network graphs.

On the basis of the graphs and node mobility values from M3, an approach was introduced for characterization of interfacial hydration networks via their decomposition into static and dynamic subnets. It was found that static subnets consisting of nodes of low (≤50) mobility appear at buried sites and around highly charged ligand moieties. Static subnets usually have a high density of edges, and they can dominate small hydration networks. They can also form the cores of interfacial hydration networks of large protein−protein complexes and are often essential for strongly linking of the partners. For extended interfaces with several interacting water molecules, the presence of dynamic subnets becomes considerable. Dynamic subnets connect water and solute nodes at even large topological distances and provide access channels between the static and bulk regions.

Our results on exploration and characterization of interfacial hydration networks are implemented in a standalone, open source software tool, which can be used for prediction of structure and stability of biomolecular complexes and engineering of new ligands (Figure S15).

METHODS

Preparation of systems. The protein databank (PDB) structure of the target (M1 and M3) or the target−ligand complex (M2) was used as primary input of the calculations. All crystallographic waters (M1, M2, and M3) and the ligand molecule (M1 and M3) were removed. Missing atoms of solute side-chains (both protein and ligand) were reconstructed with Swiss PDB Viewer.[61] Solute amino acids absent in the crystallographic structure were not remodeled. The structure was placed in a dodecahedral box using a distance criterion of 1 nm between the solute and the box. Void spaces of the box were filled by explicit TIP3P water molecules[51] with the standard gmx solvate routine (Figure S1) of GROMACS.[62,63] In the case of the M3 complex, input geometry of water molecules of the interface region were obtained as described in Figure 1 and in the Calculation of Interfacial Hydration Structure section. Counterions (sodium or chloride) were added to neutralize the system.

Energy Minimization. A uniform procedure was applied for molecular mechanics energy minimization in all cases prior to the MD steps. In the first step, a steepest descent (sd) optimization was carried out, with convergence threshold set to 10^3 kJ mol^{-1} nm^{-1}. This was followed by a conjugate gradient (cg) calculation, where the convergence threshold was set 10 kJ mol^{-1} nm^{-1}. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ mol^{-1} nm^{-2} in both steps. Distance restraints were applied between structural ions and coordinating amino acid residues at a force constant of 10^3 kJ mol^{-1} nm^{-2} in the cases of systems 2H2D, 2H2H, 2H2G, 3QL9, and 3U43. All calculations were performed with programs of the GROMACS software package,[63] using the AMBER99SB-ILDN force field.[52] During the M3 protocol, the above energy-minimization was performed twice (Figure 1), that is, once for the target and once for the reassembled target–ligand complex. Before network analyses, a four-step protocol was applied for energy minimization of predicted water positions following an sd-cg-sd-cg pattern with parameters of sd and cg methods described above. During the first two steps, all solute heavy atoms and the oxygen of the predicted interfacial water molecules were position restrained, and bulk waters and ions were released. In the last two steps, position restraints were not applied on predicted waters, and only solute heavy atoms were position restrained.

Parameters of Nonstandard Residues. For nonstandard (nonamino-acid) residues of atazanavir (2O4K), phosphotyrosine (1JYR), acetyl-lysine (2H2D, 2H2G, 2H2H), 2S-2-aminopropan-1-ol (3QGJ), and trimethylated lysine (3QL9), molecular mechanics force field parameters were obtained from the GAFF force field.[64] The Lewis structure of these residues can be found in Table S8. The nonstandard residues (except atazanavir) were first capped on both terminals, with acetyl and *N*-methyl groups and preminimized with PC Model 9[65] using MMFF94 force field.[66] Subsequently, semi-empirical quantum mechanics optimization was performed with MOPAC-2009[67] using the PM6 parametrization.[68] Then, the completely minimized molecule was uploaded to RED server[69] to perform *ab initio* geometry optimization to obtain partial charges by RESP-A1B charge fitting (compatible with the AMBER99SB-ILDN force field). The calculations were performed with the Gaussian09 software,[70] using the HF/6-31G* split valence basis set.[71] The caps on the termini were excluded from charge derivation, and charge restraints were applied on these atoms. Normal mode analysis was performed using GAMESS[72] to ensure that the final geometry is in energy minimum. Bond stretching, angle bending, and torsional parameters were assigned with the parmchk utility of AmberTools 1.5 program package[73] and used together with the partial charges to build GROMACS residue topology entries for the nonstandard residues.

Molecular Dynamics. After energy minimization, 1 ns-long NPT MD simulations were carried out with a time step of 2 fs. For temperature-coupling, the velocity rescale algorithm[74] was used. Solute and solvent were coupled separately with a reference temperature of 300 K and a coupling time constant of 0.1 ps. Pressure was coupled the Parrinello–Rahman algorithm[75−77] with a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1}, and reference pressure of 1 bar. Particle Mesh–Ewald summation was used for long-range electrostatics. van der Waals and Coulomb interactions had a cutoff at 11 Å. Coordinates were saved at regular time intervals of 1 ps yielding 1.001×10^3 frames. Position and distance restraints were applied as described in the Energy-Minimization section. After MD, all frames were extracted from the trajectory. Periodic boundary conditions were treated before analysis to make the solute whole and recover hydrated solute structures centered in the box. Each frame was fit to the original protein crystal structure using C_α atoms. A detailed description of this procedure can be found in the Supporting Information. The final trajectory including all atomic coordinates of all frames was saved as portable binary files and was used for subsequent calculation of hydration structures. During the M3 protocol, the above MD simulation was performed twice (Figure 1), that is, once for the target and once for the reassembled target–ligand complex. During reproducibility tests (Figure 2), MD simulations were performed in triplicate using three different sets of initial velocities. In the case of M3, triplicate MD simulations were also performed both for the target and for the reassembled target–ligand complex.

Calculation of Interfacial Hydration Structure. To select water molecules residing in the target–ligand interface, distances between its oxygen atom and the closest heavy atom of the solute molecules were measured. A water molecule was considered interfacial if such a distance was smaller than or equal to a predefined maximal distance limit (dmax) for both the ligand and target partners. The program MobyWat[28] was updated for the present study to handle interfaces and do editing at the soaking step. The new version 1.1 of MobyWat was used for prediction of position of water molecules from MD trajectories. The source code, executable binary, and user's manual of MobyWat 1.1 will be made freely available at the program's Web site www.mobywat.com at the time of publication. In the case of predictions of positions of interface water molecules (M2 and M3), a dmax of 3.5 Å an identity-based (IDa) clustering were applied with clustering and prediction tolerances of 1.0 and 2.5 Å, respectively. In the cases of M1 and M3, predictions were also performed for the entire surface of the target with dmax, clustering, and prediction tolerances of 5.0, 1.5, and 2.5 Å, respectively, where dmax refers to the distance from the target only. During M3, the ligand–target complex was reassembled after hydration of the target surface (Figure 1). For this, the target part of the holo and the hydrated apo systems were fitted on the top of each other, and the ligand was used together with the hydrated target (soaking). Water molecules conflicting with the ligand structure were excluded using the editing mode of MobyWat at a minimum distance limit (dmin) of 1.75 Å prior to the second MD simulation. The quality of predictions was checked using the validation submode of MobyWat and expressed as success rates according to eq 1.

$$\mathrm{SR} = 100\frac{\text{Count of matches}}{\text{Count of reference water positions}}\% \qquad (1)$$

Crystallographic positions of water molecules within a dmax of 3.5 Å measured from both target and ligand were used as references irrespective of their B-factors. Matches were identified if the distance between predicted and reference water oxygen atoms was below a predefined tolerance of 1.5 Å (match tolerance). For further information on the algorithms of MobyWat and a collection of the result files of the present study, please, refer to the user's manual, Figures S7−S9, and results files in the Supporting Information. Descriptions of success and failure cases are provided in Figure S14.

Network Analyses. The NetDraw mode (Figure S10) of MobyWat ver. 1.1. was written for the present study and was used for analysis and characterization of the hydration network of a target−ligand interfaces. Using Protein Databank (PDB) files including energy-minimized structures including solute, predicted interface molecules, and bulk water molecules as an input, NetDraw produces a two-dimensional interaction network graph of the interface as lists of edges and nodes. Here, a water molecule or a residue of the solute was considered as a node. NetDraw detected and listed atomic pairs of the partner groups (ligand, target, interfacial water, bulk water) with heavy atom distances up to maximal distance limit of 3.0 Å. The lists are stored in distance files. The list of edges of the graph is distilled from the distance files by eliminating redundancies and distances (edges) to carbon atoms (C-filtering). The number of edges per node is limited to four, using the top four shortest edges only (four-filtering). The list of nodes was produced by simple book-keeping from the list of edges. All lists are produced with and without considering bulk water nodes. Finally, NetDraw produces a classification of the nodes and edges. A node is classified static if it is a solute (ligand/target) node or connected to four nodes of any type or is connected to at least three solute nodes or has a mobility value smaller than or equal to 50. Otherwise, the node is classified as dynamic. An edge is defined static if it connects two static nodes; otherwise, it is dynamic. The mobility values were produced by MobyWat for every (i^{th}) predicted water position during the prediction steps and can be transferred to the network analysis steps as B-factors in the PDB files. Mobility values were calculated from the corresponding occupancy (O_i) values by eq 2, where O_i is the occurrence of a water molecule with the same ID in the frames of a trajectory divided by the number of frames (1.001×10^3) in our case. Mobility scales between 0 and 100 where zero corresponds to the least mobile predicted interface water position. The predicted positions are listed in an increasing order of mobility.

$$\text{Mobility}_i = 100\frac{O_{\max} - O_i}{O_{\max} - O_{\min}},\ O_{\max/\min} = \max_i/\min_i O_i \qquad (2)$$

With the above classification, NetDraw helps distinguish between static and dynamic subnetworks of the entire interfacial network graph, which can be used for characterization of integrity of the interface and prediction of complex stability as described in Results and Discussion (Figure S12). MobyWat provides a hydration network graph with node mobility and dynamicity information for the nodes and edges in file formats commonly used by network visualization and analysis programs. In the present study, Gephi,[78] a graph analyzing tool, was used for visualization of the graphs produced by NetDraw. Layout was created with the ForceAtlas layout option, with "Attraction distribution". Node sizes were scaled by size with mobility information. In the case of systems 4H9N and 4H9O, a core graph of the interfacial hydration network surrounding residue G90[41] was separated for further analyses (Figure 4). The separation of the graph was uniformly done for the two systems by cutting edges beyond the topological distance of the three edges measured from the reference nodes, key residues of the interface (Table S7).[41] In the case of Figure 3, interactions (edges) between interface water and other nodes were accumulated pair-wisely for all types of nodes and plotted as a function of increasing list serial number (increasing mobility) of the predicted interface waters. Original files of network analyses are provided as a results file in the Supporting Information.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00638.

Figures S1−S15 and Tables S1−S8. (PDF)

Output files including results of calculation of hydration structures of all systems for Method 1. (ZIP)

Output files including results of calculation of hydration structures of all systems for Method 2. (ZIP)

Output files including results of calculation of hydration structures of all systems for Method 3. (ZIP)

Network analyses. (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: csabahete@yahoo.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge a grant of computer time from the CSCS Swiss National Supercomputing Centre and the NIIF Hungarian National Information Infrastructure Development Institute. Part of the simulations were carried out on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the 'Abisko' supercomputer of the High Performance Computing Center North (HPC2N, Sweden, grant SNIC2013-26-6). The work was supported by the Hungarian Scientific Research Fund (OTKA K112807). We are thankful to the Gedeon Richter Pharmaceutical Plc. for a predoctoral scholarship to N.J.

■ REFERENCES

(1) Zhou, X. Z.; Menche, J.; Barabasi, A. L.; Sharma, A. Human symptoms-disease network. *Nat. Commun.* **2014**, *5*, 4212.

(2) Barabasi, A. L. Network science luck or reason. *Nature* **2012**, *489*, 507−508.

(3) Bullmore, E. T.; Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **2009**, *10*, 186−198.

(4) Majdandzic, A.; Podobnik, B.; Buldyrev, S. V.; Kenett, D. Y.; Havlin, S.; Stanley, H. E. Spontaneous recovery in dynamical networks. *Nat. Phys.* **2014**, *10*, 34−38.

(5) Barzel, B.; Barabasi, A. L. Universality in network dynamics. *Nat. Phys.* **2013**, *9*, 673−681.

(6) Ghoshal, G.; Barabasi, A. L. Ranking stability and super-stable nodes in complex networks. *Nat. Commun.* **2011**, *2*, 394.

(7) Barik, A.; Bahadur, R. P. Hydration of protein-RNA recognition sites. *Nucleic Acids Res.* **2014**, *42*, 10148−10160.

(8) Ahmad, M.; Gu, W.; Geyer, T.; Helms, V. Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun.* **2011**, *2*, 261.

(9) Rosenbaum, D. M.; Rasmussen, S. G. F.; Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **2009**, *459*, 356−363.

(10) Jayaram, B.; Jain, T. The role of water in protein-DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 343−361.

(11) Schwabe, J. W. R. The role of water in protein-DNA interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 126−134.

(12) Garcia-Sosa, A. T.; Mancera, R. L. Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex. *Mol. Inf.* **2010**, *29*, 589−600.

(13) Roberts, B. C.; Mancera, R. L. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.* **2008**, *48*, 397−408.

(14) Mancera, R. L. Molecular modeling of hydration in drug design. *Curr. Opin. Drug Discovery* **2007**, *10*, 275−280.

(15) Garcia-Sosa, A. T.; Mancera, R. L. The effect of a tightly bound water molecule on scaffold diversity in the computer-aided de novo ligand design of CDK2 inhibitors. *J. Mol. Model.* **2006**, *12*, 422−431.

(16) Garcia-Sosa, A. T.; Firth-Clark, S.; Mancera, R. L. Including tightly-bound water molecules in de novo drug design. Exemplification through the in silico generation of poly (ADP-ribose)polymerase ligands. *J. Chem. Inf. Model.* **2005**, *45*, 624−633.

(17) Garcia-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.* **2003**, *9*, 172−182.

(18) Lloyd, D. G.; Garcia-Sosa, A. T.; Alberts, I. L.; Todorov, N. P.; Mancera, R. L. The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 89−100.

(19) GrandPre, T.; Li, S. X.; Strittmatter, S. M. Nogo-66 receptor antagonist peptide promotes axonal regeneration. *Nature* **2002**, *417*, 547−551.

(20) Thirumalai, D.; Reddy, G.; Straub, J. E. Role of water in protein aggregation and amyloid polymorphism. *Acc. Chem. Res.* **2012**, *45*, 83−92.

(21) Pommier, Y.; Marchand, C. Interfacial inhibitors: targeting macromolecular complexes. *Nat. Rev. Drug Discovery* **2011**, *11*, 25−36.

(22) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450*, 1001−1009.

(23) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. The solvent component of macromolecular crystals. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71*, 1023−1038.

(24) Afonine, P. V.; Grosse-Kunstleve, R. W.; Adams, P. D.; Urzhumtsev, A. Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *69*, 625−634.

(25) Badger, J. Modeling and refinement of water molecules and disordered solvent. *Methods Enzymol.* **1997**, *277*, 344−352.

(26) Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **1996**, *3*, 973−980.

(27) Halle, B. Biomolecular cryocrystallography: Structural changes during flash-cooling. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 4793−4798.

(28) Jeszenoi, N.; Horvath, I.; Balint, M.; van der Spoel, D.; Hetenyi, C. Mobility-based prediction of hydration structures of protein surfaces. *Bioinformatics* **2015**, *31*, 1959−1965.

(29) Lensink, M. F.; Moal, I. H.; Bates, P. A.; Kastritis, P. L.; Melquiond, A. S. J.; Karaca, E.; Schmitz, C.; van Dijk, M.; Bonvin, A. M. J. J.; Eisenstein, M.; Jimenez-Garcia, B.; Grosdidier, S.; Solernou, A.; Perez-Cano, L.; Pallara, C.; Fernandez-Recio, J.; Xu, J. Q.; Muthu, P.; Kilambi, K. P.; Gray, J. J.; Grudinin, S.; Derevyanko, G.; Mitchell, J. C.; Wieting, J.; Kanamori, E.; Tsuchiya, Y.; Murakami, Y.; Sarmiento, J.; Standley, D. M.; Shirota, M.; Kinoshita, K.; Nakamura, H.; Chavent, M.; Ritchie, D. W.; Park, H.; Ko, J.; Lee, H.; Seok, C.; Shen, Y.; Kozakov, D.; Vajda, S.; Kundrotas, P. J.; Vakser, I. A.; Pierce, B. G.; Hwang, H.; Vreven, T.; Weng, Z. P.; Buch, I.; Farkash, E.; Wolfson, H. J.; Zacharias, M.; Qin, S. B.; Zhou, H. X.; Huang, S. Y.; Zou, X. Q.; Wojdyla, J. A.; Kleanthous, C.; Wodak, S. J. Blind prediction of interfacial water positions in CAPRI. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 620−632.

(30) Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J. Am. Chem. Soc.* **2011**, *133*, 18903−18910.

(31) Halle, B. Protein hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc., B* **2004**, *359*, 1207−1223.

(32) Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M. Water networks contribute to enthalpy/entropy compensation in protein-ligand binding. *J. Am. Chem. Soc.* **2013**, *135*, 15579−15584.

(33) Genheden, S.; Mikulskis, P.; Hu, L.; Kongsted, J.; Soderhjelm, P.; Ryde, U. Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. *J. Am. Chem. Soc.* **2011**, *133*, 13081−13092.

(34) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.

(35) Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102*, 3531−3541.

(36) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of graph-theoretical parameters in quantifying molecular similarity and structure-activity-relationships. *J. Chem. Inf. Model.* **1994**, *34*, 270−276.

(37) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting properties of molecules using graph invariants. *J. Math. Chem.* **1991**, *7*, 243−272.

(38) Rahat, O.; Alon, U.; Levy, Y.; Schreiber, G. Understanding hydrogen-bond patterns in proteins using network motifs. *Bioinformatics* **2009**, *25*, 2921−2928.

(39) Schrodinger, L. *The PyMOL Molecular Graphics System*, version: 1.7.4. http://www.pymol.org/ (accessed December 2015).

(40) Musselman, C. A.; Lalonde, M. E.; Cote, J.; Kutateladze, T. G. Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1218−1227.

(41) Elsasser, S. J.; Huang, H. D.; Lewis, P. W.; Chin, J. W.; Allis, C. D.; Patel, D. J. DAXX envelops a histone H3.3-H4 dimer for H3.3-specific recognition. *Nature* **2012**, *491*, 560−565.

(42) Jenuwein, T.; Allis, C. D. Translating the histone code. *Science* **2001**, *293*, 1074−1080.

(43) Vega, C.; Abascal, J. L. F. Simulating water with rigid non-polarizable models: a general perspective. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19663−19688.

(44) Nayar, D.; Chakravarty, C. Sensitivity of local hydration behaviour and conformational preferences of peptides to choice of water model. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10199−10213.

(45) Nerenberg, P. S.; Head-Gordon, T. Optimizing protein-solvent force fields to reproduce intrinsic conformational preferences of model peptides. *J. Chem. Theory Comput.* **2011**, *7*, 1220−1230.

(46) Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M. Explicit water models affect the specific solvation and dynamics of unfolded peptides while the conformational behavior and flexibility of folded peptides remain intact. *J. Chem. Theory Comput.* **2010**, *6*, 3569−3579.

(47) Matthes, D.; de Groot, B. L. Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes. *Biophys. J.* **2009**, *97*, 599−608.

(48) Best, R. B.; Mittal, J. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B* **2010**, *114*, 14916−14923.

(49) Hess, B.; van der Vegt, N. F. A. Hydration thermodynamic properties of amino acid analogues: A systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **2006**, *110*, 17616−17626.

(50) Fadda, E.; Woods, R. J. On the role of water models in quantifying the binding free energy of highly conserved water molecules in proteins: the case of Concanavalin A. *J. Chem. Theory Comput.* **2011**, *7*, 3391−3398.

(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(52) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.

(53) Rossato, G.; Ernst, B.; Vedani, A.; Smiesko, M. AcquaAlta: A directional approach to the solvation of ligand-protein complexes. *J. Chem. Inf. Model.* **2011**, *51*, 1867−1881.

(54) Petsko, G. A.; Ringe, D. *Protein Structure and Function*; Oxford University Press: Oxford, U.K., 2008.

(55) Kim, H.; Anderson, R. Temporal node centrality in complex networks. *Phys. Rev. E* **2012**, *85*, 026107.

(56) Hulovatyy, Y.; Chen, H.; Milenkovic, T. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics* **2015**, *31*, 171−180.

(57) Hyre, D. E.; Amon, L. M.; Penzotti, J. E.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P.; Stayton, P. S. Early mechanistic events in biotin dissociation from streptavidin. *Nat. Struct. Biol.* **2002**, *9*, 582−585.

(58) Schwartzentruber, J.; Korshunov, A.; Liu, X. Y.; Jones, D. T.; Pfaff, E.; Jacob, K.; Sturm, D.; Fontebasso, A. M.; Quang, D. A.; Tonjes, M.; Hovestadt, V.; Albrecht, S.; Kool, M.; Nantel, A.; Konermann, C.; Lindroth, A.; Jager, N.; Rausch, T.; Ryzhova, M.; Korbel, J. O.; Hielscher, T.; Hauser, P.; Garami, M.; Klekner, A.; Bognar, L.; Ebinger, M.; Schuhmann, M. U.; Scheurlen, W.; Pekrun, A.; Fruhwald, M. C.; Roggendorf, W.; Kramm, C.; Durken, M.; Atkinson, J.; Lepage, P.; Montpetit, A.; Zakrzewska, M.; Zakrzewski, K.; Liberski, P. P.; Dong, Z.; Siegel, P.; Kulozik, A. E.; Zapatka, M.; Guha, A.; Malkin, D.; Felsberg, J.; Reifenberger, G.; von Deimling, A.; Ichimura, K.; Collins, V. P.; Witt, H.; Milde, T.; Witt, O.; Zhang, C.; Castelo-Branco, P.; Lichter, P.; Faury, D.; Tabori, U.; Plass, C.; Majewski, J.; Pfister, S. M.; Jabado, N. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **2012**, *482*, 226−231.

(59) Jiao, Y.; Shi, C.; Edil, B. H.; de Wilde, R. F.; Klimstra, D. S.; Maitra, A.; Schulick, R. D.; Tang, L. H.; Wolfgang, C. L.; Choti, M. A.; Velculescu, V. E.; Diaz, L. A., Jr.; Vogelstein, B.; Kinzler, K. W.; Hruban, R. H.; Papadopoulos, N. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **2011**, *331*, 1199−1203.

(60) Heaphy, C. M.; de Wilde, R. F.; Jiao, Y.; Klein, A. P.; Edil, B. H.; Shi, C.; Bettegowda, C.; Rodriguez, F. J.; Eberhart, C. G.; Hebbar, S.; Offerhaus, G. J.; McLendon, R.; Rasheed, B. A.; He, Y.; Yan, H.; Bigner, D. D.; Oba-Shinjo, S. M.; Marie, S. K.; Riggins, G. J.; Kinzler, K. W.; Vogelstein, B.; Hruban, R. H.; Maitra, A.; Papadopoulos, N.; Meeker, A. K. Altered telomeres in tumors with ATRX and DAXX mutations. *Science* **2011**, *333*, 425.

(61) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714−2723.

(62) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to super-computers. *SoftwareX* **2015**, *1*, 19−25.

(63) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(64) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(65) Gille, A. L.; Dutmer, B. C.; Gilbert, T. M. PCMODEL 9.2. *J. Am. Chem. Soc.* **2009**, *131*, 5714−5714.

(66) Halgren, T. A. Merck molecular force field 0.1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(67) Stewart, J. J. P. *MOPAC2009*, version: 2009; Steward Computational Chemistry; Colorado Springs, CO, U.S.A. http://openmopac.net/ (accessed December 2015).

(68) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(69) Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **2011**, *39*, W511−517.

(70) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, version: Gaussian, Inc.: Wallingford, CT, U.S.A. http://www.gaussian.com/g_prod/g09.htm (accessed December 2015).

(71) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular-orbital methods 0.20. Basis set for correlated wave-functions. *J. Chem. Phys.* **1980**, *72*, 650−654.

(72) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic-structure system. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(73) Case, D.; Darden, T.; Cheatham, T., III; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Walker, R.; Zhang, W.; Merz, K. *AmberTools*, version: 15. http://ambermd.org/ (accessed December 2015).

(74) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(75) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(76) Nose, S.; Klein, M. L. Constant pressure molecular-dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055−1076.

(77) Parrinello, M.; Rahman, A. Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(78) Bastian, M.; Heymann, S.; Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*; San Jose, CA, U.S.A., 2009.

D18

hetenyi.csaba_83_23

Journal of Molecular Graphics and Modelling 82 (2018) 117–128

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Analysis of the influence of simulation parameters on biomolecule-linked water networks

Norbert Jeszenői [a], Gabriella Schilli [b], Mónika Bálint [b, c], István Horváth [d], Csaba Hetényi [b, *]

[a] MTA NAP-B Molecular Neuroendocrinology Group, Institute of Physiology, Szentágothai Research Center, Center for Neuroscience, Medical School, University of Pécs, Szigeti út 12, 7624, Pécs, Hungary
[b] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624, Pécs, Hungary
[c] Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117, Budapest, Hungary
[d] Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720, Szeged, Hungary

ARTICLE INFO

Article history:
Received 3 November 2017
Received in revised form
27 March 2018
Accepted 21 April 2018
Available online 27 April 2018

Keywords:
Water
Mobility
Temperature
Networks
Molecular dynamics

ABSTRACT

Advancement of computational molecular dynamics allows rapid calculation of large biomolecular systems in their water surroundings. New approaches of prediction of hydration networks of biomolecular surfaces and complex interfaces are also based on molecular dynamics (MD). Calculations with explicit solvent models can trace thousands of water molecules individually on a real time scale, yielding information on their mobility, and predicting their networking with biomolecular solutes and other water partners. Here, we investigate the effect of key parameters of molecular dynamics simulations on the quality of such predictions. Accordingly, systematic scans on temperature, pressure, force field, explicit water model and thermodynamic ensemble are performed. Explanations of optimal parameter values are provided using structural examples and analyses of the corresponding hydration networks.

1. Introduction

Structure and function of water is a central question of science and technology. Water is important as bulk solvent, reaction/interaction partner, and building block. There are specific examples where networks of hydrogen bonds formed between water and protein molecules stabilize active protein conformations [1,2], and promote the folding process [3–5]. Water molecules mediate the assembly of β-amyloid protofilaments of Alzheimer's disease [6–10], participate in the binding process of protein-RNA association [11] and playing a fundamental role in proton transfer reactions [12–20]. Entrapped in the protein interior, they have a special stabilizing effect [21,22]. Precise structure-based drug design requires the inclusion of water molecules influencing ligand binding [23–35].

There are static and dynamic computational methods of prediction of the above hydration networks of protein surfaces and interfaces. Among static methods there are knowledge-based [36,37], structural [38,39], and docking-based [40] approaches. Molecular dynamics (MD) combined with explicit water models can handle all interactions of water molecules including not only solute-water, but also water-water contacts. They also account for the exchange between the bulk and bound waters. MD has become a powerful engine of various methods of prediction of hydration structure of proteins targets and their complexes with ligands [41–45]. It is widely applied in drug discovery [46–49], and the analysis of protein-protein interactions [50]. With the continuous development in hardware and software technology modelling of structural changes occurring at computationally demanding time scales and large molecular systems has become feasible. GPU calculations of non-bonded interactions [51–53], and the appearance of dedicated hardware [54,55] have also expanded the frontiers of MD calculations.

Present MD-based predictors of hydration structure can be further categorized. There are approaches for calculation of average solvent densities and proximal radial distribution functions around different atom types derived from MD simulations to predict hydration shells around proteins [42,56–61], or to analyse hydration structure around macromolecules [59,61]. Other, occupancy-based

[*] Corresponding author.
E-mail address: csabahete@yahoo.com (C. Hetényi).

https://doi.org/10.1016/j.jmgm.2018.04.011

algorithms utilize all-atom MD with explicit solvent for analysis [62] and prediction of hydration sites [41,63–69]. The latter MD algorithms work with individual positions of hydrating water molecules instead of average densities and apply various occupancy-based evaluation schemes to obtain hydration sites, like time averaged positions [68] and/or identity-based clustering with mobility scores [70]. The latter approach was found useful for prediction of complete hydration structure of protein surfaces [70] and interfacial hydration networks of target-ligand complex interfaces [71] and it was implemented in a standalone program MobyWat. Hydration structures predicted by the program showed on average 81% and 90% agreements with crystallographic water positions, in the cases of protein surfaces and interfaces, respectively.

Accuracy of predicted hydration structures produced by both approaches may be largely influenced by the various properties of MD simulations. Intensive parameters such as temperature and pressure, the choice of the right force-field, type of ensemble and explicit water model are all determinants of prediction efficiency. For example, explicit water models are widely applied in biomolecular simulations and several solvent models were developed [72]. However, it is often not trivial, which water model is the most suitable for certain simulations and the calculations of water-protein and water-water energetics. It was demonstrated, that the water model can significantly affect the results of the simulations of A-RNA duplexes [73–77], peptide and protein folding [73–77], the calculation of hydration thermodynamics [78,79] and the binding free energy of ligands to proteins [78]. The present study investigates the effect of the above key parameters of MD on prediction quality and provides explanations on the resulted optimal conditions via analyses of corresponding examples on hydration networking.

2. Results and discussion

Standardized methods [70,71] were applied for the prediction of hydration structures. Consecutive MD and a MobyWat steps were used (Methods, Fig. 1) for predictions on protein surfaces (once) and in interfaces (twice). The present study is based on the results of extensive, repeated MD simulations (Methods). A standard validation protocol (Methods) [70,71] was applied for an automated and unbiased qualification of predictions by comparison of predicted and experimental (reference) hydration structures. Formerly, the validation protocol had been applied for calibration of tolerance values and selection of the best clustering scheme for the program MobyWat [70]. Briefly, the validation protocol is based on the identification of matches between experimental and predicted water positions. A match is identified if the distance between predicted and reference water oxygen atoms is below a pre-defined value of match tolerance (mtol, Methods). From the matches, a success rate (SR) value is calculated (Eq. (1), Methods), where the denominator contains the count of reference water positions in the crystal structure which are within the maximal distance limit (d_{max}) as defined in Methods. An SR value of 100% indicates a complete match, and optimal MD conditions. Besides MobyWat predictions, SR values were also calculated for individual MD frames, as well. In these cases, symbol SR_n is used throughout in this text, where n denotes the serial number of an MD frame. Thus, SR_n values are based on raw data referring to the MD frames and used for comparison only.

$$SR = 100\frac{\text{Count of matches}}{\text{Count of reference water positions}}\% \qquad (1)$$

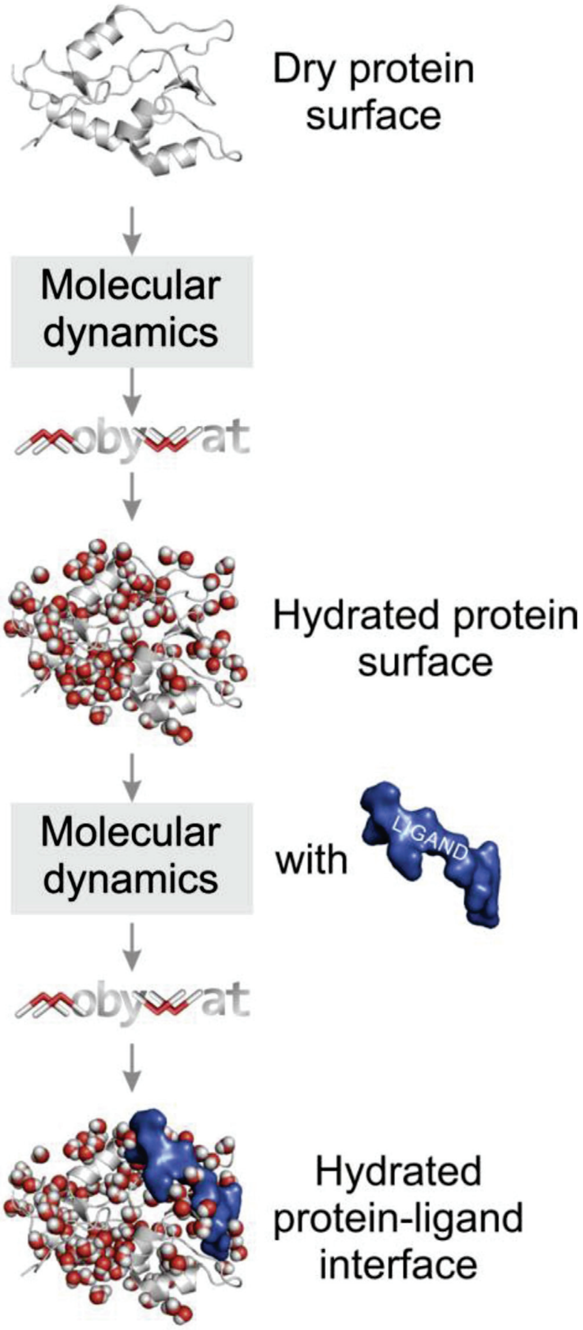


Fig. 1. Prediction of hydration structures of proteins and protein-ligand interfaces using standardized methods with MD simulation and MobyWat steps. In the case of an interface, MD and MobyWat steps were repeated in the presence of the ligand molecule.

2.1. *Simulation temperature*

Beyond algorithmic tolerances and schemes [70,71], the outcome of predictions is also influenced by various parameters of the MD simulations (Introduction). Simulation temperature (T) is a key intensive parameter of MD calculations. It determines the starting velocity distribution of the system, and is inversely proportional to the self-diffusion coefficient of water as proposed by the Stokes-Einstein equation and revisited by recent measurements [80,81].

A branch of MD-based prediction methods such as MobyWat uses occupancy/mobility information (Introduction), and therefore, their prediction quality may largely depend on T. Moreover it was found that the fraction of high mobility water molecules is directly correlating with simulation temperature [71] (Supplementary Fig. S11 of Ref 71). However, previous studies were restricted to a single value (T = 298 [67]' [82] or 300 [70,71] K) and the effect of T on SR has not been investigated. In the present study, a systematic scan was performed on a wide range between 75 K and 368 K including twenty-five individual simulation temperatures and the corresponding change in prediction quality. The Amber99SB-ILDN [83] force field combined with TIP3P water model [84] was applied at all temperatures.

2.1.1. *Surface hydration*

In the first part of the investigations, systems 2O9S, 2VB1, 3NIR were involved (Table 1). Predictions of water positions were performed on the entire protein surfaces. All predictions were reproduced five times with different initial velocity distributions (Methods), and the corresponding average and standard deviation of SR values were calculated and shown in Table S1. A gradual increase in SR was observed between 75 K and 200 K (Fig. 2, S1-S2). The average SR of predictions increased by 20, 25, and 20%, for the above three systems, respectively (Tables S1–S3). On higher temperatures SR fluctuates around 80% with the highest mean SRs of 82, 89 and 84%. These findings can be readily explained by simple considerations using Eq. (2). At low temperature (75 K) diffusivity of water molecules is low (Eq. (2)), their movement is restricted (frozen) and they mostly vibrate around their initial positions. Thus, less raw water positions are piped into the clustering algorithm of MobyWat from the trajectories.

Consequently, at a high temperature (368 K) more water positions can be predicted than at a low temperature (75 K) for the same protein (Fig. 2) which is one reason of the increase in SR. For additional explanations, network graphs of the predicted hydration structures were produced by the NetDraw mode of MobyWat (Methods). Based on graph information nodes and edges were classified (Methods) into static and dynamic categories. The ratio of dynamic nodes among all nodes was found more than 15% higher at 368 K than at 75 K (Table S4). This resulted in extensive dynamic subnet regions at 368 K (red edges in Fig. 3) interconnecting and stabilizing water positions in the static subnets [71], and an

Table 1
Systems investigated in the present study.

PDB code	System (ligand)	Resolution (Å)	Crystallization T (K)	Data collection T (K)	Data collection p (Mpa)[a]	No. of waters[a] [b]
Surface hydration						
2O9S	SH3 domain from ponsin	0.83	na[g]	100	na	108
2VB1[c]	Hen egg white lysozyme	0.65	292	100	na	144
3NIR	Crambin	0.48	na	100	na	65
Case study						
4WLD[d]	Hen egg white lysozyme	1.54	293	298	0.1	93
4WLT[d]	Hen egg white lysozyme	1.60	293	298	190	109
4WLX[d]	Hen egg white lysozyme	1.60	293	298	280	117
4WLY[d]	Hen egg white lysozyme	1.62	293	298	380	127
4WM1[d]	Hen egg white lysozyme	1.60	293	298	500	129
4WM2[d]	Hen egg white lysozyme	1.60	293	298	600	135
4WM3[d]	Hen egg white lysozyme	1.55	293	298	710	137
4WM4[d]	Hen egg white lysozyme	1.60	293	298	800	144
4WM5[d]	Hen egg white lysozyme	1.60	293	298	890	151
1BGI[c]	Hen egg white lysozyme	1.70	310	283	na	96
1IEE[c]	Hen egg white lysozyme	0.94	293	110	na	151
1LPI[c]	Hen egg white lysozyme	2.00	na	278	na	64
1V7S[c]	Hen egg white lysozyme	1.14	313	290	na	138
2F2N[c]	Hen egg white lysozyme	1.60	300	277	na	133
2Z18[c]	Hen egg white lysozyme	1.15	Na	90	na	143
3LZT[c]	Hen egg white lysozyme	0.93	296	120	na	182
3WPJ[c]	Hen egg white lysozyme	2.00	293	300	na	61
3 ZEK[c]	Hen egg white lysozyme	1.43	na	298	na	77
4AGA[c]	Hen egg white lysozyme	1.50	na	63	na	136
4LYO[c]	Hen egg white lysozyme	2.05	>297[f]	285	na	47
4LZT[c]	Hen egg white lysozyme	0.95	296	295	na	124
4NGI[c]	Hen egg white lysozyme	1.70	288	125	na	113
4ZIX[c]	Hen egg white lysozyme	1.89	na[e]	273	na	82
Interface hydration						
3RO3	G-protein-signaling modulator 2 (QVDSVQRWMEDLKLMTE)	1.10	288	100	na	12
3U43	colicin-E2 immunity protein (colicin-E2)	1.72	na	100	na	22
4H9O	death domain-associated protein 6 (histone H3.3 G90 M mutant/ H4)	2.05	277	na	na	35

Notes to Table 1.
[a] Atmospheric pressure was assumed, where no pressure data was provided.
[b] Number of crystallographic water molecules under bmax = 100 Å^2 and dmax = 3.5 Å.
[c] HEWLs used in the simulations investigating the efficacy of simulating on data collecting temperature.
[d] HEWLs used in the simulations investigating the effect of simulation pressure.
[e] Room temperature was specified.
[f] According to Wang et al.
[g] Non-available.

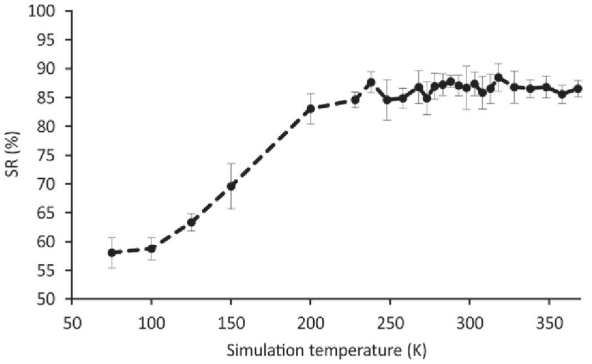


Fig. 2. The effect of simulation temperature on success rates of prediction of hydration structure of protein surface (system 2VB1). Each data point represents an average succes rate value calculated from five simulations with different starting velocity distributions. Error bars denote standard deviations.

increase of SR. Such a stabilization effect of dynamic networking cannot work at 75 K, where rarely interconnected static (blue) regions dominate in a relatively small graph. A close-up of a representative situation is shown in Fig. 4, where static water positions, 3611, 3666 and 3674 connected to D835 and E839 and positioned in deep pockets were reproduced at both temperatures, and the dynamic ones were found only at 368 K (Fig. 4). In the hydration graph produced at 75 K small separated sub-graphs can be observed. On 368 K the number of sub-graphs was increased, densely interconnected by dynamic water nodes (3663 and 3729) located on flat surfaces.

Whereas efficient exploration of available binding sites requires the increase of temperature, very high temperature may also have an antagonistic effect on SR resulting in high mobility, low clustering occupancy, and finally, the saturation of SR curves (Fig. 2, S1-S2) beyond 200 K. Besides per-trajectory SR values of predictions, such differences between temperatures can be illustrated by per-frame SR_n values calculated by the Analysis mode of MobyWat (Methods) and the corresponding statistics is provided in Tables S5 and S6. For each system and starting velocity distribution combination (MD trajectory) there is a single SR value and 1001 SR_ns. Thus, SR_n reflects fluctuation (evolution) of hydration layer during an MD simulation. Statistics of SR_n shows that its fluctuation is temperature-dependent (Tables S5 and S6). An example of protein

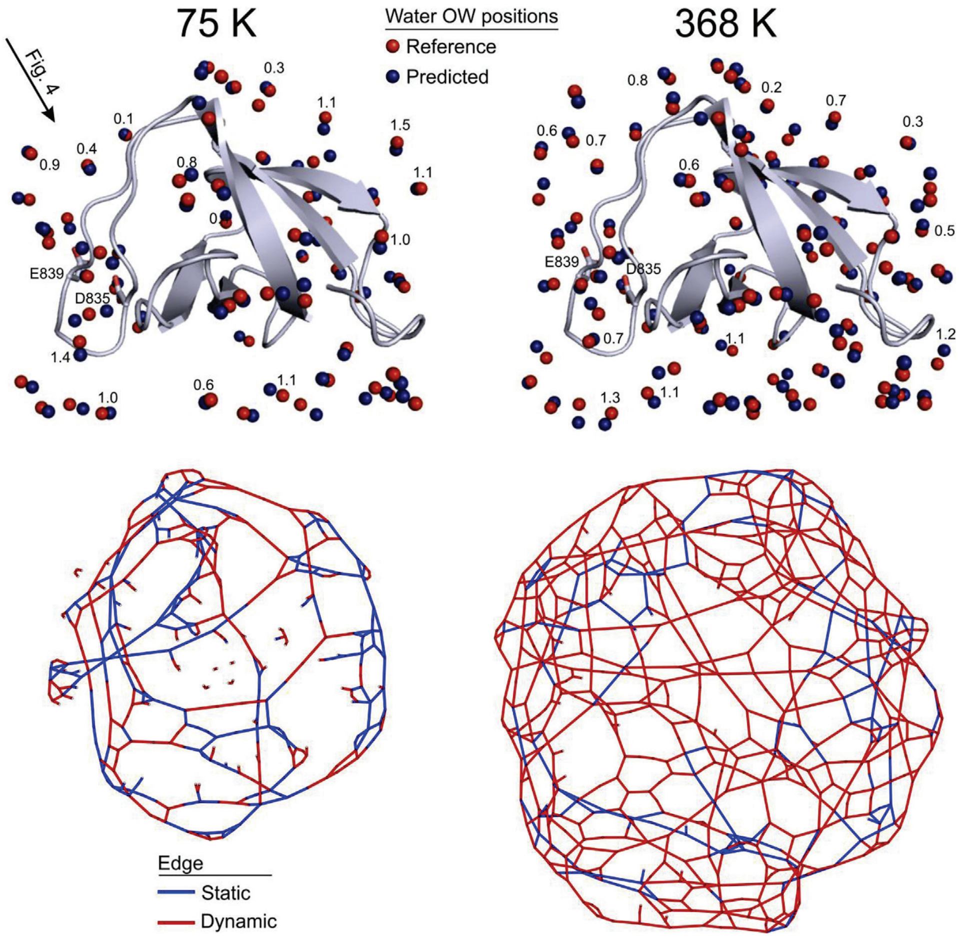


Fig. 3. Top: matches of predicted and experimental water positions on the surface of SH3 domain from ponsin (2O9S) at 75 and 368 K, respectively. Match distances are shown in Å for some of the matches for comparability. Bottom: the corresponding hydration network graphs produced from predicted water positions. The arrow points to a region around negatively charged amino acids D835 and E839 further discussed in Fig. 4. Further details of the analysis of temperature dependency of the hydration network are provided in Supporting Table S14 and Fig. S12.

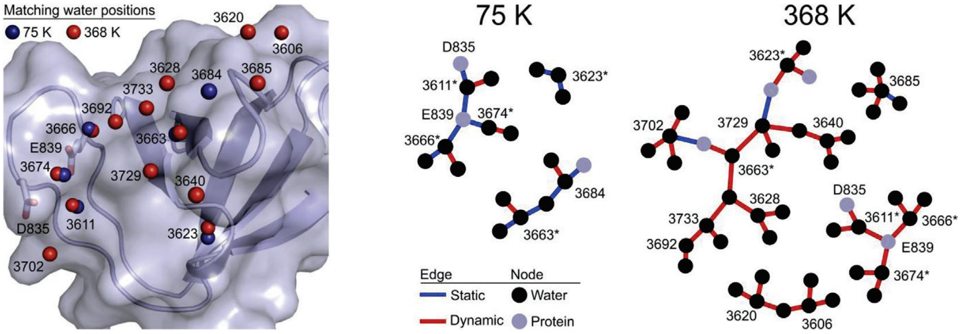


Fig. 4. A close-up of predicted water positions at 75 and 368 K matching with the crystallographic reference positions of system 2O9S. The figure corresponds to a negatively charged region of the protein as indicated by an arrow in Fig. 3. The numbers correspond to residue numbers of waters in 2O9S. The subgraphs contain the first connection sphere of predicted water molecules. Positions found at both temperatures are marked with asterisk on the 75 K subgraph.

HEWL (system 2VB1) in Fig. 5 shows that the fluctuation of per-frame SR_n values is significantly higher at 368 K than at 75 K, due to the above-mentioned differences in water mobility. The average SR_n values are also higher on 368 K than on 75 K (Tables S5 and S6) remain below the average SR (84%) achieved with IDa clustering.

2.1.2. *Interface hydration*

In the cases of complex (protein-protein and protein-peptide) interfaces of systems 3RO3, 3U43, 4H9O (Table 1) a similar trend can be observed (Fig. 6, S3-S4) as in the previous three cases of hydration of protein surfaces. On the same interval (T = 75–368 K) the increment of average prediction SRs is significant. Analysis runs were performed on trajectories of interface calculations too. Fluctuations in SR_n followed similar patterns to analysis of protein surface simulations (Tables S10 and S11).

However, the standard deviation of SRs of the five simulations of interface predictions is generally larger than that of surface predictions (Tables 1–3 and S7-9). This can be explained by the small count of reference water positions in the interfaces, where misprediction of 1–2 positions can cause large drop in the SR value (Eq. (1)). Even the largest complex interface (4H9O) has ca. half of the reference positions of the smallest surface system (3NIR, Table 1). Interestingly, the standard deviation of SR values of 3U43 predictions decreases with increasing temperature (Fig. 7), similarly to other interface systems (Figs. S5 and S6). This trend means that high temperature is beneficial for reproducibility and robustness in the cases of interface predictions. This trend was not observed for surface predictions (Figs. S7–S9). The differences in the above trends of surface and interface predictions can be explained by several factors. First of all, the faces of target and ligand solutes restrict water molecules in the narrow volume of the interface, and therefore, their translational and rotational freedoms are limited if compared to those of surface water molecules. There is also a moderate exchange between interface and bulk regions

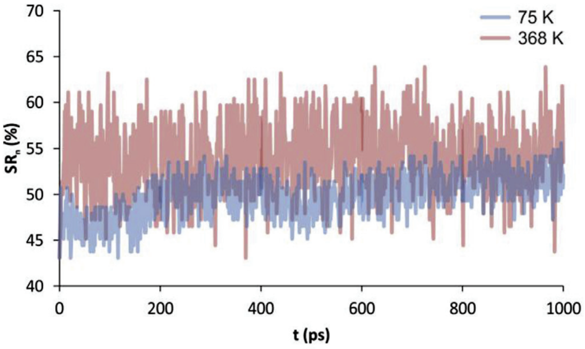


Fig. 5. Per-frame SR_n values of system 2VB1 (initial velocity distribution 1), at 75 and 368 K.

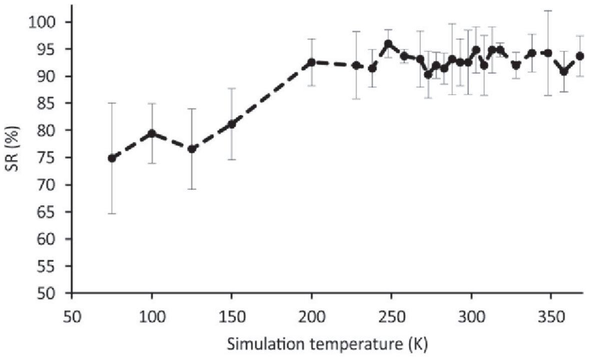


Fig. 6. The effect of simulation temperature on success rates of prediction of hydration structure of complex interface (system 4H9O). Each data point represents an average success rate value calculated from five simulations with different starting velocity distributions. Error bars denote standard deviations.

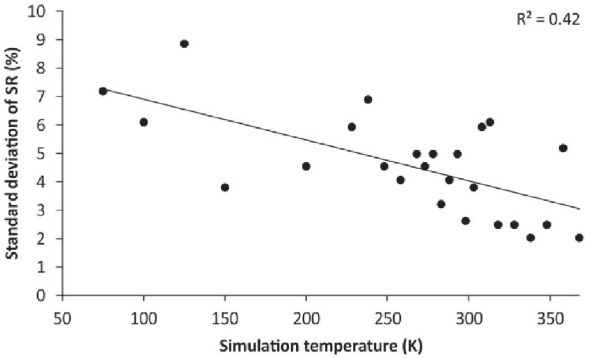


Fig. 7. The effect of simulation temperature on the standard deviation of success rates of prediction of hydration structure of a complex interface (system 3U43). Each data point represents a standard deviation calculated from five simulations with different starting velocity distributions.

due to the close topography of the target-ligand interface. Thus, the overall mobility of interface water molecules is topologically restricted. At high temperatures, high water mobility (Eq. (2)) overrides these restrictions, the interface will become accessible and a good ensemble of water positions are provided in each of the five trajectories piped into MobyWat clustering. In other words, the use of high temperature results in an increased mobility of water molecules which improves the efficiency of the occupancy-based algorithm of MobyWat. Thus, the difference between the resulted SR values and its standard deviation will decrease. In the cases of surface hydration where such topological restrictions did not apply, water molecules can freely occupy hydration sites in all five trajectories and the increase of T will not correlate with standard deviation of SR.

At 75 K higher SRs could be observed for interfaces than for surfaces (Figs. 2 and 6), and the highest average SR for surfaces is 60% (3NIR) 15% lower than the same value for interfaces (75%, 4H9O). Even full match with references was achieved (3RO3, T = 100 K, seed 5, Table S7). Finding all interfacial waters was not uncommon, while the best SR of surface predictions was 91% (system 2VB1, T = 318 K, seed3, Table S1). With high SRs found at both low and high temperatures (system 3RO3, 92% on 75 K and 100% on 368 K, Table S7), differences can be observed in the corresponding hydration graphs of a representative region at residue E82 (Fig. 8). Extensive static subnets are formed at 75 K with five static of six total edges starting from central E82. The network is scattered to several subnets, and dynamic waters linking static regions to each other are missing. At 368 K, the situation changes, mostly dynamic links are formed around ligand residue E82, instead of the static one and the majority of the hydration network became dynamic. The high SRs of interface predictions and their relative immunity to temperature changes is a consequence of the use of the ligand molecule and duplicate steps (Fig. 1, Methods) in the prediction scheme.

2.1.3. A systematic case study

The structure of Hen egg-white lysozyme (HEWL) had been resolved at fifteen different data collecting temperatures (Table 1) ranging from 63 K to 300 K. Hydration structures of all fifteen HEWL structures were calculated for each data collecting temperature using mobility data of altogether 225 MD simulations (=15 protein structure × 15 temperatures), using Amber99SB-ILDN [83] force field in combination with TIP3P water model [84]. The results in Fig. 9 show that larger match can be obtained with reference water positions using simulation temperatures over 273 K than at cryogenic conditions (<125 K). SRs above 80% can be found mostly in the range of 273–300 K. The maximal SR obtained between 63 and 125 K was only 77%, while the overall maximum SR (97%) was achieved at 283 K (Table S13).

Apparently, simulating at cryogenic conditions, is not the best condition for prediction of hydration structure. There are at least two reasons for this. (i) In cryo-crystallography a flash-cooling concept is applied including fast freezing of a molecular structure equilibrated at the crystallization (usually room) temperature [85–88] to a cryoscopic temperature. Thus, the structure is representative for a crystallization- and not cryoscopic temperature. Accordingly, hydration structures of lysozyme were better calculated from simulations at crystallization temperatures than those at cryogenic temperatures, in terms of SR. (ii) The reduced mobility of water molecules at low temperatures can also result in low SRs as it was described in the previous Sections. Thus, this systematic case study also showed that good predictions can be achieved with simulations close to crystallization (room) temperature instead of cryogenic data collecting temperatures.

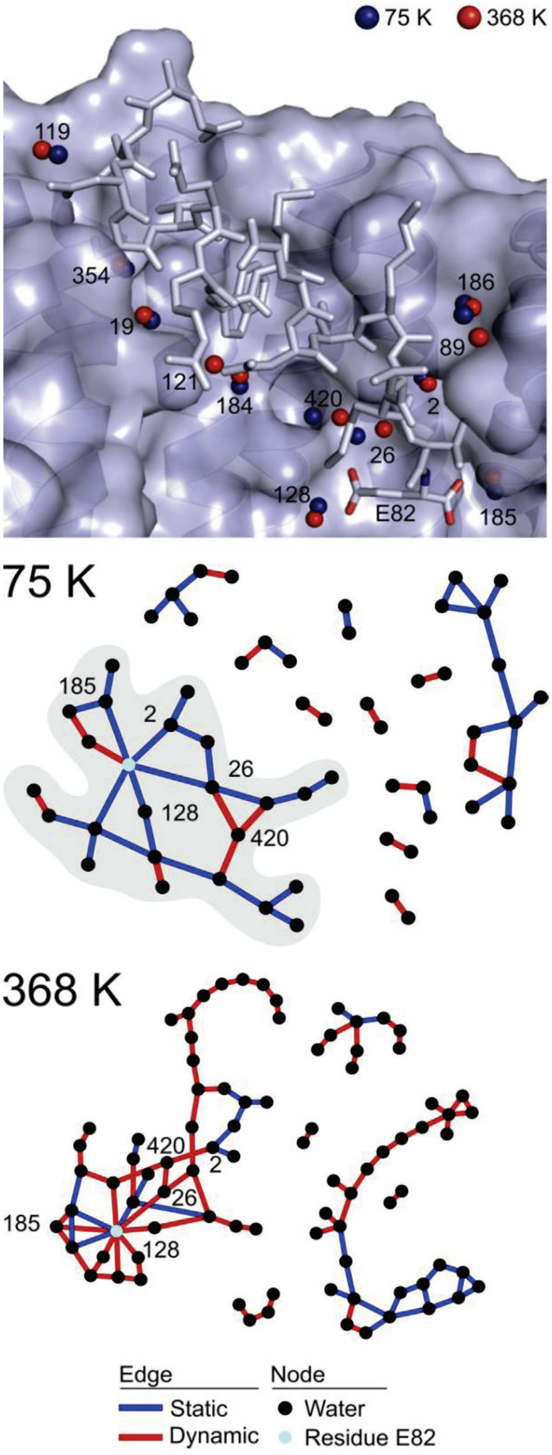


Fig. 8. Predicted water positions at 75 and 368 K matching to the crystallographic reference positions of system 3RO3. The numbers correspond to residue numbers of waters in 3RO3. Static subnet around ligand residue E82 can be observed at 75 K (grey area). This core region turns into dynamic at 368 K, where dynamic waters surround fragmented static subnet.

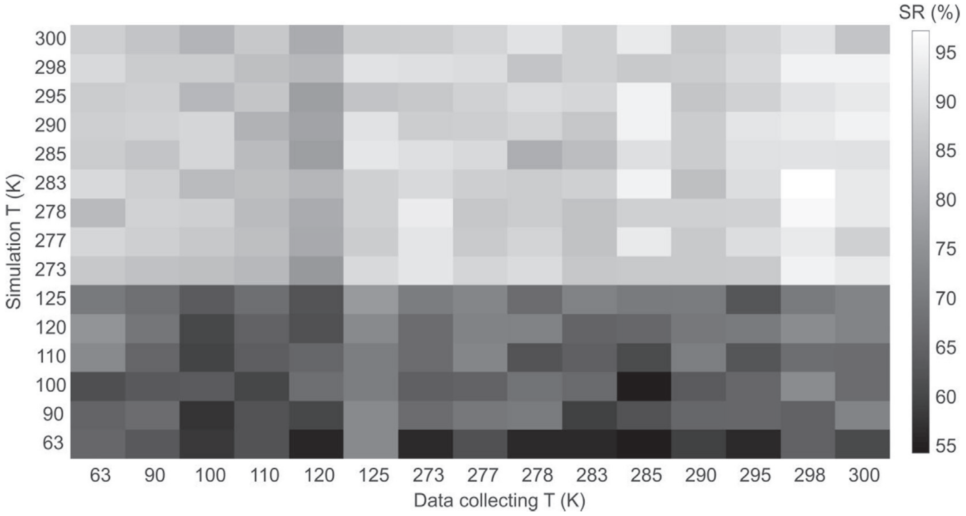


Fig. 9. Success rate matrix from predictions of lysozyme structures solved at different data collecting temperatures.

2.2. Force field and water model

The selection of force field/water model combination may also affect prediction quality. Besides Amber99SB-ILDN/TIP3P force field-water model combination used in previous MobyWat [70,71] predictions, three other water models TIP4P, TIP4P-Ew, and TIP5P were also investigated in the present study. In the four-point TIP4P [84] and TIP4P-Ew [89] models, a virtual site with negative charge was added along the bisector of the H-*O*-H angle to improve the electrostatic distribution around the molecule. The five site model, TIP5P [90] has also two dummy atoms representing the lone pairs of the oxygen. In a previous study [70], an OPLS-AA/TIP4P combination was investigated for prediction of surface hydration. Here, results on interface hydration are provided. System 4H9O, with extensive interface network was calculated with the aforementioned five force field/water model combinations at a constant 300 K. The five-point TIP5P model gave the best success rate (Table 2), nevertheless the differences with others are marginal. The reproduction of the experimental hydration layer with different initial velocities was successful regardless the combination as demonstrated by low errors in Table 2. The OPLS-AA/TIP4P combination also performed well, reinforcing our previous results [70]. Mobility-based prediction shows little dependence from force field/water model combination at room temperature.

Comparison of TIP3P and TIP5P water models was also performed on an extended temperature scale of 75–368 K in combination with the Amber99SB-ILDN force field in both cases. In the interval of 75 and 268 K SRs calculated with TIP5P are lower than SRs with TIP3P (Fig. 10.). While increments in SR are observable in both cases, SR of TIP5P calculations continues growing after 200 K, where growth in TIP3P stops until 268 K. SRs by the two models become indistinguishable after 268 K. The observed difference at low temperature regions is understandable considering that the mobility of TIP3P [75] water is higher than that of TIP5P [91] also reflected by the relatively low melting temperature of water in TIP3P model.

Table 2
Effect of different force field/water model combinations on interface prediction.

Force field	Water model	Mean SR (%)[a]
Amber99SB-ILDN	SPC/HW[b,c]	82.6
Amber99SB-ILDN	TIP3P[c]	87.0
Amber99SB-ILDN	TIP3P[d]	92.6 ± 5.6
Amber99SB-ILDN	TIP4P[d]	93.7 ± 4.7
Amber99SB-ILDN	TIP4P-Ew[d]	92.6 ± 3.3
Amber99SB-ILDN	TIP5P[d]	96.5 ± 4.7
OPLS	TIP4P[d]	93.7 ± 3.1

Notes to Table 2 [a]Prediction parameters were: ctol = 1.0 Å, ptol = 2.5 Å, bmax = 100 Å^2 and dmax = 3.5 Å, IDa clustering. [b]Modified SPC/E water model. [c]The examined structure (pdb ID: 1V7S) contain D_2O. [d] [c]The examined structure (pdb ID: 4H9O) contain H_2O.

2.3. Simulations with heavy water

As various experimental techniques apply heavy water (D_2O) as a solvent, it may be interesting to check the outcomes of MobyWat predictions on systems measured with hydrating heavy water molecules. The experimental structure of a system of the above case study 1V7S (Hen egg-white lysozyme) had been determined

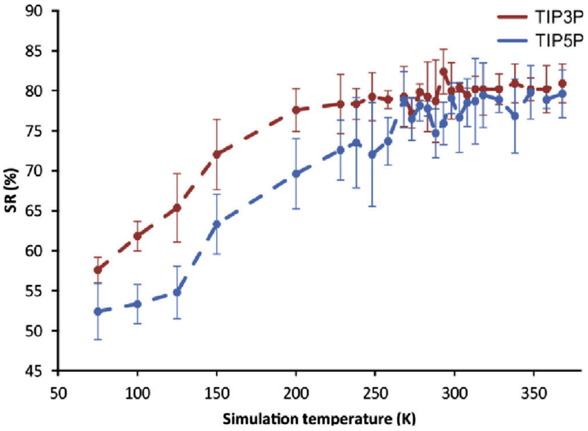


Fig. 10. Success rates of prediction of hydration structure of a protein surface (system 2O9S) using two different explicit water models. Each data point represents an average success rate value calculated from five simulations with different initial velocity distributions. Error bars denote standard deviations.

with heavy waters [92]. The reproduction of the experimental reference positions was successful with H_2O. In a next step, the hydration structure was re-calculated with D_2O, as well. Notably, the use of D_2O instead of H_2O should have no effect on the observed configurations as sampling time is sufficient in this case. However, the use of D_2O provided a good example for involvement of a modified SPC/E water model SPC/HW [93] in the study, applied with Amber99SB-ILDN force field [83] (Table 2). As the 1V7S system was investigated on total of fifteen different simulation temperatures in the above case study, fifteen runs with D_2O were also done in this temperature range. Other MD and prediction parameters were not altered. It was found that the reproduction of experimental hydration structure was successful with the heavy water simulations too. Average of differences of SRs calculated on fifteen different temperatures between predictions with H_2O and is 5.5% (Table S13) showing marginal differences in reproduction of reference water positions. At 290 K, the data collecting temperature of 1V7S the difference between light and heavy water-based predictions of surface hydration was also marginal (4%, Table S13).

2.4. Ensemble type

Similar to assessment of different temperature, pressure and force field/water model combinations calculations of an MD ensemble other than the NPT could be also informative. In the NPT ensemble, the temperature is coupled to an external heat bath and pressure is kept constant. Simulations can also produce an NVT (or canonical) ensemble, where the volume is kept constant instead of pressure. The hydration structures of two surface (2O9S, 3NIR) and one interface examples (4H9O) were calculated under NVT conditions with Amber99SB-ILDN [83] force field in combination with TIP3P water model [84]. The results are presented on Table S14, showing that surface predictions are not influenced by the type of the ensemble, as the differences in SR are marginal. For interface predictions NVT showed somewhat better performance (average SR) and reproducibility (smaller standard deviations) than NPT.

2.5. Pressure

As data collection temperature, pressure can also vary in crystallographic measurements. For example, the structure of lysozyme was solved at several pressures with high pressure protein crystallography (HPPX) [94]. At high pressures, the internal cavity volumes are compressed, implicating changes in the hydration network on the surfaces [94]. The number of (reference) water positions determined in crystal structures is increasing with pressure. As described in the original publications of the lysozyme structures, at high pressures, waters can appear on hydrophobic surfaces, and intrude into previously unfilled cavities [94]. The present predictions were done with Amber99SB-ILDN [83] combined with TIP3P water model [84]. It was found that SR is still decent on the highest pressure, 890 MPa (82%), but finding waters deep in the protein structure is harder than those situated on the surface. SR shows a surprisingly strong anti-correlation with simulation pressure (Fig. 11, $r^2 = 0.93$). As pressure also influences the self-diffusion of water [95,96], the decrease of mobility of water with increasing pressure may explain the decrease of SR similarly to the considerations detailed previously in the Section about the effect of temperature.

3. Conclusions

Among the investigated parameters of MD simulations, temperature has the largest influence on the efficiency and reproducibility of prediction of hydration structure. The MD-based prediction algorithm of MobyWat is not very sensitive to pressure and ensemble type. Various combinations of force fields and explicit water models can provide adequate trajectories for the predictions. By default, an NPT simulation at crystallization conditions of 300 K and 0.1 MPa combined with a TIP3P water model and a decent force field is sufficient for good predictions. The reproducibility of the results at various conditions demonstrated the robustness of the occupancy-based approach and the identity-clustering scheme of MobyWat. Five different combination of force field-water model systems were investigated and the results suggest that the approach is fairly independent on these combinations, and probably other known force fields and water models could be also used in future predictions with similar success. Furthermore, the validation protocol of the present study can be recommended as a standard tool of testing new force fields and explicit water models for their efficiency in calculation of hydration structure. Based on our results, the Mobywat is powerful in predicting hydration structure from MD trajectories, and constructing networks of water molecules.

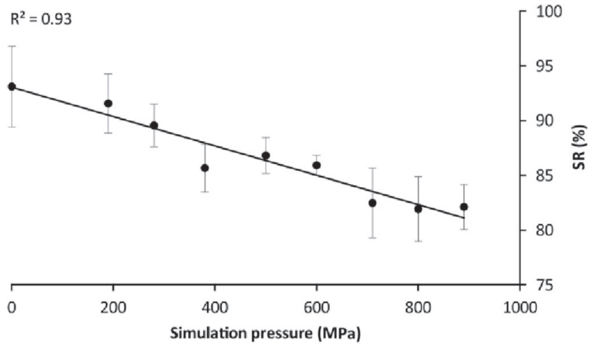


Fig. 11. The effect of simulation pressure on success rates of prediction of hydration structure of surfaces of lysozyme (Table 1). Each data point represents an average success rate value calculated from five simulations with different seeds. Error bars denote standard deviations.

4. Methods

4.1. Systems

Three protein systems (2O9S, 2VB1, 3NIR) were selected randomly from previous study [70] for surface predictions, and another three protein-ligand complexes (3RO3, 3U43, 4H9O) for interface predictions [71]. For a case study and investigations of pressure-dependence of prediction quality, a series of lysozyme structures were used as references (Table 1).

4.2. Preparation of an MD run

A two-step energy minimization protocol prior to molecular dynamics simulations was applied for all systems of Table 1. The following standard protocol was applied for all proteins throughout this study. All calculations were performed with programs of the GROMACS software package, using the AMBER99SB-ILDN force field [83] and OPLS force fields. Water molecules were explicitly calculated using TIP3P [84], TIP4P, TIP4P-EW, SPC/E, SPC/HW and TIP5P water models. In the first step, crystallographic water molecules were removed and the dehydrated (dry) protein was placed in a simulation box. The distance between the solute and the box was set to 10 Å. The box was filled with water molecules and counter-ions were added to neutralize the system.

4.3. Energy minimization

A two-step energy minimization protocol prior to molecular dynamics simulations was applied for all systems of Table 1. In the first step a steepest descent (sd) optimization was done, with the convergence threshold set to 10^3 kJmol^{-1}nm^{-1}. Conjugate gradient (cg) calculation was done in the second step, the convergence threshold was changed to 10 kJ mol^{-1}nm^{-1}.

Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJmol^{-1}nm^{-2} in both steps. Distance restraints (in case of system 3U43) were applied between structural ions and coordinating amino acid residues with force constant of 10^3 kJmol^{-1}nm^{-2}. The optimization protocol was performed twice for interface predictions, once for the pure target and once for the re-assembled target-ligand complex.

Before network analyses, a 4-step protocol was applied for energy minimization of predicted water positions following an sd-cg-sd-cg pattern with parameters of sd and cg methods described above. During the first two steps, all solute heavy atoms and the oxygen of the predicted interfacial water molecules were position restrained and bulk waters and ions were released. In the last two steps, position restraints were not applied on predicted waters, only solute heavy atoms were position restrained.

4.4. Molecular dynamics simulations

After energy-minimization and setting up velocity distribution, 1-ns-long MD simulations were carried out with a time step of 2 fs. Simulations of 2VB1 and 2O9S at 75 K were also extended to 10 ns. The velocity rescale algorithm [97] was applied for temperature coupling, solute and solvent were coupled separately with a coupling time constant of 0.1 ps. Fifteen lysozyme structures (marked with ^{c} in Table 1) were simulated at their data collecting temperatures (Table 1). The three complex systems, proteins 2O9S, 2VB1 and 3NIR were simulated on twenty-five different temperatures (75, 100, 150, 200, 228, 238, 248, 258, 268, 273, 278, 283, 288, 293, 298, 303, 308, 313, 318, 328, 338, 348, 358 and 368 K). Investigations on the influence of simulation temperature, pressure and ensemble type were performed using Amber99SB-ILDN [83] force field in combination with TIP3P water model [84].

Lysozyme structures resolved different pressures were simulated on 300 K (marked with ^{c} in Table 1). In NPT simulations, pressure was coupled the Parrinello-Rahman algorithm [98–100] with a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1} and reference pressure of 1 bar. Particle Mesh-Ewald summation was used for long range electrostatics. Van der Waals and Coulomb interactions had a cut-off at 11 Å. Coordinates were saved at regular time-intervals of 1 ps yielding 1.001×10^3 frames. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJmol^{-1}nm^{-2}. MD calculations were performed twice for interface predictions, once for the pure target and once for the re-assembled target-ligand complex. Periodic boundary conditions were treated before predictions, to centre the solute in the box and make molecules whole. The calculations were done with the Gromacs 5 package [101].

4.5. Initial velocity distribution setup

MD trajectories produced for the same system may differ substantially from each other [70] because of the applications due to hardware-dependent rounding of floating point calculations, the use of dynamic load balancing in parallel execution and so on. To test the reproducibility of the methods, all MD calculations were reproduced with five different initial velocity distributions. Practically, this can be done by selecting different seed numbers of the velocity generator routine The distribution of initial velocities of particles are generated from the Maxwell-Boltzmann distribution. In the five runs, the random number generator was seeded with five different seed numbers, and velocity generation temperature was set to simulation temperature. In the case of interface predictions, the five different seed numbers were set for the interface and surface MD run pairs.

4.6. Mobility-based prediction of hydration structures

Throughout the present study, an open source program MobyWat [70,71] was used for prediction of water positions. MobyWat is accessible at www.mobywat.com, and derives mobility information of movements of solvent molecules from MD calculations. The mobility information obtained from the trajectory is transformed into the hydration structure during the prediction process. For MobyWat evaluations reference and candidate pools of water molecules were separated with possible structural role to distinguish them from bulk waters of no use. A maximal distance limit (d_{max}) was used for the distinction. In the case of surface hydration, a water molecule is selected for the reference or candidate pools if a distance measured between its oxygen atom and the closest heavy atom of the target is less than or equal to d_{max}. In case of interface hydration, the same distance criterion was checked for the target and ligand molecules at the same time. Occurrence of a molecule in the pools with the same ID (the atom and the residue serial numbers of the water oxygen atoms) is counted during the whole trajectory and the count is registered in the list as an occupancy number corresponding to the ID. That is, the value of an occupancy number in the list is increased if a pool includes the water with the ID in question. After evaluating all candidate pools, occupancy lists are sorted by decreasing occupancies. Spatial clustering step was introduced Water molecules of different pools with the same ID (belonging to the same row of the occupancy list) are collected into a cluster using a pre-defined clustering tolerance (ctol) value. In the final step, MobyWat creates prediction lists from the cluster lists. Prediction lists contain the Cartesian atomic coordinates of water positions and the corresponding mobility (M_i) values as final outcomes of the prediction process. Mobility (M_i) value is calculated for each row of the prediction lists from normalized occupancy (O_i) values (Eq. (2)).

$$M_i = 100\frac{O_{max} - O_i}{O_{max} - O_{min}}, \quad O_{max/min} = \underset{i}{\max/\min} O_i \tag{2}$$

M_i values scale between 0 and 100. Zero corresponds to the least mobile predicted water position, one hundred corresponds to the most mobile one. Detailed description of the prediction algorithm can be found in Section S2.1.5 of Ref. [70] and in the User's Manual of MobyWat.

4.6.1. Surface predictions

Hydration structures of protein surfaces were deducted from a single MD simulation with 1001 coordinate snapshots. After treating periodic boundary conditions, the trajectory was processed with MobyWat. The prediction tolerances (ptol) were set accordingly to previous study [70], run parameters (dmax, clustering and prediction tolerances) were set to 3.5, 1.0 and 2.5 Å, respectively, where dmax refers to the maximum distance of a predicted water oxygen from closest target heavy atoms. The IDa (all-inclusive) clustering algorithm was applied in prediction process. IDa clustering identifies a candidate water molecule by its atom or residue serial numbers and uses the history of residence of each molecule on target surface for mobility calculations. Detailed description of the clustering algorithm can be found in Section S2.1.5 of Ref. [70]

and in the User's Manual of MobyWat. All 1001 coordinate snapshots from MD trajectories were used.

4.6.2. Interface predictions

Hydration structures of complex interfaces were calculated with the M3 protocol of MobyWat [71]. M3 is an advanced protocol for prediction of void-free hydration structure of the target surface which reduces the amount of cavities and to produces a complete hydration structure of the interface [71]. Briefly, in M3, the surface of the free target molecule is hydrated without the ligand using short MD and a MobyWat steps. Having an (over)loaded target surface, the ligand is positioned back so as to form the interface with the target. This interface can be considered soaked, having as many water molecules as physically possible to reduce the volume of unwanted void spaces. Finally, additional MD and MobyWat steps are performed to re-equilibrate all interactions of water molecules in the presence of the ligand.

Thus, in the beginning of M3, the target structure without ligand and crystallographic waters were simulated and surface prediction was performed for the entire surface of the target with dmax, clustering and prediction tolerances of 5.0, 1.5 and 2.5 Å, respectively, where dmax refers to the maximum distance from closest target heavy atoms. The IDa-clustering algorithm was applied. The ligand-target complex was re-assembled after hydration of the target surface. For this, the target part of the holo and the hydrated apo systems were fitted on the top of each-other and the ligand was used together with the hydrated target (soaking). To select water molecules residing in the target-ligand interface distances between its oxygen atom and the closest heavy atom of target and ligand molecules were measured. A water molecule was considered interfacial if such a distance was smaller than/equal to a pre-defined maximal distance limit (dmax) for both the ligand and target partners. Water molecules conflicting with the ligand structure were excluded using the Editing mode of MobyWat at a minimum distance limit (dmin) of 1.75 Å prior the second MD simulation. After the second MD, the second and final prediction was performed to get the final prediction list. A dmax of 3.5 Å, IDa clustering was applied with clustering and prediction tolerances of 1.0 and 2.5 Å, respectively. In both steps all frames, altogether 1001 snapshots from MD trajectories were used. Additional details can be found in the Results and discussion and Fig. S1 of Ref. [71] and in the User's Manual of MobyWat.

4.6.3. Validation protocol

Validation sub-mode of MobyWat was used for an unbiased, automated comparison of the MobyWat-predicted (Section 4.6.2) and reference (experimental, crystallographic) positions of water molecules. After MD, all frames were extracted from the trajectory. Each frame was fit to the original protein crystal structure, using Cα atoms. In validation sub-mode, predicted waters are compared to reference, experimental waters (reference pool), their proportion gives the Success Rate (Eq. (1)). The reference water molecules are crystallographic positions of water molecules within a dmax of 3.5 Å measured from target (surface predictions), or from both target and ligand (interface predictions) irrespective of their B-factors. The quality of predictions was checked using the Validation sub-mode of MobyWat and expressed as success rates. Matches were identified if the distance between predicted and reference water oxygen atoms was below a pre-defined tolerance of 1.5 Å (match tolerance, mtol). The higher the SR value, the more successful a prediction is in comparison with crystallographic water positions.

4.6.4. Analysis mode

In analysis mode, MobyWat compares the positions of water molecules of a reference structure with positions of water molecules in each (nth) frame of a molecular dynamics calculation. Per frame success rates (SR_n) are calculated according to Eq. (1), with “Count of matches in the nth frame” used in the numerator. That is, the calculation is performed for each (nth) frame of the trajectory without clustering of the frames.

4.6.5. Classification of hydration networks

The NetDraw mode of MobyWat was used for analysis and characterization of the hydration network of protein surface or target-ligand interface. The characterization protocol of hydration network was implemented as the NetDraw mode of program MobyWat (see Methods and Fig. S10). NetDraw determines the interactions of interfacial water molecules with each other, and the bulk water and solute (ligand + target) molecules result in network graphs with mobility assigned to all water nodes. Thus, a water molecule or a residue of the solute was considered as a node. Using PDB files of energy-minimized structures including solute, predicted surface or interface-, and bulk water molecules as an input, NetDraw produces the two-dimensional interaction network graph of the interface as lists of edges and nodes. List of edges of the graph is distilled from the distance files by eliminating redundancies and distances (edges) to carbon atoms (C-filtering). Number of edges per node is limited to four, using the top four shortest edges only (4-filtering).

The list of nodes was produced from the list of edges. All lists are produced with and without considering bulk water nodes. Finally, NetDraw produces a classification of the nodes and edges. A node is classified static if it is a solute (ligand/target) node or connected to four nodes of any type or connected to at least three solute nodes or has a mobility value smaller than or equal to 50. Otherwise the node is classified dynamic. An edge is defined static if it connects two static nodes, otherwise it is dynamic. The mobility values were produced by MobyWat for every (ith) predicted water position during the prediction steps and can be transferred to the network analysis steps as B-factors in the PDB files. Subnetworks are built from edges. Connected static/dynamics edges yield static/dynamic subnetworks. With the above classification NetDraw helps distinguishing between static and dynamic sub-networks of the entire surface or interfacial network. MobyWat provides the hydration network graph with mobility information for the nodes and edges in file formats commonly used by network visualization and analysis programs. Gephi [102] was used for visualization of the graphs produced by NetDraw. Layout was created with the ForceAtlas layout option, with “Attraction distribution”. For Systems 2O9S and 3RO3, NetDraw output files are provided as Supporting Files.

Acknowledgements

We acknowledge a grant of computer time from CSCS Swiss National Supercomputing Centre, and NIIF Hungarian National Information Infrastructure Development Institute. We acknowledge that the results of this research have been achieved using the DECI resource Archer based in the UK at the National Supercomputing Service with support from the PRACE aisbl. The work was supported by the K123836, K112807, K120391grants from the National Research, Development, and Innovation Office. The University of Pécs is acknowledged for a grant PTE ÁOK_KA/2017 and also support in the frame of "Viral Pathogenesis" Talent Centre program. We are thankful to the Gedeon Richter Pharmaceutical Plc. for a predoctoral scholarship to N.J.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jmgm.2018.04.011.

References

[1] L. Nisius, S. Grzesiek, Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network, Nat. Chem. 4 (2012) 711–717.

[2] L. Zhao, W. Li, P. Tian, Reconciling mediating and slaving roles of water in protein conformational dynamics, PLoS One 8 (2013), e60553.

[3] M.S. Cheung, A.E. Garcia, J.N. Onuchic, Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse, P Natl Acad Sci USA 99 (2002) 685–690.

[4] Y. Levy, J.N. Onuchic, Water mediation in protein folding and molecular recognition, Annu. Rev. Biophys. Biomol. Struct. 35 (2006) 389–415.

[5] H. Frauenfelder, P.W. Fenimore, G. Chen, B.H. McMahon, Protein folding is slaved to solvent motions, Proc. Natl. Acad. Sci. U. S. A. 42 (2006) 15469–15472.

[6] D. Thirumalai, G. Reddy, J.E. Straub, Role of water in protein aggregation and amyloid polymorphism, Accounts Chem. Res. 45 (2012) 83–92.

[7] S.H. Chong, S. Ham, Distinct role of hydration water in protein misfolding and aggregation revealed by fluctuating thermodynamics analysis, Acc. Chem. Res. 48 (2015) 956–965.

[8] L. Grisanti, D. Pinotsi, R. Gebauer, G.S. Kaminski Schierle, A.A. Hassanali, A computational study on how structure influences the optical properties in model crystal structures of amyloid fibrils, Phys. Chem. Chem. Phys. 19 (2017) 4030–4040.

[9] K. Jong, L. Grisanti, A. Hassanali, Hydrogen bond networks and hydrophobic effects in the amyloid $\beta_{30\text{-}35}$ chain in water: a molecular dynamics study, J. Chem. Inf. Model. 57 (2017) 1548–1562.

[10] D. Thirumalai, G. Reddy, J.E. Straub, Role of water in protein aggregation and amyloid polymorphism, Acc. Chem. Res. 1 (2012) 83–92.

[11] Y.Y. Li, B.T. Sutch, H.H. Bui, T.K. Gallaher, I.S. Haworth, Modeling of the water network at protein-RNA interfaces, J. Chem. Inf. Model. 51 (2011) 1347–1352.

[12] Y.-T. Kao, X. Guo, Y. Yang, Ultrafast dynamics of Nonequilibrium electron transfer in photoinduced redox cycle: solvent mediation and conformation flexibility, J. Phys. Chem. B 30 (2012) 9130–9140.

[13] S. Sappati, A. Hassanali, R. Gebauer, P. Ghosh, Nuclear quantum effects in a HIV/cancer inhibitor: the case of ellipticine, J. Chem. Phys. 20 (2016) 205102.

[14] Y.K. Law, A.A. Hassanali, Role of quantum vibrations on the structural, electronic, and optical properties of 9-methylguanine, J. Phys. Chem. 44 (2015) 10816–10827.

[15] F. Giberti, A.A. Hassanali, The excess proton at the air-water interface: the role of instantaneous liquid interfaces, J. Chem. Phys. 24 (2017) 244703.

[16] C.A. Daly, L.M. Streacker, Y. Sun, S.R. Pattenaude, A.A. Hassanali, P.B. Petersen, S.A. Corcelli, D. Ben-Amotz, Decomposition of the experimental Raman and infrared spectra of acidic water into proton, special pair, and counterion contributions, J. Phys. Chem. Lett. 21 (2017) 5246–5252.

[17] M. Chen, L. Zheng, B. Santra, H.Y. Ko, R.A. DiStasio, M.L. Klein, R. Car, X. Wu, Hydroxide diffuses slower than hydronium in water because its solvated structure inhibits correlated proton transfer, Nat. Chem. 4 (2018) 413–419.

[18] J. Cuny, A.A. Hassanali, Ab initio molecular dynamics study of the mechanism of proton recombination with a weak base, J. Phys. Chem. B 48 (2014) 13903–13912.

[19] G. Zundel, Hydrogen bonds with large proton polarizability and proton transfer processes in electrochemistry and biology, Adv. Chem. Phys. 111 (2000) 1–217.

[20] H. Ishikita, K. Saito, Proton transfer reactions and hydrogen-bond networks in protein environments, J. R. Soc. Interface 11 (2013) 20130518.

[21] M.A. Williams, J.M. Goodfellow, J.M. Thornton, Buried waters and internal cavities in monomeric proteins, Protein Sci. 3 (1994) 1224–1235.

[22] O. Carugo, Structure and function of water molecules buried in the protein core, Curr. Protein Pept. Sci. 16 (2015) 259–265.

[23] A.T. Garcia-Sosa, S. Firth-Clark, R.L. Mancera, Including tightly-bound water molecules in de novo drug design. Exemplification through the in silico generation of poly (ADP-ribose)polymerase ligands, J. Chem. Inf. Model. 45 (2005) 624–633.

[24] A.T. Garcia-Sosa, R.L. Mancera, The effect of a tightly bound water molecule on scaffold diversity in the computer-aided de novo ligand design of CDK2 inhibitors, J. Mol. Model. 12 (2006) 422–431.

[25] A.T. Garcia-Sosa, R.L. Mancera, Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex, Mol Inform 29 (2010) 589–600.

[26] A.T. Garcia-Sosa, R.L. Mancera, P.M. Dean, WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes, J. Mol. Model. 9 (2003) 172–182.

[27] D.G. Lloyd, A.T. Garcia-Sosa, I.L. Alberts, N.P. Todorov, R.L. Mancera, The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models, J. Comput. Aided Mol. Des. 18 (2004) 89–100.

[28] R.L. Mancera, Molecular modeling of hydration in drug design, Curr. Opin. Drug Discov. Dev 10 (2007) 275–280.

[29] B.C. Roberts, R.L. Mancera, Ligand-protein docking with water molecules, J. Chem. Inf. Model. 48 (2008) 397–408.

[30] S.E. Wong, F.C. Lightstone, Accounting for water molecules in drug design, Expet Opin. Drug Discov. 6 (2011) 65–74.

[31] R. Abel, T. Young, R. Farid, B.J. Berne, R.A. Friesner, Role of the active-site solvent in the thermodynamics of factor Xa ligand binding, J. Am. Chem. Soc. 130 (2008) 2817–2831.

[32] G.V. DeLucca, S. EricksonViitanen, P.Y.S. Lam, Cyclic HIV protease inhibitors capable of displacing the active site structural water molecule, Drug Discov. Today 2 (1997) 6–18.

[33] S.B.A. de Beer, N.P.E. Vermeulen, C. Oostenbrink, The role of water molecules in computational drug design, Curr. Top. Med. Chem. 10 (2010) 55–66.

[34] A.T. Garcia-Sosa, Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies, J. Chem. Inf. Model. 53 (2013) 1388–1405.

[35] C.N. Cavasotto, A.T. García-Sosa, Role of Water Molecules and Hydration proPerties in Modeling Ligand–Protein Interaction and Drug Design, in Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications, CRC Press, 2015, pp. 393–410.

[36] W.R. Pitt, J. Murrayrust, J.M. Goodfellow, Aquarius2-Knowledge-Based modeling of solvent sites around proteins, J. Comput. Chem. 14 (1993) 1007–1018.

[37] W.R. Pitt, J.M. Goodfellow, Modeling of solvent positions around polar groups in proteins, Protein Eng. 4 (1991) 531–537.

[38] A. Vedani, D.W. Huhta, An algorithm for the systematic solvation of proteins based on the directionality of hydrogen-bonds, J. Am. Chem. Soc. 113 (1991) 5860–5862.

[39] G. Rossato, B. Ernst, A. Vedani, M. Smiesko, AcquaAlta: a directional approach to the solvation of ligand-protein complexes, J. Chem. Inf. Model. 51 (2011) 1867–1881.

[40] G.A. Ross, G.M. Morris, P.C. Biggin, Rapid and accurate prediction and scoring of water molecules in protein binding sites, PLoS One 7 (2012).

[41] J.S. Mason, A. Bortolato, D.R. Weiss, F. Deflorian, B. Tehan, F.H. Marshall, High end GPCR design: crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks, Silico Pharmacol 1 (2013) 23.

[42] B.M. Pettitt, M. Karplus, The structure of water surrounding a peptide - a theoretical approach, Chem. Phys. Lett. 136 (1987) 383–386.

[43] P.J. Rossky, M. Karplus, Solvation - molecular-dynamics study of a dipeptide in water, J. Am. Chem. Soc. 101 (1979) 1913–1937.

[44] W.F. Vangunsteren, H.J.C. Berendsen, J. Hermans, W.G.J. Hol, J.P.M. Postma, Computer-simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data, P Natl Acad Sci-Biol 80 (1983) 4315–4319.

[45] G. Copie, F. Cleri, R. Blossey, M.F. Lensink, On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes, Sci. Rep. 6 (2016) 38259.

[46] D.W. Borhani, D.E. Shaw, The future of molecular dynamics simulations in drug discovery, J. Comput. Aided Mol. Des. 26 (2012) 15–26.

[47] Y.B. Shan, E.T. Kim, M.P. Eastwood, R.O. Dror, M.A. Seeliger, D.E. Shaw, How does a drug molecule find its target binding site? J. Am. Chem. Soc. 133 (2011) 9181–9183.

[48] R.O. Dror, A.C. Pan, D.H. Arlow, D.W. Borhani, P. Maragakis, Y.B. Shan, H.F. Xu, D.E. Shaw, Pathway and mechanism of drug binding to G-protein-coupled receptors, P Natl Acad Sci USA 108 (2011) 13118–13123.

[49] R.O. Dror, H.F. Green, C. Valant, D.W. Borhani, J.R. Valcourt, A.C. Pan, D.H. Arlow, M. Canals, J.R. Lane, R. Rahmani, J.B. Baell, P.M. Sexton, A. Christopoulos, D.E. Shaw, Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs, Nature 503 (2013) 295.

[50] Y.B. Shan, K. Gnanasambandan, D. Ungureanu, E.T. Kim, H. Hammaren, K. Yamashita, O. Silvennoinen, D.E. Shaw, S.R. Hubbard, Molecular basis for pseudokinase-dependent autoinhibition of JAK2 tyrosine kinase, Nat. Struct. Mol. Biol. 21 (2014) 579–584.

[51] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers, Software 1 (2015) 19–25.

[52] R. Salomon-Ferrer, A.W. Gotz, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald, J. Chem. Theor. Comput. 9 (2013) 3878–3888.

[53] A.W. Gotz, M.J. Williamson, D. Xu, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born, J. Chem. Theor. Comput. 8 (2012) 1542–1555.

[54] A.C. Pan, T.M. Weinreich, S. Piana, D.E. Shaw, Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems, J. Chem. Theor. Comput. 12 (2016) 1360–1367.

[55] D.E. Shaw, M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Ierardi, I. Kolossvary, J.L. Klepeis, T. Layman, C. Mcleavey, M.A. Moraes, R. Mueller, E.C. Priest, Y.B. Shan, J. Spengler, M. Theobald, B. Towles, S.C. Wang, Anton, a special-purpose machine for molecular dynamics simulation, Commun. ACM 51 (2008) 91–97.

[56] J.J. Virtanen, L. Makowski, T.R. Sosnick, K.F. Freed, Modeling the hydration layer around proteins: HyPred, Biophys. J. 99 (2010) 1611–1619.

[57] V. Lounnas, B.M. Pettitt, G.N. Phillips, A global-model of the protein-solvent interface, Biophys. J. 66 (1994) 601–614.

[58] V. Makarov, B.M. Pettitt, M. Feig, Solvation and hydration of proteins and mucleic acids: a theoretical view of simulation and experiment, Accounts Chem. Res. 35 (2002) 376–384.

[59] V.A. Makarov, B.K. Andrews, B.M. Pettitt, Reconstructing the protein-water interface, Biopolymers 45 (1998) 469–478.

[60] V.A. Makarov, B.K. Andrews, P.E. Smith, B.M. Pettitt, Residence times of water molecules in the hydration sites of myoglobin, Biophys. J. 79 (2000) 2966–2974.

[61] B.M. Pettitt, V.A. Makarov, B.K. Andrews, Protein hydration density: theory, simulations and crystallography, Curr. Opin. Struct. Biol. 8 (1998) 218–221.

[62] B.P. Schoenborn, A. Garcia, R. Knott, Hydration in protein crystallography, Prog. Biophys. Mol. Biol. 64 (1995) 105–119.

[63] M.F. Lensink, I.H. Moal, P.A. Bates, P.L. Kastritis, A.S.J. Melquiond, E. Karaca, C. Schmitz, M. van Dijk, A.M.J.J. Bonvin, M. Eisenstein, B. Jimenez-Garcia, S. Grosdidier, A. Solernou, L. Perez-Cano, C. Pallara, J. Fernandez-Recio, J.Q. Xu, P. Muthu, K.P. Kilambi, J.J. Gray, S. Grudinin, G. Derevyanko, J.C. Mitchell, J. Wieting, E. Kanamori, Y. Tsuchiya, Y. Murakami, J. Sarmiento, D.M. Standley, M. Shirota, K. Kinoshita, H. Nakamura, M. Chavent, D.W. Ritchie, H. Park, J. Ko, H. Lee, C. Seok, Y. Shen, D. Kozakov, S. Vajda, P.J. Kundrotas, I.A. Vakser, B.G. Pierce, H. Hwang, T. Vreven, Z.P. Weng, I. Buch, E. Farkash, H.J. Wolfson, M. Zacharias, S.B. Qin, H.X. Zhou, S.Y. Huang, X.Q. Zou, J.A. Wojdyla, C. Kleanthous, S.J. Wodak, Blind prediction of interfacial water positions in CAPRI, Proteins 82 (2014) 620–632.

[64] R.H. Henchman, J.A. McCammon, Structural and dynamic properties of water around acetylcholinesterase, Protein Sci. 11 (2002) 2080–2090.

[65] H.C. Huang, D. Jupiter, M. Qiu, J.M. Briggs, V. VanBuren, Cluster analysis of hydration waters around the active sites of bacterial alanine racemase using a 2-ns MD simulation, Biopolymers 89 (2008) 210–219.

[66] M.S. Madhusudhan, S. Vishveshwara, Deducing hydration sites of a protein from molecular dynamics simulations, J. Biomol. Struct. Dyn. 19 (2001) 105–114.

[67] R. Abel, T. Young, R. Farid, B.J. Berne, R.A. Friesner, Role of the active-site solvent in the thermodynamics of factor Xa ligand binding, J. Am. Chem. Soc. 130 (2008) 2817–2831.

[68] T. Lazaridis, Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory, J. Phys. Chem. B 102 (1998) 3531–3541.

[69] T. Lazaridis, Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to simple fluids, J. Phys. Chem. B 102 (1998) 3542–3550.

[70] N. Jeszenoi, I. Horvath, M. Balint, D. van der Spoel, C. Hetenyi, Mobility-based prediction of hydration structures of protein surfaces, Bioinformatics 31 (2015) 1959–1965.

[71] N. Jeszenoi, M. Balint, I. Horvath, D. van der Spoel, C. Hetenyi, Exploration of interfacial hydration networks of target ligand complexes, J. Chem. Inf. Model. 56 (2016) 148–158.

[72] C. Vega, J.L.F. Abascal, M.M. Conde, J.L. Aragones, What ice can teach us about water interactions: a critical comparison of the performance of different water models, Faraday Discuss 141 (2009) 251–276.

[73] D. Nayar, C. Chakravarty, Sensitivity of local hydration behaviour and conformational preferences of peptides to choice of water model, Phys. Chem. Chem. Phys. 16 (2014) 10199–10213.

[74] P.S. Nerenberg, T. Head-Gordon, Optimizing protein-solvent force fields to reproduce intrinsic conformational preferences of model peptides, J. Chem. Theor. Comput. 7 (2011) 1220–1230.

[75] P. Florova, P. Sklenovsky, P. Banas, M. Otyepka, Explicit water models affect the specific solvation and dynamics of unfolded peptides while the conformational behavior and flexibility of folded peptides remain intact, J. Chem. Theor. Comput. 6 (2010) 3569–3579.

[76] D. Matthes, B.L. de Groot, Secondary structure propensities in peptide folding simulations: a systematic comparison of molecular mechanics interaction schemes, Biophys. J. 97 (2009) 599–608.

[77] R.B. Best, J. Mittal, Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse, J. Phys. Chem. B 114 (2010) 14916–14923.

[78] E. Fadda, R.J. Woods, On the role of water models in quantifying the binding free energy of highly conserved water molecules in proteins: the case of Concanavalin A, J. Chem. Theor. Comput. 7 (2011) 3391–3398.

[79] O. Rahaman, M. Kalimeri, M. Katava, A. Paciaroni, F. Sterpone, Configurational Disorder of Water Hydrogen-bond Network at the Protein Dynamical Transition, vol. 121, 2017, pp. 6792–6798.

[80] A. Einstein, Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, Ann. Phys. 322 (1905) 549–560.

[81] C.C. Miller, The Stokes-Einstein law for diffusion in solution, proceedings of the royal society of london. Series a, Containing Papers of a Mathematical and Physical Character 106 (1924) 724–749.

[82] R.H. Henchman, J.A. McCammon, Extracting hydration sites around proteins from explicit water simulations, J. Comput. Chem. 23 (2002) 861–869.

[83] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, D.E. Shaw, Improved side-chain torsion potentials for the Amber ff99SB protein force field, Proteins 78 (2010) 1950–1958.

[84] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935.

[85] B. Halle, Biomolecular cryocrystallography: structural changes during flash-cooling, P Natl Acad Sci USA 101 (2004) 4793–4798.

[86] H. Hope, Cryocrystallography of biological macromolecules - a generally applicable method, Acta Crystallogr. B 44 (1988) 22–26.

[87] J.W. Pflugrath, Practical macromolecular cryocrystallography, Acta Crystallographica Section F-Structural Biology Communications 71 (2015) 622–642.

[88] E.F. Garman, T.R. Schneider, Macromolecular cryocrystallography, J. Appl. Crystallogr. 30 (1997) 211–237.

[89] H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, T. Head-Gordon, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew, J. Chem. Phys. 120 (2004) 9665–9678.

[90] M.W. Mahoney, W.L. Jorgensen, A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions, J. Chem. Phys. 112 (2000) 8910–8922.

[91] C. Vega, E. Sanz, J.L.F. Abascal, The melting temperature of the most common models of water, J. Chem. Phys. 122 (2005).

[92] K. Harata, T. Akiba, Phase transition of triclinic hen egg-white lysozyme crystal associated with sodium binding, Acta Crystallogr. D 60 (2004) 630–637.

[93] J.R. Grigera, An effective pair potential for heavy water, J. Chem. Phys. 114 (2001) 8064–8067.

[94] H. Yamada, T. Nagae, N. Watanabe, High-pressure protein crystallography of hen egg-white lysozyme, Acta Crystallogr. D 71 (2015) 742–753.

[95] F.X. Prielmeier, E.W. Lang, R.J. Speedy, H.D. Ludemann, The pressure-dependence of self-diffusion in supercooled light and heavy-water, Ber Bunsen Phys Chem 92 (1988) 1111–1117.

[96] K.R. Harris, L.A. Woolf, Pressure and temperature-dependence of the self-diffusion coefficient of water and O-18 water, J. Chem. Soc. Faraday. Trans. 1 (76) (1980) 377–385.

[97] G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling, J. Chem. Phys. 126 (2007).

[98] T. Darden, D. York, L. Pedersen, Particle Mesh Ewald - an N.log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.

[99] S. Nose, M.L. Klein, Constant pressure molecular-dynamics for molecular systems, Mol. Phys. 50 (1983) 1055–1076.

[100] M. Parrinello, A. Rahman, Polymorphic transitions in single-crystals - a new molecular-dynamics method, J. Appl. Phys. 52 (1981) 7182–7190.

[101] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, Bioinformatics 29 (2013) 845–854.

[102] M. Bastian, S. Heymann, M. Jacomy, Gephi: an Open Source Software for Exploring and Manipulating Networks, 2009. San Jose, CA, USA.

D19

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

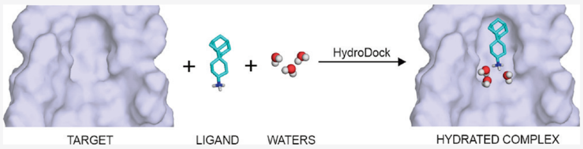
pubs.acs.org/jcim Article

Determination of Ligand Binding Modes in Hydrated Viral Ion Channels to Foster Drug Design and Repositioning

Balázs Zoltán Zsidó, Rita Börzsei, Viktor Szél, and Csaba Hetényi*

Cite This: J. Chem. Inf. Model. 2021, 61, 4011−4022 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Target-based design and repositioning are mainstream strategies of drug discovery. Numerous drug design and repositioning projects have been launched to fight the ongoing COVID-19 pandemic. The resulting drug candidates have often failed due to the misprediction of their target-bound structures. The determination of water positions of such structures is particularly challenging due to the large number of possible drugs and the diversity of their hydration patterns. To answer this challenge and help correct predictions, we introduce a new protocol HydroDock, which can build hydrated drug−target complexes from scratch. HydroDock requires only the dry target and drug structures and produces their complexes with appropriately positioned water molecules. As a test application of the protocol, we built the structures of amantadine derivatives in complex with the influenza M2 transmembrane ion channel. The repositioning of amantadine derivatives from this influenza target to the SARS-CoV-2 envelope protein was also investigated. Excellent agreement was observed between experiments and the structures determined by HydroDock. The atomic resolution complex structures showed that water plays a similar role in the binding of amphipathic amantadine derivatives to transmembrane ion channels of both influenza A and SARS-CoV-2. While the hydrophobic regions of the channels capture the bulky hydrocarbon group of the ligand, the surrounding waters direct its orientation parallel with the axes of the channels via bridging interactions with the ionic ligand head. As HydroDock supplied otherwise undetermined structural details, it can be recommended to improve the reliability of future design and repositioning of antiviral drug candidates and many other ligands with an influence of water structure on their mechanism of action.

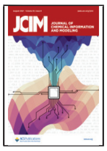
INTRODUCTION

The COVID-19 pandemic has generated a tsunami in target-based drug design[1] and repositioning.[2] Target-based design is a widely used approach[3−7] where the target structure serves as a reference point for fitting and selection of drug candidates. Repositioning is a cheap and fast strategy of drug discovery, as the pharmacological profile of known drugs is readily available with detailed information on their pharmacodynamics, pharmacokinetics, toxicity, interactions, and side effects. The clinical repositioning trials of a number of known drugs were launched in the past year[8−11] to test their applicability against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Although a few drugs were approved for clinical use, the repositioning trials have not led to real breakthroughs against SARS-CoV-2.

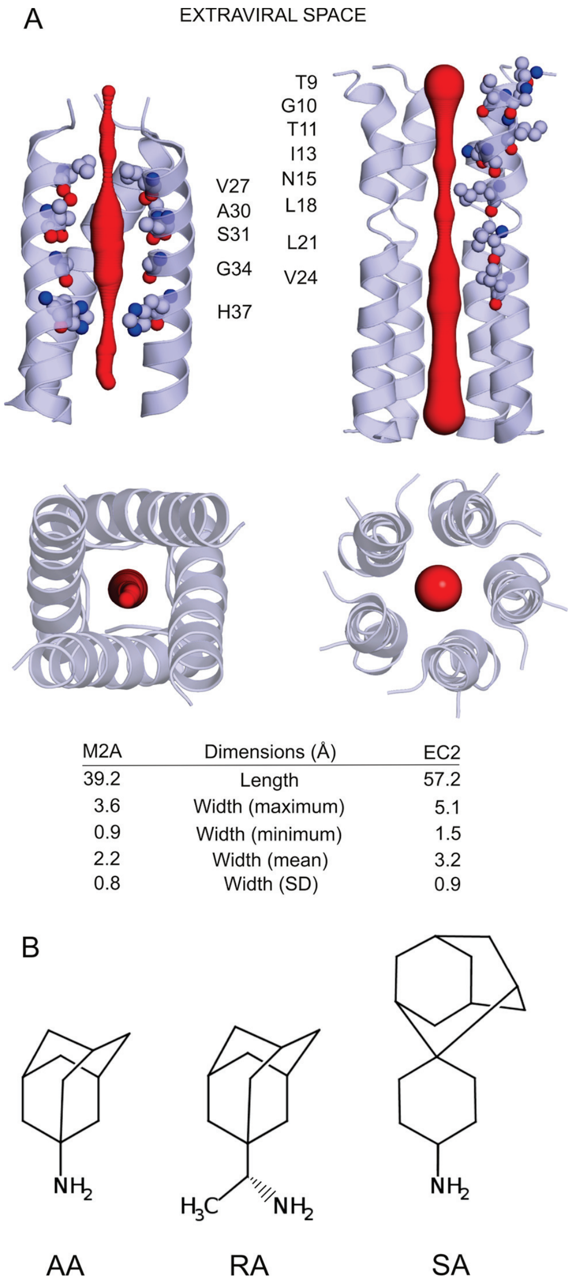
The failure of repositioning trials can be largely attributed to the structural differences between the old and new targets. For example, the structural dissimilarities between the active sites of proteases of HIV-1 and SARS-CoV-2 forecasted[12,13] the failure of recent repositioning trials[8,14] of HIV-1 protease inhibitors lopinavir and ritonavir to SARS-CoV-2. Such painful lessons highlight the necessity of a careful structure-based design and repositioning to reduce the number of failed clinical trials.

In the present study, we investigate the structural basis of repositioning of FDA-approved drugs amantadine (AA, Gocovri, Symmetrel) and its derivatives, rimantadine (RA, Flumadine) and spiroadamantyl amine (SA),[15−20] (Figure 1b) to the ion channel formed by the transmembrane domain of the SARS-CoV-2 envelope protein (EC2, Figure 1a) as a possible "new" target. These AA derivatives were shown to inhibit the cation conductance of the M2 transmembrane ion channel of influenza A virus (M2A, Figure 1a),[21] and the "old" target was also used as a reference in this study. AA was originally suggested[16] against SARS-CoV and showed various beneficial effects in patients infected by the SARS-CoV-2[18−20] as well. EC2 is homologous to the envelope protein of SARS-CoV[16] and also functions as a cation-selective ion channel like M2A, playing a role in virus budding, release, and host inflammation response.[15] The blocking of EC2 by AA derivatives or similar amphipathic molecules is a promising drug design strategy[22,23] even on a longer term due to the low mutagenicity of EC2 found in mutated SARS-CoV-2 lineages collected from patients in

Received: May 6, 2021
Published: July 27, 2021



https://doi.org/10.1021/acs.jcim.1c00488



M2A	Dimensions (Å)	EC2
39.2	Length	57.2
3.6	Width (maximum)	5.1
0.9	Width (minimum)	1.5
2.2	Width (mean)	3.2
0.8	Width (SD)	0.9

Figure 1. (A) M2A (left) and EC2 (right) ion channels shown as cartoon. The red cone represents the diameter of the ion channels. Interacting amino acids are labeled and shown as spheres in the side views at the top (a helix was deleted to show the interior of the channels). Top views from the extraviral space and lists of dimensions of the channels are shown at the bottom. (B) Lewis structures of the three AA derivatives investigated in the present study. Under physiological conditions, the amino group is protonated, resulting in a net charge of +1. *R*-rimantadine was used in the study, referred to as RA.

India.[24] Recently, the atomic resolution structure of EC2 (Figure 1a) was determined[15] using solid-state NMR, providing a starting point for target-based design. The same study demonstrated the binding of fluorinated AA to EC2 as well. The large pore[15] of EC2 is formed in a pentameric helical bundle stabilized by interhelical aromatic stacking interactions.

The pore size of EC2 is comparable to that formed by the tetrameric bundle in M2A[21] (Figure 1a), which captures the AA derivatives. The similar pore geometry of M2A and EC2 is just one structural factor if considering the repositioning of ligands between the two ion channels.

Their amino acid composition and water structure[25] are also key factors of ligand binding. The mediating role of water molecules was highlighted in the binding mechanism of AA derivatives to M2A.[21,25] Considering the above similarity between M2A and EC2, one may expect that understanding the role of water molecules will be important in the case of EC2 as well. The available EC2 structure[15] is an apo form without water and ligand molecules (Figure 1a), and therefore, it cannot supply any information on the possible mediating role of water molecules in ligand binding to EC2. Thus, an atomic resolution structure of the full complex with a bound ligand and water molecules (a hydrated holo structure) is necessary to foster correct repositioning and design to EC2.

As the full complex has not been solved at atomic resolution, we have to calculate the binding of the AA derivatives and the water structure from scratch, which is a challenging task for current methods.[25] To answer this challenge, we introduce a new protocol that will supply the water structure of the EC2 channel and also adopt docking and molecular dynamics steps to produce the representative binding modes of AA derivatives. The protocol will be tested on the old M2A target with available experimental complex structures as references and will be transferred to the new EC2 target. In this way, we will explore the role of water in binding of the AA derivatives and produce their key binding modes on the new EC2 target, supplying the necessary atomic resolution structures for repositioning and design.

■ METHODS

Input Structures. The atomic coordinates of M2A complexed with AA (6BKK), RA (6BKL), and SA (6BMZ)[21] and the ligand-free structure of M2A (3LBW)[26] were acquired from the Protein Databank (PDB). A, B, C, and D chains and their corresponding ligand (except for the apo structure) and water molecules were used for protocol development and validation purposes (Sections "The Effect of Interfacial Water Molecules on Ligand Docking to the Influenza A M2A Channel" and "Construction of the Ligand-Bound, Hydrated Influenza A M2A Channel Structures from Scratch"). The EC2 NMR structure (first model of the 20) from ref 15 (7K3G) was used in Section "Ligand Binding Modes and the Water Structure in the EC2 Channel of SARS-CoV-2" to create the hydration structure and ligand binding modes from scratch.

Ligand Preparation. Ligands were built in Maestro.[27] The raw structures were energy-minimized using a semiempirical quantum chemistry program package, MOPAC[28] with PM7 parametrization.[29] The gradient norm was set to 0.001. The energy-minimized structures were submitted to force calculations; the force constant matrices were positive definite. Restrained electrostatic potential (RESP) charges were calculated with RED-vIII.52[30] after geometry optimization by GAMESS.[31] Acpype[32] and antechamber[32,33] were used to assign bound parameters and atom types for topology of ligands.

Target Preparation. The N-terminal ends of the ion channels were capped with acetyl groups and the C-terminal ends with imino-methyl groups using Maestro[27] and were subjected to

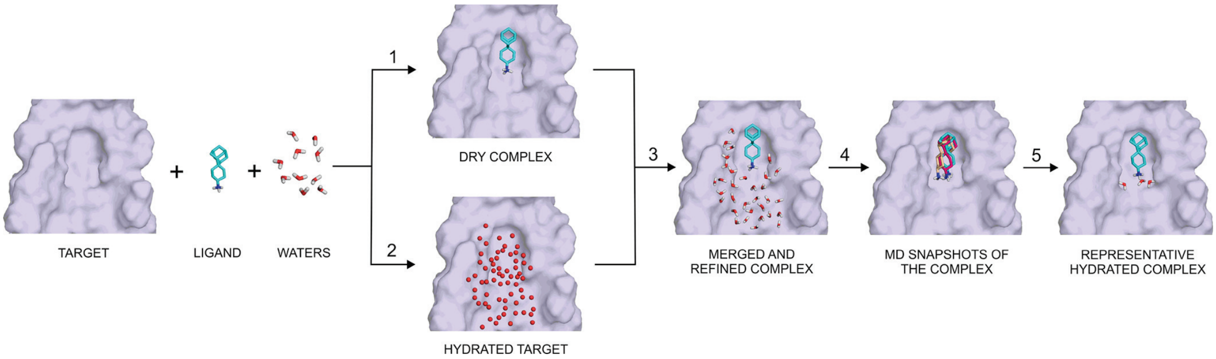

TARGET
LIGAND
WATERS
1
2
DRY COMPLEX
HYDRATED TARGET
3
MERGED AND REFINED COMPLEX
4
MD SNAPSHOTS OF THE COMPLEX
5
REPRESENTATIVE HYDRATED COMPLEX

Figure 2. Assembly of the hydrated complex of the M2A channel (target, surface) and SA (ligand, sticks) from scratch using the HydroDock protocol. The numbering of steps of HydroDock follows the explanation in the main text. After the first step, nonminimized water positions from MobyWat[40,41] are shown as red spheres; otherwise, sticks representation is used for hydrogenated and minimized waters. During the third step, some of the water positions are replaced by the ligand. For clarity, only a few MD snapshots of ligand binding modes are shown after the fourth step. Coordinate files of all snapshots are accessible in the Supporting Information.

energy minimization in the merging step (Step 3). Hydrogen atoms and Gasteiger−Marsili partial charges[34] were added to the targets with AutoDock Tools.[35] After ligand and target preparation, the dry target and respective ligands were used as starting points of HydroDock (next section).

HydroDock. HydroDock is a new protocol shortly featured in Section 2 of Results and Discussion. The steps of HydroDock are numbered in Figure 2 and referred to in the following detailed descriptions using the same numbering as in Results and Discussion.

Step 1. Dry Docking. Blind docking was performed as described before[36] for both targets M2A and EC2 (Box B, Table S5). During blind docking, the docking box covered the whole surface of the target. Focused docking was also used for EC2 when the box only covered the upper half of the protein (Box A, Table S5). The unliganded M2A and EC2 structures were used as targets of the blind and focused docking runs. No explicit water molecules were adopted from the PDB structures. The target was treated as a rigid body except that the flexibility of the N15 amino acid side chains was allowed on all helices of EC2 to allow the entrance of the ligand toward the intraviral regions (Table S5). AutoGrid 4.2[35] was used for grid map calculations. Grid boxes were generated around the entire M2A target. The grid boxes were centered on the target, and 70 (M2A) and 90 (EC2) grid points along all axes were set with 0.503 Å grid spacing (0.375 Å in Box A). The resulting docking box covered the entire M2A and EC2 in the case of blind docking and allowed the entrance of the ligands from both extra- and intraviral regions. To avoid artefacts and allow ligand entrance only from the extraviral space (Figure 1a), the docking box was reduced to only cover the upper half of EC2 (Box A, Table S5).

Molecular docking calculations were performed by AutoDock 4.2.[35] Hydrogen atoms and Gasteiger−Marsili[34] partial charges were added to the ligands with an OpenBabel[37] program package. All chemically relevant torsions of the ligands were enabled. One hundred blind docking runs were performed. The Lamarckian genetic algorithm and the pseudo-Solis and Wets local search with a maximum number of 300 iterations and 25 million energy evaluations and 150 population size were applied as in refs 38 and 39. The generated 100 ligand binding modes were clustered and ranked (see Section "Evaluation Criteria" for details) based on their calculated free energy of binding values and structural similarity. Representative ligand structures of each rank in complex with their dry target structures were used as dry complexes. Due to the symmetry of both M2A and EC2, from among identical, symmetry-related rank representatives, the one with the lowest calculated binding free energy was selected and forwarded to the next steps of HydroDock.

In the case of M2A, a total of six representatives were found, one−one for all three AA derivatives on both holo and apo target forms (Table 1). In the case of EC2, five (AA1, ..., AA5, Table S5), two, and one representatives of AA, RA, and SA were found (eight in total) and forwarded to Step 3.

Table 1. Comparison of Computationally Docked and Experimental Binding Positions of Ligands AA, RA, and SA to a Dry M2A Target

ligand	M2A conformation	RMSD (Å)	rank[a]
AA	holo	3.3	1/1
AA	apo	3.7	1/1
RA	holo	3.8	1/2
RA	apo	3.6	1/1
SA	holo	4.8	1/3
SA	apo	2.9	1/1
mean		3.7	
SD		0.7	

[a] Serial number of rank/count of all ranks.

Step 2. Building the Water Structure of the Inner Surface of the Target Channels. The water structure of the inner surface of the target channels was built using MobyWat,[40,41] which requires an MD trajectory of a target in explicit water as an input. The MD-based evaluation of MobyWat allows consideration of all solute−water and water−water interactions and results in high success rates if compared with experimental structures.[40,41]

Generation of MD Trajectories. The dry M2A (6BKK) and EC2 (7K3G) targets were energy-minimized by steepest descent and conjugate gradient algorithms as in Step 3 of HydroDock to prepare them for the 1 ns-long MD simulations. The simulation box was filled with explicit TIP3P[42] water molecules, and counterions (sodium or chloride) were added to neutralize the system. Exit tolerance levels were set to 10^3 and 10 kJ·mol^{-1}·nm^{-1}, while maximum step sizes were set to 0.5 and 0.05 nm for the steepest descent and conjugate gradient steps, respectively.

https://doi.org/10.1021/acs.jcim.1c00488
J. Chem. Inf. Model. 2021, 61, 4011−4022

Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ·mol^{-1}·nm^{-2}. Calculations were performed with programs of the GROMACS[43] software package using the AMBER99SB-ILDN[44] force field. After energy minimization, 1 ns-long NPT MD simulation was carried out with a time step of 2 fs. For temperature coupling, the velocity rescale[45] algorithm was used. The solute and solvent were coupled separately with a reference temperature of 300 K and a coupling time constant of 0.1 ps. Pressure was coupled by the Parrinello−Rahman algorithm[46,47] and a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1}, and reference pressure of 1 bar. Particle mesh-Ewald summation[48] was used for long-range electrostatics. Van der Waals and Coulomb interactions had a cutoff at 11 Å. Coordinates were saved at regular time intervals of 1 ps, yielding 1×10^3 frames. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ·mol^{-1}·nm^{-2}. Periodic boundary conditions were treated before analysis to make the solute whole and recover hydrated solute structures centered in the box. Each frame was fit to the original protein crystal structure using Cα atoms. The final trajectory including all atomic coordinates of all frames was converted to portable XDR binary files equipped with name extension xtc.

MobyWat Calculations. From the MD trajectory, surface water positions were calculated with Mobywat's[40] all-inclusive identity-based (IDa) prediction algorithm. The maximum distance from the target (d_{max}), prediction, and clustering tolerances were set to 5, 2.5, and 1 Å, respectively. The MobyWat algorithm was described earlier.[40,41] Briefly, candidate water molecules for all frames are selected based on a desired distance limit (d_{max}) from the target, and then an occupancy list is constructed containing every different water IDs on every line and the respective number of occurrences as candidates among all frames. Clustering is applied to all rows (all different water IDs) of the occupancy list using the ctol parameter to define the distance between elements of the same cluster. The largest cluster is selected from all to give the first predicted water molecule by averaging the spatial coordinates of included molecules. In the further steps, clusters are selected in a descending order size-wise and checked if their distance is larger than the prediction tolerance from previously predicted water positions. After the above clustering, a list of water positions (prediction list) was produced as the O atom coordinates covering the surface of the EC2 (7K3G) and M2A (6BKK) channels. The hydrogens were added to the predicted water O atoms in a later step (Step 3 of HydroDock).

In the case of M2A, the predicted water oxygen positions were compared to the reference water molecules in the PDB structure 6BKK using the validation mode of MobyWat. The above settings were used with a match tolerance of 1.5 Å.

Step 3. Merging and Refinement. Merging. The outcomes of Steps 1 and 2 were combined to build the raw complex structures, that is, the hydrated, ligand-bound targets. For this, the complexes were placed in a common coordinate system by alignment of the target structure of the dry complex from Step 1 and the hydrated target structure from Step 2 using PyMol.[49] After alignment, a raw complex still contains all surface water molecules predicted by MobyWat. However, after the placement of the dry docked ligand structure into the fully hydrated target, some water molecules overlap with the ligand. The overlapping water molecules were removed by the editing mode of MobyWat,[40] and only interfacial water molecules were retained. The merged structures (see Step 1. Dry Docking) of the eight EC2 complexes (Table S5) and six M2A complexes (Table 1) were then subjected to robust refinement.

Soft Refinement (Not Part of HydroDock and Used during Protocol Development) (Figure S1). The interfacial crystallographic water oxygen atoms within a d_{max} of 5.0 Å distance limit from both the ligand and the target were kept, as they bridge between the ligand and the amino acid residues of the protein; other waters were removed. The structure of the M2A channel with the water O atoms was placed in a dodecahedral box using a distance criterion of 1 nm between the solute and the box. Void spaces of the box were filled by explicit TIP3P water molecules by GROMACS.[43] Hydrogen atoms were added to water oxygen and solute atoms by the GROMACS program pdb2gmx. The system was neutralized by counterions. A steepest descent (steepest descent1) optimization was carried out,[40] with convergence threshold set to 10^3 kJ·mol^{-1}·nm^{-1} followed by a conjugate gradient (conjugate gradient1) calculation, where the convergence threshold was set 10 kJ·mol^{-1}·nm^{-1}. Position restraints at a force constant of 10^3 kJ·mol^{-1}·nm^{-2} were applied on all heavy atoms in both steps. An AMBER99SB-ILDN[44] force field was used for the calculations. The steepest descent and conjugate gradient minimization steps were carried out once again (steepest descent2, conjugate gradient2), with the same settings[40] as in steepest descent1 and conjugate gradient1, with the exception that only backbone Cα atoms were position restrained.

Robust Refinement Was Adopted as an Appropriate Protocol of HydroDock Based on the Good Docking Results (Table 3). Robust refinement has only one difference when compared to soft refinement; the steepest descent1+conjugate gradient1 step is not immediately followed by the steepest descent2+conjugate gradient2 steps, but first, a 100 ps-long MD simulation (md) is carried out (steepest descent1+conjugate gradient1+md+steepest descent2+conjugate gradient2). In the MD simulation, only backbone Cα atoms were position-restrained. Notably, in a general application of HydroDock for systems with large flexibility on the target backbones, the use of a membrane model would be advisable instead of position restraining of the backbone. For temperature coupling, the velocity rescale algorithm was used. The solute and solvent were coupled separately with a reference temperature of 300 K and a coupling time constant of 0.1 ps. Pressure was coupled with the Parrinello−Rahman algorithm with a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1}, and reference pressure of 1 bar. Particle mesh-Ewald summation was used for long-range electrostatics. Van der Waals and Coulomb interactions had a cutoff at 11 Å. Robust refinement resulted in the correct position of the experimental water molecules of M2A, with the right orientation of H atoms that led to the formulation of two water networks. Based on the success, robust refinement was adopted in Step 3 of HydroDock after merging.

Wet Docking (Not Part of HydroDock and Used during Protocol Development) to Choose the Sufficient Refinement Protocol and Validate It. In wet docking, every detail was set as in dry docking (Step 1 of HydroDock) except that refined water molecules were included. When compared to the experimental ligand positions, the Gasteiger−Marsili partial charges on the atoms of the water molecules yielded incorrect results (Figure S2). Thus, partial charges of the TIP3P explicit water model were used on all water molecules instead.

Step 4. Generating MD Snapshots of the Target−Ligand Complex. The MD simulations of the merged and refined complexes were carried out with the same settings described in

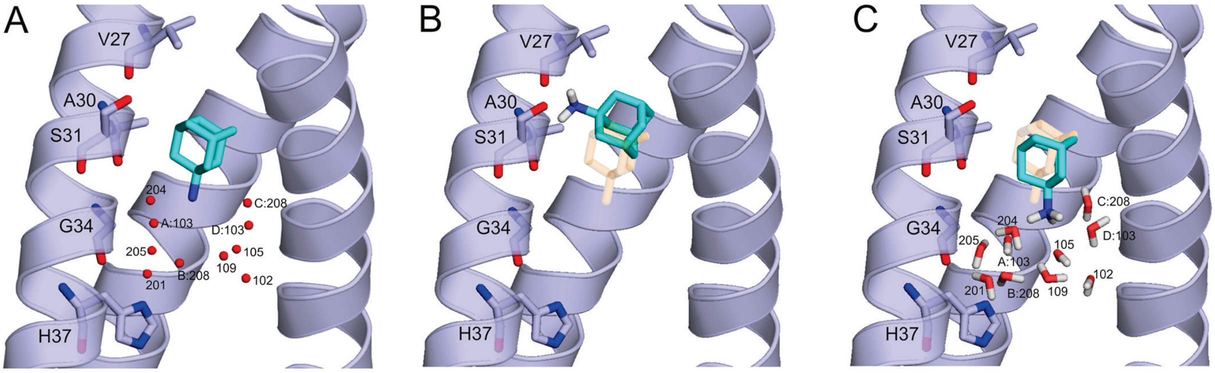


Figure 3. Complex of AA (sticks with teal carbon) bound to the M2A channel (cartoon and sticks with gray carbon, a frontal helix turned off for clarity). (A) Experimental binding mode in the PDB structure 6bkk with three H-bonds formed between the ligand protonated amino N and water O atoms. Water molecules are represented as red spheres and labeled by their chain IDs and/or residue numbers. (B) Result of "dry" docking of AA shows a shift of the positive protonated amino group of AA. Instead of the missing water molecules, interactions with the partially negative backbone carbonyl groups of V27 and A30 were formed. (C) Result of "wet" docking of AA is in a good agreement with the experimental binding mode of AA shown in panel A. The minimized water molecules are shown as thick lines and labeled according to the residue numbering in the PDB structure 6bkk. The crystallographic ligand binding position in (A) is also shown in (B) and (C) with transparent orange sticks for comparability.

the minimization procedure (robust refinement). The simulations were performed as listed in Table S4 and for 100 ns in the cases of M2A and EC2, respectively. Only the Cα atoms of the proteins were restrained. The movements of the amino acid side chains, the ligand, and the solvent were allowed. The refined hydration structure was kept in the MD simulations; the rest of the simulation box was filled with water molecules by GROMACS. Complex snapshots were aligned by a GROMACS tool trjconv using their target Cα atoms, and the bound ligand snapshots were separately generated as individual files from the MD trajectory file by 0.1 ns steps (conformation pool).

Step 5. The Selection of the Representative Ligand Binding Modes from the MD Trajectory File. An average ligand conformation was calculated from the conformation pool using a shell script provided in the Supporting Information file. RMSD values between the individual ligand pool structures and the average ligand pool structure were calculated according to eq 1, where the average pool conformation was used instead as a reference **C** in this case. A pool structure with the lowest RMSD value was selected as the representative ligand binding mode from the MD trajectory. A representative binding mode of the ligand is the suggested final binding mode to the target (M2A, EC2). Distinct binding modes produced by dry docking (Step 1) usually result in more than one representative structure after HydroDock.

Evaluation Criteria. Standard criteria[50−54] were applied to evaluate the results of dry and wet docking and HydroDock. In all cases, the structural match of the calculated (docked or HydroDock representative, **D** in eq 1) binding mode to the crystallographic reference (**C**) was expressed as a root-mean-square deviation (RMSD) value according to eq 1[51]

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |\mathbf{D}_n - \mathbf{C}_n|^2} \tag{1}$$

In eq 1, N is the number of ligand heavy atoms, **C** is the space vector of the nth heavy atom of the crystallographic reference ligand molecule, and **D** is the space vector of the nth heavy atom of the calculated ligand conformation. Overlapping ligand conformations resulted by 120° turns around the trigonal vertical axis were considered identical during RMSD calculations.

The ranking order was also shown in the cases of dry and wet docking trials. The docked ligand conformations were structurally clustered and ranked according to their AutoDock 4.2 binding free energy values, and the serial numbers of ranks are listed in Results and Discussion. During this procedure, the ligand structure with the lowest calculated free energy of binding was selected, and the neighboring docked ligand structures within 2 Å[38] were collected in the rank; then, a new rank is opened starting with an unused structure of the lowest calculated free energy of binding from the remaining structures, etc. until all 100 ligand structures were collected into ranks.[40] Ranks with a low serial number indicate an energetically favorable binding conformation. Note that in the case of HydroDock, representative binding modes were selected (Step 5) without the need of further ranking.

Calculation of Interaction Energy Values of SA-EC2 Complexes. The Lennard-Jones interaction energy (E_{LJ}) was calculated between the target and ligand molecules according to eq 2

$$E_{\mathrm{IJ}} = \sum_{\mathrm{ij}}^{N_{\mathrm{T}}N_{\mathrm{L}}} \left[\frac{A_{\mathrm{ij}}}{r_{\mathrm{ij}}^{12}} - \frac{B_{\mathrm{ij}}}{r_{\mathrm{ij}}^{6}} \right]$$

$$A_{\mathrm{ij}} = \varepsilon_{\mathrm{ij}} R_{\mathrm{ij}}^{12};\ B_{\mathrm{ij}} = 2\varepsilon_{\mathrm{ij}} R_{\mathrm{ij}}^{6};\ R_{\mathrm{ij}} = R_{\mathrm{i}} + R_{\mathrm{j}};\ \varepsilon_{\mathrm{ij}} = \sqrt{\varepsilon_{\mathrm{i}}\varepsilon_{\mathrm{j}}} \tag{2}$$

In eq 2, ε_{i} and ε_{j} are the potential well depths in the equilibrium distance of atom pairs of identical types; $\varepsilon_{\mathrm{ij}}$ is the potential well depth in equilibrium between the ith (ligand) and jth (target) atoms; R_{ij} is the internuclear distance at equilibrium between ith (ligand) and jth (target) atoms; R_{i} and R_{j} are half equilibrium distances between ii and jj atom pairs of identical types, respectively; r_{ij} is the actual distance between the ith (ligand) and jth (target) atoms; N_{T} is the number of target atoms; and N_{L} is the number of ligand atoms. The Amber 2012 force field parameters were used.[56] The calculations were performed for dry and hydrated targets as well. In the case of the hydrated target, explicit water molecules were considered as part of the target.

RESULTS AND DISCUSSION

(1) The Effect of Interfacial Water Molecules on Ligand Docking to the Influenza A M2A Channel. Water molecules play a key role[21] in binding AA and its derivatives to the influenza A M2A channel. For example, water (w) molecules A:w103, B:w204 (at A30) and B:w201, and C:w205 (at G34) form bridges between the positive protonated amino group of AA and the carbonyl oxygens of the amino acids (Figure 3a, the numbering of the PDB structure 6bkk is used). Together with other water molecules at H37, a static H-bonding network of 10 water molecules is formed, filling the channel cavity below AA (Figure 3a). Incorporation of such water molecules in docking calculations can be essential[57−60] to obtain precise results.

To check this assumption, a systematic investigation of computational docking of all three ligands (AA, RA, and SA) was performed to the M2A channel using different approaches of handling interfacial water molecules. Targeting the dry M2A channel without any surface water molecules (Table 1) is the simplest approach and provides a basis for comparisons throughout this study. An average of 3.7 ± 0.7 Å root-mean-square deviation (RMSD) was calculated between the docked and crystallographic ligand conformations with the latter ones used as references. This value is above the RMSD of 1.5−2.0 Å considered acceptable in the literature[50−54] and indicates that the dry M2A channel may not be an appropriate target for docking. The dry M2A channels in holo (ligand-bound) conformations did not yield significantly better results than the apo ones as docking targets. This follows from the high identity between the holo and apo target structures with an average superposition RMSD of 0.3 Å ± 0.1 (Table S2). Thus, there is no considerable induced fit during ligand binding to the M2A channel, and the rigidity (Methods) of the target structure did not affect the result in these cases. In the docked structure (AA-Holo in Table 1), the adamantyl group of AA was close to the crystallographic position (Figure 3b). However, the lack of the abovementioned (Figure 3a)[21] bridging water molecules resulted in a miscoordination of the protonated amino group to the carbonyl oxygen of V27 and the hydroxyl group of S31 (Figure 3b) and the large RMSD values of Table 1.

In the wet docking calculations, a set of functional water positions (Table 2 and Figure 3c) of the crystal structures was used together with the M2A channel as a target. As the coordinates of water hydrogen atoms were not available, a theoretical refinement was necessary to add and optimize their positions. During the refinements, the ligand was kept in the holo structure to help in the correct arrangements of water hydrogen atoms in contact with the protonated amino group.

Two refinement protocols (a soft and a robust one) were investigated. During the soft protocol (soft refinement), simple energy minimization steps were applied (Methods) for the water hydrogen atoms while the positions of all heavy atoms (including water oxygen) were restrained in their crystallographic positions. The docking of AA to the ligand-free, S-refined target still did not result in an acceptable RMSD (2.7 Å), which can be attributed to the incorrectly positioned water hydrogen atoms (Figure S1). A closer inspection of the S-refined target structure showed that the incorrect positioning of water hydrogen atoms was a consequence of several close contacts (Table 2) in the original crystallographic water structure.[21] The close contacts were maintained by the position restraints during soft refinement, resulting in relatively small shifts from their crystallographic positions (Table 2), hindered reconstruction of the interfacial H-bonding network, and atomic positions preformed to interact with AA (Figure S1). As docking of AA to the wet M2A target with soft refinement did not improve the dry results (Table 1), a robust protocol was also tested (robust refinement) including a molecular dynamics step with no restraints on the atoms. Robust refinement appropriately shifted half of the water molecules of Table 2 (A:w103, D:w105, D:w109, B:w201, and B:w204) to 1 Å or a larger distance (Table S3) from their crystallographic positions. In this way, their erroneous close contacts were eliminated, and their hydrogen atoms were arranged into correct orientations, resulting in a perfect H-bonding network. Some experimenting with the partial charge system on water molecules showed that TIP3P[42] outperformed Gasteiger−Marsilli[34] partial charges (Figure S2). Robust refinement and TIP3P charges on water molecules yielded excellent docking results with an average RMSD of 1.2 ± 0.3 Å (Table 3) for all ligands. The low serial numbers/counts of the corresponding ranks indicate that the structural precision reflected by the low RMSD values was accompanied by the best

Table 2. Deviations of Refined Crystallographic and MobyWat-Predicted Water Positions Used in Wet Docking Calculations Measured from the Original Crystallographic Positions (PDB ID 6bkk) with Their Close Contacts Also Listed

water #[a]	close contact[b]	soft refinement (Å)	robust refinement (Å)	predicted (Å)[c]
A:w102		0.3	0.2	0.2
A:w103		0.6	1.6	0.9
D:w103		0.5	0.5	0.9
D:w105	D:w109	0.7	2.1	0.6
D:w109	C:w208, D:w105	0.9	1.0	2.1
B:w201	B:G34	1.0	1.0	0.6
B:w204		0.8	1.6	0.7
C:w205		0.4	0.9	0.5
B:w208		0.8	0.8	1.0
C:w208	D:w109	0.9	0.2	0.3

[a]The numbering of PDB structure 6bkk is used (see Table S1 for details of selection of reference structures). [b]Close contacts of the crystallographic structure were listed if the distance between the oxygen atom of the actual water molecule and a heavy atom of a neighboring residue or the oxygen of the neighboring water molecule was below 2.75 Å. [c]Crystallographic water positions of PDB structure 6bkk were used as reference; see also Table S1 for details on selection of reference crystallographic structures.

Table 3. Comparison of Computationally Docked and Experimental Binding Positions of Ligands AA, RA, and SA to the M2A Target Covered by Crystallographic Water Positions Subjected to a Robust Refinement Protocol and Equipped with Partial Charges of the TIP3P Explicit Water Model

ligand	M2A conformation	RMSD (Å)	rank[a]
AA	holo	1.2	1/1
AA	apo	1.0	1/1
RA	holo	1.0	1/2
SA	holo	1.7	1/1
mean (holo)		1.2	
SD (holo)		0.3	

[a]Serial number of the rank/count of all ranks.

calculated binding free energies (or a single, homogeneous rank was produced). In the case of AA, docking to the wet, apo M2A channel structure was also performed after robust refinement. Similar to the holo results, an excellent RMSD of 1.0 Å was obtained (Figure 3c).

The results of Table 3 showed that appropriately placed and oriented water molecules are keys to precise docking results if compared with the insufficient outcomes of dry docking (Table 1). It was also found (Table 2) that the availability of crystallographic water positions alone cannot guarantee the success for two reasons. (1) Often, only oxygen positions are supplied, and water orientations are obviously not assigned due to the lack of hydrogen atoms. (2) There are also other limitations[41,61−69] of assignation of the crystallographic density map, resulting in missing or too many water molecules (overfitting). Such problems often result in crystallization artefacts[67] and close contacts similar to those listed in Table 2. Thus, a robust theoretical refinement of experimental water structure is necessary in general and for correct calculation of complexes of all three ligands with the M2A channel in the present case.

(2) Construction of the Ligand-Bound, Hydrated Influenza A M2A Channel Structures from Scratch. In agreement with other studies,[25] the results of the previous section showed that docking calculations are very sensitive to even small errors in the water structure. In the previous examples (Table 2), a robust refinement of the measured water positions was necessary to achieve good docking results. In a real drug screening project,[36,55] experiments cannot supply interfacial water positions and holo structures for all possible ligand molecules designed for the target binding pocket, and only an apo target structure is available for the docking calculations. Thus, only atomic coordinates of the individual components (ligand, target, and water) can be used for the construction work. It is a real challenge to bring all these partners together into a hydrated complex structure due to the difficulties of correct positioning of interfacial water molecules.[25]

To address this challenge, we introduce HydroDock, a hybrid protocol that supplies the hydrated complex structure from scratch. HydroDock is composed of five steps (Figure 2 and Methods) and was tested on the M2A target and its ligands (Figure 1). Step 1 involved a fast docking calculation with results described in Table 1. In Step 2, the water structure of the surface of the target was built by MobyWat[40,41] with high precision. MobyWat is a molecular dynamics (MD)-based method that can predict solute−water and water−water interactions as well. In the present case, the inner surface of the M2A target was completely hydrated and the calculated water positions were compared to the crystallographic reference ones as listed in Table 2. Nine out of ten water molecules were successfully predicted at a match threshold of 1.0 Å (see also Figure S3). The predicted hydration structure was a priori close contact-free and equipped with hydrogen atoms, which is necessary for correct docking calculations (Section "The Effect of Interfacial Water Molecules on Ligand Docking to the Influenza A M2A Channel" and Table 3). In Step 3, the results of the first two steps were merged into one structure and surface water molecules overlapping with the docked ligand were eliminated using the Editing mode[40] of MobyWat. In Step 4, the hydrated M2A−ligand complexes were subjected to molecular dynamics (MD) in a simulation box filled with explicit water molecules to generate a pool of several hundreds (N_{pool} in Table S4) of member conformations. Step 5 of HydroDock produces a representative complex conformation statistically selected from the pool (see Methods for the details of all steps).

The matches of the representative ligand conformations to the crystallographic ones are listed in Table 4 and shown in Figure 4.

Table 4. Comparison of Computational and Experimental Binding Modes of Ligands AA, RA, and SA to the M2A Target[a]

ligand	M2A conformation	RMSD of representative (Å)	mean RMSD (Å)	SD RMSD (Å)
AA	holo	0.7	1.8	0.7
AA	apo	1.1	1.9	0.5
RA	holo	4.0	2.0	0.7
RA	apo	1.5	1.8	0.6
SA	holo	2.6	1.7	0.9
SA	apo	0.3	1.1	0.7

[a] The computational binding modes were produced by the HydroDock protocol introduced in the present study.

For these small ligands (Figure 1b), the conformation pools were generated in relatively short MD simulations of 40−100 ns (Methods and Table S4) appropriate for the selection of the representatives. The search space was also restricted by the helical boundaries of the narrow M2A channel (Figure 1a), and therefore, the selection of representatives was not particularly challenging from the ligand conformation pools containing fairly uniform binding modes (Table 4 and Table S4). Notably, in our previous study,[55] we found that the generation of conformation pools in the cases of large, flexible ligands may require longer MD simulation times, especially if they bind to the target surface.

The final results (Table 4 and Figure 4) show excellent agreement with the experimental ligand conformations[21] in all three cases. A closer inspection of the changes during the MD simulations (Step 4 of HydroDock) shows that ligand binding modes underwent considerable rearrangements due to their interactions with water molecules generated in Step 2. Due to the lack of the anchoring water molecules, dry docking (Step 1 of Hydrodock) produced misdocked binding modes exemplified by Figure 3a. During Step 4, all three ligands entered hydration networks of surrounding water molecules via their protonated amino groups that formed hydrogen bonds with water oxygen atoms (Figure 4). They also adopted their appropriate binding positions (Figure S4) with a rapid rotation and a slight downward movement toward the middle of the channel. Interestingly, besides the crystallographic binding mode, RA also adopted an alternative, parallel orientation corresponding to the higher RMSD of RA, holo in Table 4.

(3) Ligand Binding Modes and the Water Structure in the EC2 Channel of SARS-CoV-2. A recent study[15] explored the interactions of fluoro-AA with EC2 on the basis of chemical shift perturbations from nuclear magnetic resonance (NMR) spectroscopic measurements.

They also used docking calculations to map the anchoring residues during ligand entry from the extraviral space down to N15 of EC2 (Figure 1a).

The study identified a group of apolar entry residues T11...I13 by NMR (asterisks in Figure 5a) and others like N15 by docking calculations (empty circles), respectively. The fluoro-AA in ref 15 is only a slight modification of AA, both having a largely hydrophobic head group and a positively charged tail moiety (Figure 1b), and therefore, similar binding modes can be expected for both ligands on EC2.

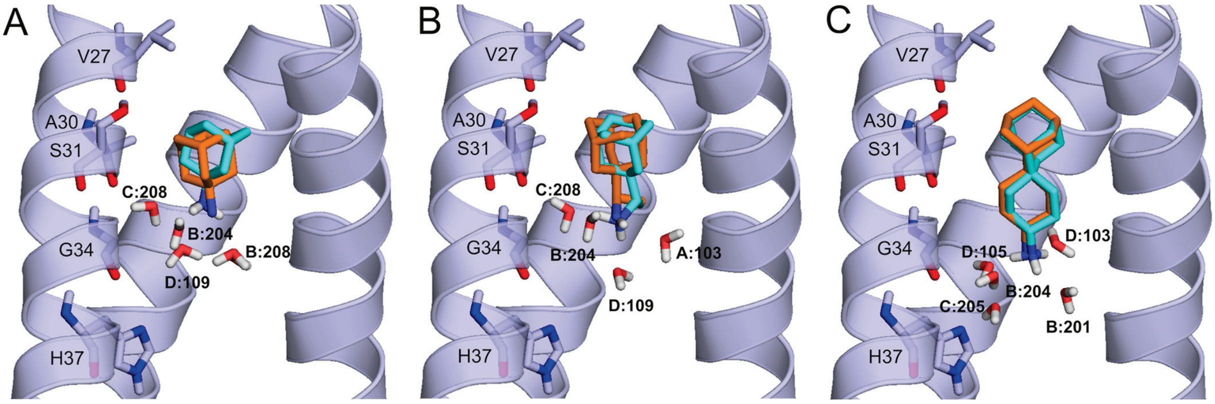


Figure 4. Representative binding modes of ligands (teal sticks for carbon atoms) (A) AA, (B) RA, and (C) SA in the complex with M2A (cartoon) produced by the HydroDock protocol. For comparison, crystallographic ligand binding modes (orange sticks for carbon atoms) are shown as references. Interacting M2A amino acids and water molecules are shown as sticks and labeled accordingly to the residue numbering of 6bkk.

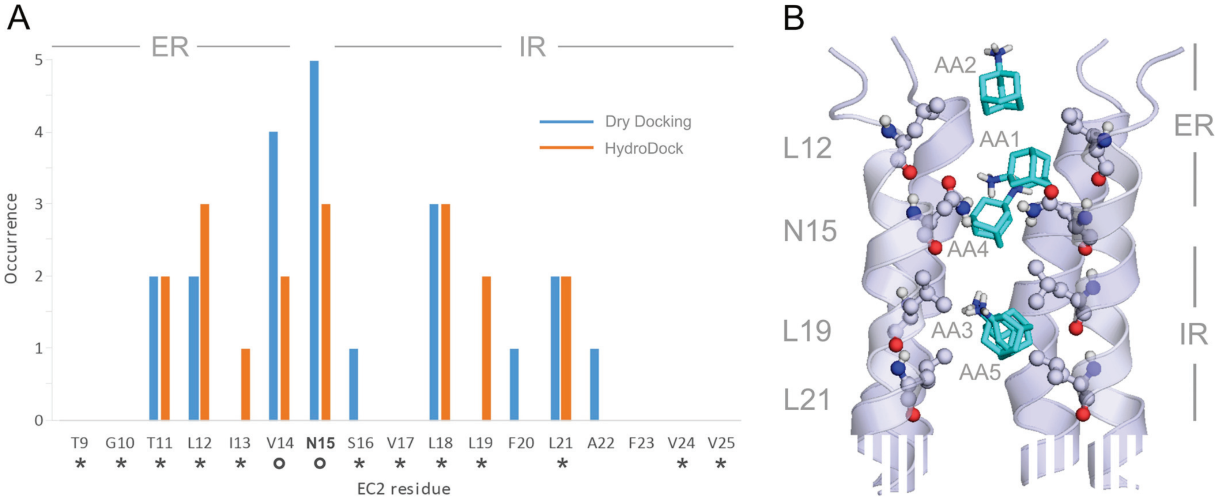


Figure 5. (A) Occurrence of EC2 amino acids interacting with AA in the five binding modes (bars) produced by dry docking (blue bars) and after refinement by HydroDock (orange bars) in the present study. Asterisks and circles indicate interacting amino acids identified by experiments and docking calculations, respectively, in a previous[15] paper. Entrance and intrachannel binding regions are marked as ER and IR, respectively, at the top of the diagrams. (B) Five representative structures of binding modes AA1, ..., AA5 (teal sticks, Table S6) on EC2 (cartoon, truncated at the bottom). The interacting EC2 amino acids are shown as balls and sticks and labeled by their identifiers according to PDB structure 7k3g. ER and IR binding regions are also shown on the right side of the figure. Raw data are provided in Tables S5 and S6 in the Supporting Information.

Inspired by the above NMR-based study[15] on EC2 and the good performance of HydroDock on the M2A channel (previous section), our protocol was applied to map the binding modes of the AA derivatives on the EC2 channel of SARS-CoV-2 (Figure 1). The binding modes of all three ligands (AA, RA, and SA) were mapped by HydroDock using the apo form EC2 as a target from ref 15. The interacting residues of EC2 were collected after dry docking (Figure 5a,b and Table S5) and for the final five representative binding modes produced by HydroDock (Figure 5a,b and Table S6) as well.

A good match was observed (Figure 5a) between the occurrence of EC2 residues involved in the binding modes of fluoro-AA identified in the NMR-based study[15] and AA found by HydroDock in the present study. The results show two main binding regions (Figure 5a,b) of EC2, that is, an entrance region (ER) toward the extraviral space and an intrachannel region (IR) roughly divided by the gating residue N15. Our dry docking calculations showed that the IR region was accessible only in the case if the side chain of the gating N15 was free to move during the docking (Table S5 and Methods), indicating that N15 has a key role in ligand binding mechanisms. The NMR-based study[15] also emphasized the role of this gating residue and concluded that small molecular drug candidates should show high binding affinity to N15 during their entry into EC2.

Water molecules significantly influence the binding modes of ligands to their targets[25] (see also previous sections). As in the above M2A examples, HydroDock refinement of EC2 systems also involved structural hydration, energy minimization, and subsequent 100 ns-long MD simulations for all binding modes found in dry docking (Step 1 in Figure 2). The comparison of the binding pattern after dry docking (blue bars in Figure 5a) with that after HydroDock refinements (orange bars in Figure 5a) may shed light on the influence of water structure on ligand binding to EC2. The hydrophobic belts of binding modes AA1, AA2, and AA4 (ER) and AA3 and AA5 (IR) maintained after HydroDock refinements (Figure 5a). The ER and IR binding regions consist of hydrophobic cores centered on residues L12 (ER) and L19 and L21 (IR), respectively. While the hydrophobic interactions appear in both dry docking and HydroDock results (Figure 5a), there are certain amino acids like L19 found by only one of the methods. In these cases, a rearrangement of the H-bonding system around the protonated amino group of AA was observed further as discussed in the next section and in Figure S5 in details.

The abovementioned hydrophobic belts of EC2 are necessary to accommodate the hydrocarbon heads of the amphipathic AA derivatives; interfacial water molecules help in the orientation of the ligands in the EC2 channel similar to their binding modes in M2A as discussed in the previous section. For example, in the first binding mode of SA (Figure 6 and Tables S5 and S6), its spiroadamantyl group is captured in a sandwich of hydrophobic side chains arranged in several belts in the EC2 channel (Figure 6). However, the hydrophobic interactions alone are not enough to obtain the final orientation of the ligand. Dry docking positioned SA perpendicular to the helical axes of the EC2 channel, and the only H-bonding interaction was formed with a backbone amide group of V24. HydroDock refinements that introduced explicit water molecules yielded a parallel orientation, and the protonated amino group formed three H-bonds with water molecules W1, ..., W3 bridging SA with the inner wall of EC2. This bridging system of waters found by HydroDock resulted in an almost doubled SA-EC2 interaction energy if compared with dry docking (Figure 6). Similar observations can be made for the role of water molecules in the binding of ligands AA (Figure S5) and RA as well.

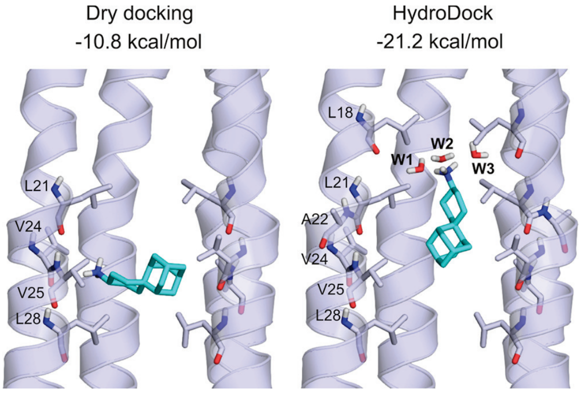


Figure 6. First binding mode of SA (teal sticks) to EC2 (cartoon) after dry docking (left) and HydroDock (right). Interacting amino acid residues are shown as sticks and labeled according to the 7k3g structure file. Water molecules are shown as red and white sticks and labeled as W1, ..., W3. Lennard-Jones interaction energies calculated (Methods) between the ligand SA and the (hydrated) target EC2 are shown at the top of the figure.

CONCLUSIONS

Determination of water molecules mediating drug−target interactions is often missing or they are erroneously positioned due to inherent limitations of structure determination methods.[25] However, the COVID-19 pandemic showed that drug repositioning or design projects often fail due to such structural errors, resulting in misprediction of drug−target interactions. The present study showed that precise positioning of interfacial water molecules is essential for correct calculation of interaction of viral channels with amphipathic ligands of the AA type. A new protocol, HydroDock, was introduced to build the hydrated target−ligand complex structures and help in the repositioning of the ligands between viral channels. In our examples, HydroDock built the hydrated complex structures from scratch and required only the apo target and ligand structures as inputs. The structures showed excellent agreements with experimental results. The atomic resolution complex structures showed that water plays a similar role in the binding of amphipathic AA derivatives to transmembrane ion channels of both influenza A (M2A) and SARS-CoV-2 (EC2). While the hydrophobic regions of the channels capture the bulky hydrocarbon group of the ligand, the surrounding waters direct its orientation parallel with the axes of the channels via bridging interactions with the ionic ligand head. Such elucidation of the role of waters is often requested,[21,25,70] and therefore, future applications of HydroDock can be expected in the design and repositioning of drug candidates.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00488.

Wet docking result, wet docking of AA to M2A, the match between crystallographic reference water positions (red spheres) and the predicted water positions, the movement of RA (teal sticks) during MD simulations, the orientation of ligand AA after dry docking and HydroDock, RA1 and RA2 dry docked binding modes, comparison of holo (6BKK) and apo (3LBW) water structures, the structural fit of M2A targets, the movement of the water molecule, the statistics of generation of the ligand conformation pool, dry docked and HydroDock reprentatives of AA, RA, and SA (PDF)

Coordinate files of all snapshots of ligand binding modes, raw data of ER and IR binding regions, detailed instruction for evaluations (ZIP)

AUTHOR INFORMATION

Corresponding Author

Csaba Hetényi − *Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, 7624 Pécs, Hungary*; orcid.org/0000-0002-8013-971X; Email: hetenyi.csaba@pte.hu

Authors

Balázs Zoltán Zsidó − *Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, 7624 Pécs, Hungary*

Rita Börzsei − *Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School and Department of Pharmacology, Faculty of Pharmacy, University of Pécs, 7624 Pécs, Hungary*

Viktor Szél − *Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, 7624 Pécs, Hungary*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.1c00488

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was funded by the Hungarian National Research, Development and Innovation Office (K123836 and K134214), 2017-1.2.1 NKP-2017-00002 (NAP-2; Chronic Pain Research Group), EFOP-3.6.1.-16-2016-0004, and GINOP 2.3.2-15-2016-00050 "PEPSYS". This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Notes

The authors declare no competing financial interest.

AutoDock (http://autodock.scripps.edu/), MobyWat (http://mobywat.com/), and GROMACS (https://www.gromacs.org/) are open source programs. Schrödinger's Maestro (https://www.schrodinger.com/products/maestro), MOPAC (http://openmopac.net/), GAMESS (https://www.msg.chem.iastate.edu/gamess/), acpype (https://pypi.org/project/acpype/), and antechamber (http://ambermd.org/antechamber/ac.html) are freely accessible programs for academic use. MOLEonline (https://mole.upol.cz/) and RED-vIII (https://upjv.q4md-forcefieldtools.org/REDServer-Development/) are accessible web servers. In-house scripts were used for average ligand structure calculation and RMSD comparisons between the average and the ligand conformations from the ligand pool and between the experimental and the ligand conformations from the ligand pool. A detailed instruction for these evaluations with a README file and the scripts are available in the Supporting Information. There is also an example for one system with all the ligand conformation pool files (sep*.pdb), the in-house scripts mentioned above, log files containing the results of the scripts, excel files created from the log files to help in evaluation, a reference pdb structure for RMSD calculation, the average ligand conformation file, and a fit.pdb file that was used for structure alignment with the gmx trjconv command line. There are three additional examples for energy minimization of the ligands in a separate folder, and all the pdb files were used to create the figures of the main text. A tutorial document, input, and output files are provided in a GitHub repository https://github.com/csabahetenyi/HydroDock.

■ ACKNOWLEDGMENTS

We acknowledge the support from the Governmental Information Technology Development Agency, Hungary. The work was supported by the ÚNKP-20-5 and ÚNKP-20-3-I, New National Excellence Program of the Ministry for Innovation and Technology, and by the PTE ÁOK-KA no:2019/KA-2019-31. The project has been supported by the European Union, cofinanced by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, EFOP-3.6.1-16-2016-00004.

■ ABBREVIATIONS

EC2, SARS-COV-2 envelope protein E; M2A, influenza 2A transmembrane protein

■ REFERENCES

(1) Fu, L.; Ye, F.; Feng, Y.; Yu, F.; Wang, Q.; Wu, Y.; Zhao, C.; Sun, H.; Huang, B.; Niu, P.; Song, H.; Shi, Y.; Xuebing, L.; Wenjie, T.; Qi, J.; Gao, G. F. Both Boceprevir and GC376 Efficaciously Inhibit SARS-CoV-2 by Targeting its Main Protease. *Nat. Commun.* **2020**, *11*, 4417.

(2) Mucke, H. A. M. COVID-19 and the Drug Repurposing Tsunami. *Assay Drug Dev. Technol.* **2020**, *18*, 211−214.

(3) Eder, J.; Sedrani, R.; Wiesmann, C. The Discovery of First-in-Class Drugs, Origins and Evolution. *Nat. Rev. Drug Discov.* **2014**, *13*, 577−587.

(4) Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery, An Industry Perspective. *Nat. Rev. Drug Discov.* **2017**, *16*, 531−543.

(5) Swinney, D. C.; Anthony, J. How Were New Medicines Discovered? *Nat. Rev. Drug Discov.* **2011**, *10*, 507−519.

(6) Lindsay, M. A. Target Discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 831−838.

(7) Schenone, M.; Wagner, B. K.; Clemons, P. A.; Program, B. Biology and Drug Discovery. *Nat. Chem. Biol.* **2017**, *9*, 232−240.

(8) Cao, B.; Wang, Y.; Wen, D.; Liu, W.; Wang, J.; Fan, G.; Ruan, L.; Song, B.; Cai, Y.; Wei, M.; Li, X.; Xia, J.; Chen, N.; Xiang, J.; Yu, T.; Bai, T.; Xie, X.; Zhang, L.; Li, C.; Yuan, Y.; Chen, H.; Li, H.; Huang, H.; Tu, S.; Gong, F.; Liu, Y.; Wei, Y.; Dong, C.; Zhou, F.; Gu, X.; Xu, J.; Liu, Z.; Zhang, Y.; Li, H.; Shang, L.; Wang, K.; Li, X.; Zhou, X.; Dong, X.; Qu, Z.; Lu, S.; Hu, X.; Ruan, S.; Luo, S.; Wu, J.; Peng, L.; Cheng, F.; Pan, L.; Zou, J.; Jia, C.; Wang, J.; Liu, X.; Wang, S.; Wu, X.; Ge, Q.; He, J.; Zhan, H.; Qiu, F.; Guo, L.; Huang, C.; Jaki, T.; Hayden, F. G.; Horby, P. W.; Zhang, D.; Wang, C. A Trial of Lopinavir−Ritonavir in Adults Hospitalized with Severe Covid-19. *N. Engl. J. Med.* **2020**, *382*, 1787−1799.

(9) Horby, P. W.; Mafham, M.; Bell, J. L.; Linsell, L.; Staplin, N.; Emberson, J.; Palfreeman, A.; Raw, J.; Elmahi, E.; Prudon, B.; Green, C.; Carley, S.; Chadwick, D.; Davies, M.; Wise, M. P.; Baillie, J. K.; Chappell, L. C.; Faust, S. N.; Jaki, T.; Jefferey, K.; Lim, W. S.; Montgomery, A.; Rowan, K.; Juszczak, E.; Haynes, R.; Landray, M. J. Lopinavir−Ritonavir in Patients Admitted to Hospital with COVID-19 (RECOVERY), a Randomised, Controlled, Open-Label, Platform Trial. *Lancet* **2020**, *396*, 1345−1352.

(10) Grein, J.; Ohmagari, N.; Shin, D.; Diaz, G.; Asperges, E.; Castagna, A.; Feldt, T.; Green, G.; Green, M. L.; Lescure, F.; Nicastri, E.; Oda, R.; Yo, K.; Quiros-Roldan, E.; Studemeister, A.; Redinski, J.; Ahmed, S.; Bernett, J.; Chelliah, D.; Chen, D.; Chihara, S.; Cohen, S. H.; Cunningham, J.; Monforte, A. D.; Ismail, S.; Kato, H.; Lapadula, G.; L'Her, E.; Maeno, T.; Majumder, S.; Massari, M.; Mora-Rillo, M.; Mutoh, Y.; Nguyen, D.; Verweij, P. E.; Zoufaly, A.; Osinusi, A. O.; DeZure, A.; Zhao, Y.; Zhong, L.; Chokkalingam, A.; Elboudwarej, E.; Telep, L.; Timbs, L.; Henne, I.; Sellers, S.; Cao, H.; Tan, S. K.; Winterbourne, L.; Desai, P.; Mera, R.; Gaggar, A.; Myers, R. P.; Brainard, B. M.; Childs, R.; Flanigan, T. Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N. Engl. J. Med.* **2020**, *382*, 2327−2336.

(11) Beigel, J. H.; Tomashek, K. M.; Dodd, L. E.; Mehta, A. K.; Zingman, B. S.; Kalil, A. C.; Hohmann, E.; Chu, H. Y.; Luetkemeyer, A.; Kline, S.; Lopez de Castilla, D.; Finberg, R. W.; Dierberg, K.; Tapson, V.; Hsieh, L.; Patterson, T. F.; Paredes, R.; Sweeney, D. A.; Short, W. R.; Touloumi, G.; Lye, D. C.; Ohmagari, N.; Oh, M.; Ruiz-Palacios, G. M.; Benfield, T.; Fätkenheuer, G.; Kortepeter, M. G.; Atmar, R. L.; Creech, C. B.; Lundgren, J.; Babiker, A. G.; Pett, S.; Neaton, J. D.; Burgess, T. H.; Bonnett, T.; Green, M.; Makowski, M.; Osinusi, A.; Nayak, S.; Lane, C. Remdesivir for the Treatment of Covid-19 — Final Report. *N. Engl. J. Med.* **2020**, *383*, 1813−1826.

(12) Li, G.; De Clercq, E. Therapeutic Options for the 2019 Novel Coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* **2020**, *19*, 149−150.

(13) Ullrich, S.; Nitsche, C. The SARS-CoV-2 Main Protease as Drug Target. *Bioorg. Med. Chem. Lett.* **2020**, *30*, 127377.

(14) Kim, J.-W.; Kim, E. J.; Kwon, H. H.; Jung, C. Y.; Kim, K. C.; Choe, J.-Y.; Hong, H.-L. Lopinavir-Ritonavir Versus Hydroxychloroquine for Viral Clearance and Clinical Improvement in Patients with Mild to Moderate Coronavirus Disease 2019. *Korean J. Intern. Med.* **2021**, S253.

(15) Mandala, V.; McKay, M.; Shcherbakov, A.; Dregni, A.; Kolocouris, A.; Hong, M. Structure and Drug Binding of the SARS-

CoV-2 Envelope Protein in Phospholipid Bilayers. *Nat. Struct. Mol. Biol.* **2020**, *27*, 1202−1208.

(16) Torres, J.; Maheswari, U.; Parthasarathy, K.; Ng, L.; Liu, D. X.; Gong, X. Conductance and Amantadine Binding of a Pore Formed by a Lysine-Flanked Transmembrane Domain of SARS Coronavirus Envelope Protein. *Protein Sci.* **2007**, *16*, 2065−2071.

(17) Brenner, S. R. The Potential of Memantine and Related Adamantanes Such as Amantadine, to Reduce the Neurotoxic Effects of COVID-19, Including ARDS and to Reduce Viral Replication Through Lysosomal Effects. *J. Med. Virol.* **2020**, *92*, 2341−2342.

(18) Abreu, G. E. A.; Aguilar, M. E. H.; Covarrubias, D. H.; Durán, F. R. Amantadine as a Drug to Mitigate the Effects of COVID-19. *Med. Hypotheses* **2020**, *140*, 109755.

(19) Aranda-Abreu, G. E.; Aranda-Martínez, J. D.; Araújo, R. Use of Amantadine in a Patient with SARS-CoV-2. *J. Med. Virol.* **2021**, *93*, 110−111.

(20) Rejdak, K.; Grieb, P. Adamantanes Might be Protective from COVID-19 in Patients with Neurological Diseases, Multiple Sclerosis, Parkinsonism and Cognitive Impairment. *Mult. Scler. Relat. Disord.* **2020**, *42*, 102163.

(21) Thomaston, J. L.; Polizzi, N. F.; Konstantinidi, A.; Wang, J.; Kolocouris, A.; Degrado, W. F. Inhibitors of the M2 Proton Channel Engage and Disrupt Transmembrane Networks of Hydrogen-Bonded Waters. *J. Am. Chem. Soc.* **2018**, *140*, 15219−15226.

(22) Jeppesen, M. G., Amantadin has Potential for the Treatment of COVID- 19 Because It Targets Known and Novel Ion Channels Encoded by SARS-CoV-2. *Research Square Preprint,* DOI: 10.21203/rs.3.rs-121743/v1.

(23) Fink, K.; Nitsche, A.; Neumann, M.; Grossegesse, M.; Eisele, K.-H.; Danysz, W. Amantadine Inhibits SARS-CoV-2 In Vitro. *Viruses* **2021**, *13*, 539−549.

(24) Hassan, S. S.; Choudhury, P. P.; Roy, B.; Jana, S. S. Missense Mutations in SARS-CoV2 Genomes from Indian Patients. *Genomics* **2020**, *112*, 4622−4627.

(25) Zsidó, B. Z.; Hetényi, C. The Role of Water in Ligand Binding. *Curr. Opin. Struct. Biol.* **2021**, *67*, 1−8.

(26) Acharya, R.; Carnevale, V.; Fiorin, G.; Levine, B. G.; Polishchuk, A. L.; Balannik, V.; Samish, I.; Lamb, R. A.; Pinto, L. H.; DeGrado, W. F.; Klein, M. L. Structure and Mechanism of Proton Transport Through the Transmembrane Tetrameric M2 Protein Bundle of the Influenza A Virus. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 15075−15080.

(27) Schrödinger. *Maestro Schrödinger*; Schrödinger, Release 2020−4.

(28) Stewart, J. J. P. *Stewart Computational Chemistry*; Stewart: Colorado Springs, CO, USA H, MOPAC. 2016

(29) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI, More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1−32.

(30) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. tools, Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821−7839.

(31) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(32) Sousa Da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 367.

(33) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

(34) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity-a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(35) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(36) Bálint, M.; Horváth, I.; Mészáros, N.; Hetényi, C. Towards Unraveling the Histone Code by Fragment Blind Docking. *Int. J. Mol. Sci.* **2019**, *20*, 422.

(37) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(38) Hetényi, C.; Van Der Spoel, D. Blind Docking of Drug-Sized Compounds to Proteins With Up to a Thousand Residues. *FEBS Lett.* **2006**, *580*, 1447−1450.

(39) Hetényi, C.; van der Spoel, D. Efficient Docking of Peptides to Proteins Without Prior Knowledge of the Binding Site. *Protein Sci.* **2002**, *11*, 1729−1737.

(40) Jeszenői, N.; Bálint, M.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Exploration of Interfacial Hydration Networks of Target-Ligand Complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148−158.

(41) Jeszenői, N.; Horváth, I.; Bálint, M.; Van Der Spoel, D.; Hetényi, C. Mobility-Based Prediction of Hydration Structures of Protein Surfaces. *Bioinformatics* **2015**, *31*, 1959−1965.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions For Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(43) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS, Fast, Flexible, And Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(44) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(45) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys* **2007**, *126*, 014101.

(46) Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials, a Molecular Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196−1199.

(47) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals, A New Molecular Dynamics Method. *J. App. Phys.* **1981**, *52*, 7182−7190.

(48) Darden, T.; Darrin, Y.; Pedersen, L. Particle Mesh Ewald, an N log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *12*, 10089−10092.

(49) De Lano, W. L. *The PyMOL Molecular Graphics System*; Version 2.0 Schrödinger, LLC, 2002.

(50) Kevener, H. E.; Zhao, W.; Ball, D. M.; Babaoglu, K.; Qi, J.; White, S. W.; Lee, R. E. Validation of Molecular Docking Programs for Virtual Screening Against Dihydropteroate Synthase. *J. Chem. Inf. Model.* **2009**, *49*, 444−460.

(51) Castro-Alvarez, A.; Costa, A. M.; Vilarrasa, J. The Performance of Several Docking Programs At Reproducing Protein-Macrolide-Like Crystal Structures. *Molecules* **2017**, 136−150.

(52) Mena-Ulecia, K.; Tiznado, W.; Caballero, J. Study of the Differential Activity of Thrombin Inhibitors Using Docking, QSAR, Molecular Dynamics, And MM-GBSA. *PLoS One* **2015**, *10*, No. e0142774.

(53) Ramírez, D.; Caballero, J.; Is, I. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data. *Molecules* **2018**, *23*, 1038.

(54) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(55) Bálint, M.; Jeszenői, N.; Horváth, I.; Van Der Spoel, D.; Hetényi, C. Systematic Exploration of Multiple Drug Binding Sites. *Aust. J. Chem.* **2017**, *9*, 65−77.

(56) Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M. J.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations. 4. van der Waals Parametrization. *J. Phys. Chem. B* **2012**, *116*, 7088−7101.

(57) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling Water Molecules in Protein-Ligand Docking Using GOLD. *J. Med. Chem.* **2005**, *48*, 6504−6515.

(58) Pastor, M.; Cruciani, G.; Watson, K. A. A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site Into

a Three-Dimensional Quantitative Structure - Activity Relationship Analysis. *J. Med. Chem.* **1997**, *40*, 4089−4102.

(59) Rarey, M.; Kramer, B.; Lengauer, T. The Particle Concept, Placing Discrete Water Molecules During Protein- Ligand Docking Predictions. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 17−28.

(60) Huang, N.; Shoichet, B. K. Exploiting Ordered Waters in Molecular Docking. *J. Med. Chem.* **2008**, *58*, 4862−4865.

(61) Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein- Ligand Binding Sites and its Potential Application to Drug Design. *Chem. Biol.* **1996**, *3*, 973−980.

(62) Carugo, O. Correlation Between Occupancy and B Factor Of Water Molecules in Protein Crystal Structures. *Protein Eng.* **1999**, *12*, 1021−1024.

(63) Kim, K. H. Outliers in SAR and QSAR , 3 . Importance of Considering the Role of Water Molecules in Protein − Ligand Interactions and Quantitative Structure − Activity Relationship Studies. *J. Comput.-Aided Mol. Des.* **2021**, 371.

(64) Kim, K. H. Outliers in SAR and QSAR, Is Unusual Binding Mode a Possible Source Of Outliers? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 63−86.

(65) Kim, K. H. Outliers in SAR and QSAR, 2. Is a Flexible Binding Site a Possible Source of Outliers? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 421−435.

(66) Maveyraud, L.; Mourey, L. Protein X-ray Crystallography and Drug Discovery. *Molecules* **2020**, *25*, 1030−1048.

(67) Søndergaard, C. R.; Garrett, A. E.; Carstensen, T.; Pollastri, G.; Nielsen, J. E. Structural artifacts in protein-ligand X-ray Structures, Implications for the Development of Docking Scoring Functions. *J. Med. Chem.* **2009**, *52*, 5673−5684.

(68) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database, Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(69) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database, Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(70) Thomaston, J. L.; Konstantinidi, A.; Liu, L.; Lambrinidis, G.; Tan, J.; Caffrey, M.; Wang, J.; Degrado, W. F.; Kolocouris, A. X-ray Crystal Structures of the Influenza M2 Proton Channel Drug-Resistant V27A Mutant Bound to a Spiro-Adamantyl Amine Inhibitor Reveal the Mechanism of Adamantane Resistance. *Biochemistry* **2020**, *59*, 627−634.

D20

hetenyi.csaba_83_23

doi:10.1016/j.jmb.2005.04.039 *J. Mol. Biol.* (2005) **350**, 156–169

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT

ELSEVIER

Extended Intermolecular Interactions in a Serine Protease–Canonical Inhibitor Complex Account for Strong and Highly Specific Inhibition

Krisztián Fodor[1], Veronika Harmat[2], Csaba Hetényi[3], József Kardos[3] József Antal[3], András Perczel[4], András Patthy[1], Gergely Katona[5] and László Gráf[1,3]*

[1]*Biotechnology Research Group of the Hungarian Academy of Sciences, Eötvös Loránd University, Budapest, H-1117 Hungary*

[2]*Protein Modeling Group of the Hungarian Academy of Sciences and Eötvös Loránd University Budapest, H-1117, Hungary*

[3]*Department of Biochemistry Eötvös Loránd University Budapest, H-1117, Hungary*

[4]*Department of Organic Chemistry, Eötvös Loránd University, Budapest, H-1117 Hungary*

[5]*Department of Biochemistry University of Leicester University Road, Leicester LE1 7RH, UK*

**Corresponding author*

We have previously shown that a trypsin inhibitor from desert locust *Schistocerca gregaria* (SGTI) is a taxon-specific inhibitor that inhibits arthropod trypsins, such as crayfish trypsin, five orders of magnitude more effectively than mammalian trypsins. Thermal denaturation experiments, presented here, confirm the inhibition kinetics studies; upon addition of SGTI the melting temperatures of crayfish and bovine trypsins increased 27 °C and 4.5 °C, respectively. To explore the structural features responsible for this taxon specificity we crystallized natural crayfish trypsin in complex with chemically synthesized SGTI. This is the first X-ray structure of an arthropod trypsin and also the highest resolution (1.2 Å) structure of a trypsin–protein inhibitor complex reported so far. Structural data show that in addition to the primary binding loop, residues P_3–P_3' of SGTI, the interactions between SGTI and the crayfish enzyme are also extended over the P_{12}–P_4 and P_4'–P_5' regions. This is partly due to a structural change of region P_{10}–P_4 in the SGTI structure induced by binding of the inhibitor to crayfish trypsin. The comparison of SGTI–crayfish trypsin and SGTI–bovine trypsin complexes by structure-based calculations revealed a significant interaction energy surplus for the SGTI–crayfish trypsin complex distributed over the entire binding region. The new regions that account for stronger and more specific binding of SGTI to crayfish than to bovine trypsin offer new inhibitor sites to engineer in order to develop efficient and specific protease inhibitors for practical use.

Keywords: serine protease; canonical inhibitor; X-ray crystallography; NMR; specificity

Introduction

Besides their involvement in extra- and intracellular breakdown of proteins to amino acids, serine proteases catalyze highly specific cleavages in a number of biological processes from blood clotting to the complement cascade of the immune system. They regulate the level of particular proteins in the organism or convert their inactive forms to active ones. While serine proteases perform a wide range of functions essential to life, they can also be harmful. This may be the reason why their activity is not only controlled by the proteolytic activation of their inactive forms, by their auto-inactivation (autolysis), and their transport, but also by their inhibition with specific protease inhibitors.

The control of trypsin activity in pancreas and intestine provides a good example of these mechanisms. As discovered more than 130 years ago[1] trypsin is produced by the pancreas in its inactive form. It was shown a few decades later that in addition to trypsinogen and other protease zymogens, pancreas also contains a protein inhibitor of pancreatic trypsin[2] that was shown to function as a protector of premature trypsinogen

Abbreviations used: DSC, differential scanning calorimetry; NOE, nuclear Overhauser enhancement; MD, molecular dynamics.

E-mail address of the corresponding author: graf@ludens.elte.hu auto-activation. Since then several pairs of serine proteases and inhibitors have been discovered and become not only targets of physiological studies but also favorite structural models for protein–protein interactions. A further aspect that makes investigation of serine protease inhibitors extremely important is their potential use in therapy. Thrombin inhibitors in use (anticoagulants) are classic examples of the practical realization of this kind of research.[3–5]

Until recently, trypsin complexed with substrate-like canonical serine protease inhibitors provided the only models for protease–protease inhibitor interactions.[6,7] With the development of biotechnology and bioinformatics there is an increasing demand for higher resolution X-ray structures of serine protease–inhibitor complexes, which reveal the specific interactions responsible for the strong and selective inhibitory effect and provide scaffolds and reliable data sets for *in silico* and *in vitro* inhibitor design. Though there are at least 18 protein families in this class of inhibitors with different overall folds, all appear to share a distinct and similar conformation of the primary binding loop.[7,8] This loop has long been thought to exhibit the same main-chain conformation in both free form and in complex with the protease and to be the major determinant of inhibition.[6,7] Our recent NMR studies on the backbone dynamics of small inhibitors of the newly discovered pacifastin family,[9–12] however, have shown that the binding loops in these inhibitors are less defined and more flexible than all the remaining part of the molecule.[13,14] Another intriguing conclusion of these studies was that SGTI (trypsin inhibitor from *Schistocerca gregaria*) is taxon-specific, inhibiting arthropod trypsins, such as the crayfish one, orders of magnitudes more effectively than the mammalian ones.[15] Our interest in the structural basis of this strong interaction between SGTI and crayfish trypsin initiated the present study.

Here we report the results of an experimental approach (differential scanning calorimetry) to demonstrate the strength of interaction between SGTI and crayfish trypsin and the crystal structure of their complex. The three-dimensional structure reported here presents the first arthropod trypsin structure and one of the highest atomic resolution of a serine-protease–protein inhibitor complex determined so far. Results from our structure-based molecular dynamics calculations are in agreement with the experimental data showing that the intermolecular interactions in the crayfish trypsin–SGTI complex are much stronger than those in the bovine trypsin–SGTI one. Our data provide experimental support to the hypothesis[16,17] that taxon specificity of inhibitors of the pacifastin family like SGTI is at least partly due to their interaction with the protease outside the commonly used interaction site of a canonical protease inhibitor, the binding loop. Engineering these newly explored sites may allow the production of highly specific inhibitors of therapeutically relevant proteases.

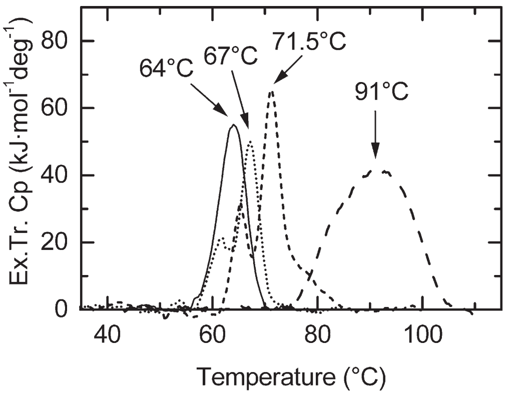


Figure 1. Thermal unfolding profiles of trypsins and their complexes with SGTI. Excess transition heat capacity of crayfish trypsin (–), bovine trypsin (· · · ·), bovine trypsin–SGTI complex (- - - -), and crayfish trypsin–SGTI complex (– – –) in 20 mM sodium phosphate, 100 mM NaCl (pH 7.0), using a heating rate of 1 deg.C/minute. The corresponding melting temperatures are indicated.

Results

Thermal stability of crayfish trypsin–SGTI and bovine trypsin–SGTI complexes

Bovine and crayfish trypsins without SGTI exhibited melting profiles with melting temperatures (T_{m}) around 67 °C and 64 °C, respectively, indicating somewhat lower structural stabilities of the crayfish enzyme (Figure 1). Addition of SGTI to bovine trypsin increased the T_{m} value to 71.5 °C, suggesting that stability of bovine trypsin is increased through interactions with SGTI. In the presence of SGTI, crayfish trypsin showed dramatically increased stability against thermal denaturation with a T_{m} value of 91 °C, indicating stronger enzyme–inhibitor interactions compared to the bovine trypsin–SGTI complex (Figure 1). Although the unfolding transitions were not reversible, differences in T_{m} values were independent of heating rate, suggesting that they reflect a real thermodynamic difference in stability.

Comparison of the amino acid sequence of crayfish and other trypsins

The nucleotide sequence was determined by DNA sequencing of recombinant crayfish trypsin. The amino acid sequences derived from this DNA sequence and from the X-ray structure of crayfish trypsin–SGTI complex using trypsin purified from *Astacus leptodactylus* were only different at a single position; in the X-ray structure a Val appears to replace an Ala residue at position 59 (Figure 2(a)). Comparing the amino acid sequences of crayfish trypsin to vertebrate trypsins, there is 41%–46%

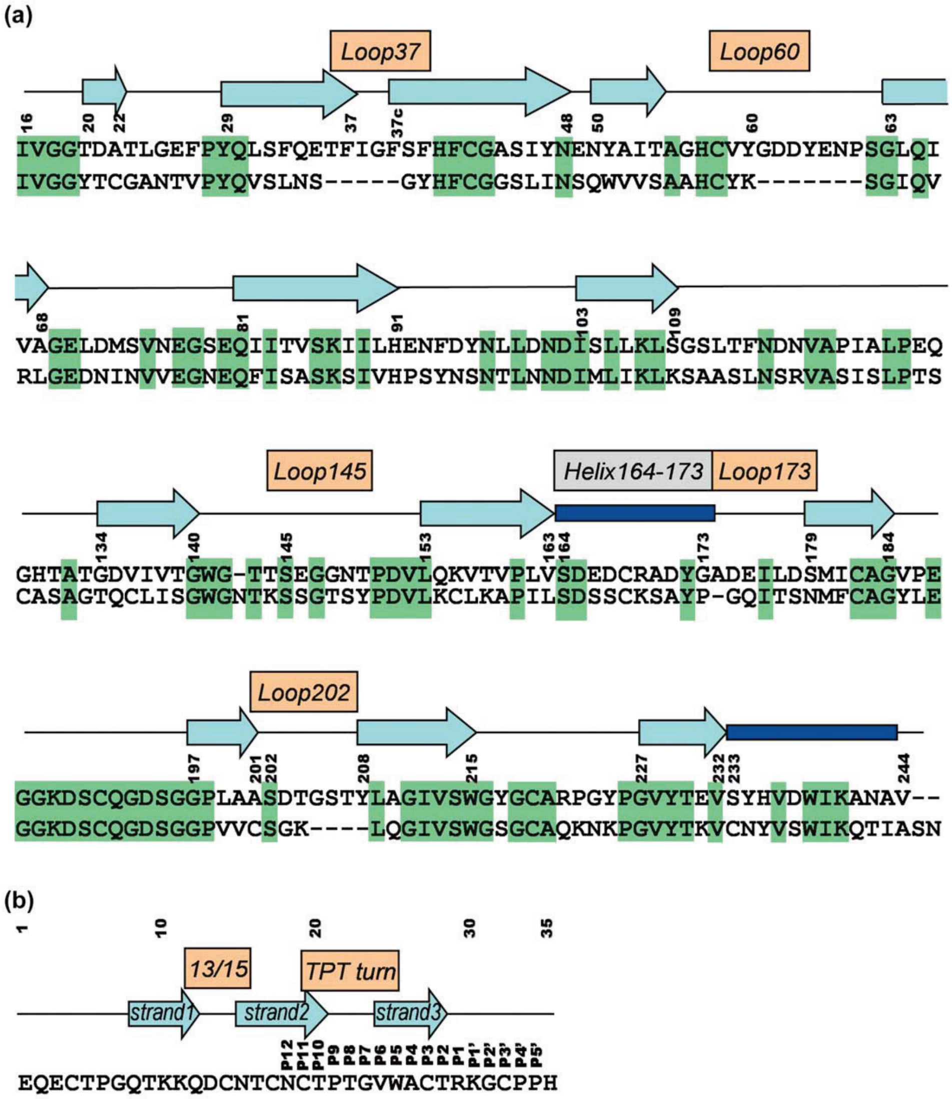

(a)
Loop37
Loop60
16 20 22 29 37 37c 48 50 60 63
IVGGTDATLGEFPYQLSFQETFIGFSFHFCGASIYNENYAITAGHCVYGDDYENPSGLQI
IVGGYTCGANTVPYQVSLNS-----GYHFCGGSLINSQWVVSAAHCYK-------SGIQV
68 81 91 103 109
VAGELDMSVNEGSEQIITVSKIILHENFDYNLLDNDISLLKLSGSLTFNDNVAPIALPEQ
RLGEDNINVVEGNEQFISASKSIVHPSYNSNTLNNDIMLIKLKSAASLNSRVASISLPTS
Loop145
Helix164-173
Loop173
134 140 145 153 163 164 173 179 184
GHTATGDVIVTGWG-TTSEGGNTPDVLQKVTVPLVSDEDCRADYGADEILDSMICAGVPE
CASAGTQCLISGWGNTKSSGTSYPDVLKCLKAPILSDSSCKSAYP-GQITSNMFCAGYLE
Loop202
197 201 202 208 215 227 232 233 244
GGKDSCQGDSGGPLAASDTGSTYLAGIVSWGYGCARPGYPGVYTEVSYHVDWIKANAV--
GGKDSCQGDSGGPVVCSGK----LQGIVSWGSGCAQKNKPGVYTKVCNYVSWIKQTIASN
(b)
1 10 20 30 35
13/15
TPT turn
strand1
strand2
strand3
P12 P11 P10 P9 P8 P7 P6 P5 P4 P3 P2 P1 P1' P2' P3' P4' P5'
EQECTPGQTKKQDCNTCNCTPTGVWACTRKGCPPH

Figure 2. Amino acid sequences of crayfish trypsin and SGTI. (a) The sequence of crayfish trypsin (upper line) was aligned with the sequence of bovine trypsin (lower line, PDB id 3PTB) based on superposition of their 3D structures. Conserved residues are shown in green boxes. Secondary structure elements of crayfish trypsin are shown (α-helices as boxes, β-sheets as arrows) with their starting and end points marked (chymotrypsin numbering). Some regions showing structural differences between the crayfish and bovine enzymes are labeled in frames. (b) The sequence of SGTI with secondary structure elements shown (β-strands as arrows with labels). The two loops at the two ends of β-strand 2 are labeled. Residues forming contacts with crayfish trypsin are labeled as P_{12}–P_5'.

sequence homology between them (e.g. 41% for bovine anionic trypsin). In case of non-crustacean arthropod trypsins, amino acid sequence homology searches resulted in a 40%–53% homology while crustacean trypsins show a homology as high as 80%–98% with trypsin from *A. leptodactylus*.

General structural features of crayfish trypsin–SGTI complex

Crystals of natural crayfish trypsin in complex with SGTI synthesized by solid phase chemical synthesis were grown (see Materials and Methods) and diffracted to 1.2 Å resolution (PDB accession no. 1YR4). With the exception of the N and C termini of SGTI, the model is well defined in electron density (Figure 3(a)). Residues Ser79, Ser104 and Lys239 of crayfish trypsin as well as Cys27 of SGTI possess dual conformations. Crayfish trypsin exhibits the conserved core structure of the chymotrypsin fold consisting of two six-stranded β-barrel domains packed against each other, with the catalytic residues located at the junction of the two barrels (Figure 3(b) and (c)). The

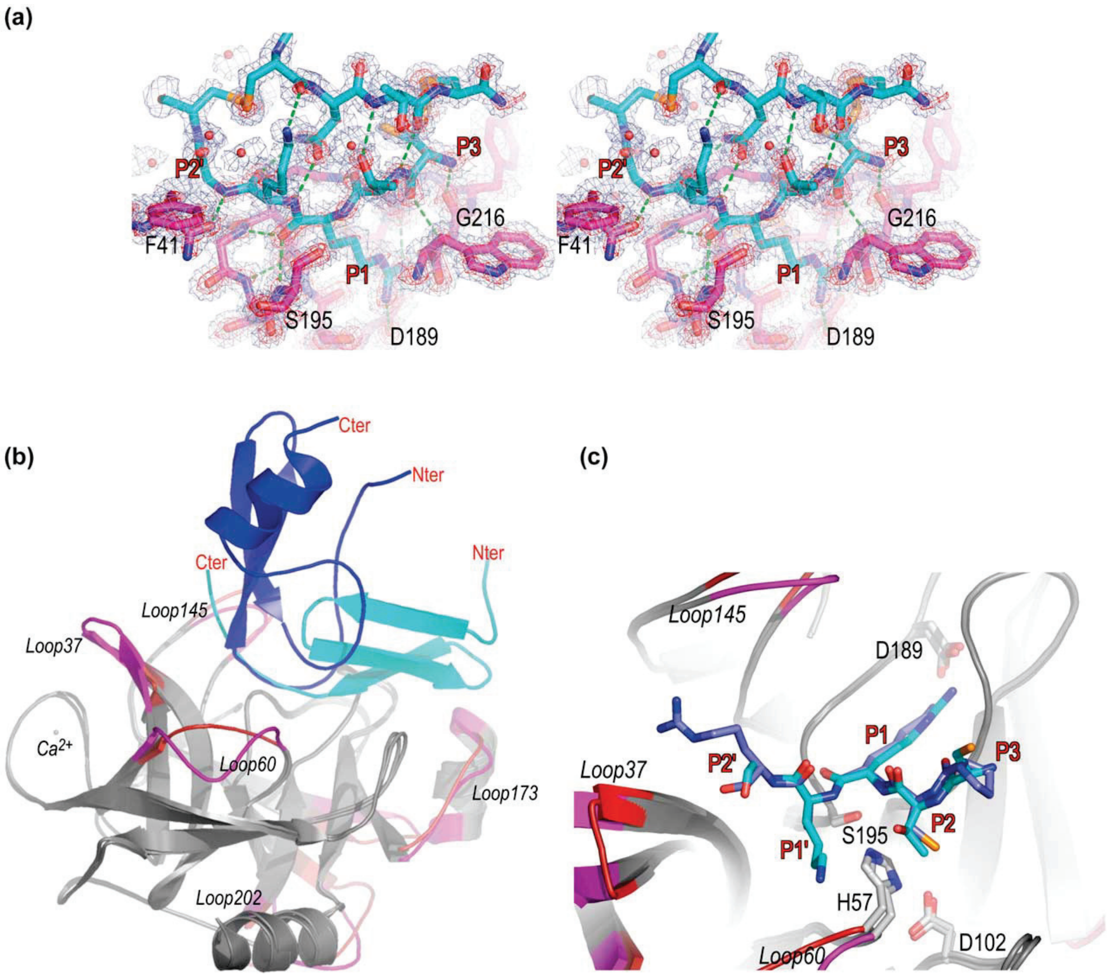


Figure 3. Crystal structure of the crayfish trypsin–SGTI complex. (a) Stereo view of the P_3–P_3' (carbon atoms in light blue), S_3–S_3' (carbon atoms in magenta) region with the $2F_o - F_c$ electron density map contoured at 1σ (blue) and 3σ (red). Protein–protein hydrogen bonds are shown as green shaded lines. (b) Overall conformation of the crayfish trypsin–SGTI complex compared with bovine trypsin–BPTI complex (PDB id 3BTK). Color codes are as described in (c). (c) Conformations of the P_2–P_2' regions of the inhibitors in the complexes, with the catalytic triad (carbon atoms in grey). In (b) and (c) the loop regions different for the two enzymes are colored magenta and red for the crayfish and bovine enzymes, respectively. Conserved structural elements are shown in grey. Carbon atoms of SGTI and BPTI are shown in light blue and dark blue, respectively. N, O and S atoms are shown in atomic colors. Black and red labels are used for the enzymes and inhibitors, respectively. For labeling of loop regions see Figure 2 and the text. Figures 3, 5(b) and 6(a) were generated by PyMOL.[51]

catalytic residues of trypsin are present in their active conformation. The overall structures of crayfish trypsin and bovine trypsin are similar. The Ca^{2+} binding loop characteristic to trypsin binds Cd^{2+} in the crystal structure as the crystallization medium contained Cd^{2+} in high concentration. (The ion in the binding loop had significantly stronger electron density than a Ca^{2+}, which was revealed by a positive $F_o - F_c$ difference Fourier peak. The nature of the bound ion was further evaluated by an anomalous difference Fourier map and *B*-factor analyses.) The conformation of SGTI is comparable to that determined by NMR spectroscopy (PDB accession no. 1KJ0).[13]

It is important to note that the major features of binding of crayfish trypsin to SGTI appear to be identical to the binding of bovine trypsin to bovine pancreatic trypsin inhibitor (BPTI) (PDB accession no. 3BTK),[18] despite the completely different fold of the inhibitors (Figure 3(c)); the antiparallel β-sheets formed between the proteases and the corresponding inhibitors are superimposable at sites S_3–S_2' (in the protease) and P_3–P_2' (in the inhibitor). A novel feature of protease–protease inhibitor interactions is that the interface region in the crayfish trypsin–SGTI complex is much more extended than in serine protease–inhibitor complexes of other inhibitor families (see Discussion).

Geometry of the scissile peptide bond in the inhibitor

The scissile peptide bond between Arg29 and Lys30 is present at full occupancy in the inhibitor complex. It is planar within the calculated coordinate error with the carbonyl carbon atom raising only 0.01 Å above the plane defined by the carbonyl oxygen and carbon alpha atom of Arg29 and the amid nitrogen of Lys30. The distance between the carbonyl carbon and the hydroxyl group of the catalytic serine is 2.69 Å, which is 0.09 Å shorter than the corresponding distance between the attacking water molecule and the carbonyl carbon of the ester bond in an atomic resolution elastase acyl-enzyme.[19] The shorter than van der Waals distance indicates significant orbital overlap between the two atoms although it is still too long for a true covalent bond. The active site is completely shielded from the solvent in the crayfish trypin–SGTI complex; the closest water molecule is 6.63 Å from the carbonyl carbon of the scissile peptide bond.

Loops of the enzyme

There are four loops of crayfish trypsin that are remarkably different in comparison with those of vertebrate trypsins (marked in Figures 2(a) and 3(b) as Loop37, Loop60, Loop145 and Loop202) from which two loops are important regarding inhibitor binding. (1) In contrast to bovine trypsin a more extended hydrophobic region is present in crayfish trypsin with a five residue insertion at position 37. The corresponding loop is referred to as Loop37. The insertion is manifested by an extension of two β-strands connected with a turn containing three phenylalanine residues and an isoleucine, which are oriented towards the inhibitor and interact with the C-terminal segment of SGTI. (2) Another insertion of seven residues occurs at position 60 (Loop60). Similar insertions could be found in some highly specified enzymes like those involved in the complement or blood clotting system (thrombin, mannose binding lectin associated serine protease, etc.).[20] However, while the so-called Loop60 of thrombin has direct influence on the S_2–P_2 interactions, this loop region of the crayfish enzyme turns away from the bound ligand and broadens the substrate binding groove, and thus plays a role in the formation of the S_1'–P_1' interaction.

Disulfide bridges of crayfish trypsin

Crayfish trypsin differs from vertebrate trypsins in its disulfide bond pattern. While bovine trypsin has six disulfide bonds, crayfish trypsin has only three. The conserved disulfide bridges are at positions 42–58, 168–182 and 191–220, respectively. All evolutionary conserved disulfide bonds are close to the active site of the enzyme, which was revealed by sequence comparison studies.[21] The 22–157 inter-domain disulfide bridge, which connects sequentially distant parts of the molecule and is suggested to be important in the structural stability of trypsins[22] is absent from crayfish trypsin. The two segments which are connected *via* this disulfide bond in bovine trypsin, are stabilized by a salt bridge between Lys157 and Glu26 in addition to main-chain hydrogen bonds in the crayfish enzyme. Absence of this disulfide bond may facilitate the relative motions of the two β-barrel domains. Another disulfide bridge is missing at positions 128–232. Ser232 connects *via* a hydrogen bond to the carbonyl oxygen of His128 while its carbonyl oxygen accepts another hydrogen bond from the Gly127 amide group. Consequently, the main-chain conformation of this part of the molecule is more similar to that of chymotrypsins that also lack this disulfide bridge. The disulfide bond at position 136–201 of bovine trypsin is also absent from crayfish trypsin. However, the main-chain conformation of this region is identical to that of the bovine enzyme. Despite the compensatory stabilizing interactions, the lack of the disulfide bridges, especially the one that connects the two domains, may cause the observed 3 °C drop in melting temperature of the crayfish trypsin compared to the bovine enzyme in the differential scanning calorimetry (DSC) study (see Figure 1).

Comparison of the interaction sites in crayfish trypsin–SGTI and bovine trypsin–SGTI complexes

The actual X-ray structure of the crayfish trypsin–SGTI complex was compared with the modeled bovine trypsin–SGTI complex, as well as with representative structures from the molecular dynamics run.

There is an important difference in the interaction pattern at the P_1' site. As seen in Figure 4(a) the P_1' lysine residue (light blue) of SGTI in the crayfish trypsin–SGTI complex is stabilized by a hydrogen bond with the Cys14 carbonyl oxygen atom of SGTI. Though its distance from Asp60b (located in Loop60) and Glu35 of crayfish trypsin (magenta) is about 8 Å in the crystal structure, the molecular dynamics simulations reveal that its position is stabilized closer to these negatively charged residues in solution establishing a weak electrostatic interaction (typical distance of charged groups of the P_1' lysine and Glu35 of crayfish trypsin is 5 Å; see Discussion). Our model of bovine trypsin–SGTI complex shows the P_1' lysine residue (green) located at a position similar to that in the crayfish trypsin–SGTI complex but the S_1' groove does not contain charged side-chains except for Lys60 (dark blue). The positively charged ε-amino group of this Lys60 forms a hydrogen bond to the Tyr39 side-chain that stabilizes its position at a distance of 6 Å from the P_1' Lys residue of SGTI.

Outside the primary binding region in both complexes there are further interactions. In crayfish trypsin–SGTI complex (Figure 4(b)) three phenylalanine residues of loop37 (magenta) bind Pro33 (P_4') of SGTI while Pro34 (P_5') turns outside. Phe39

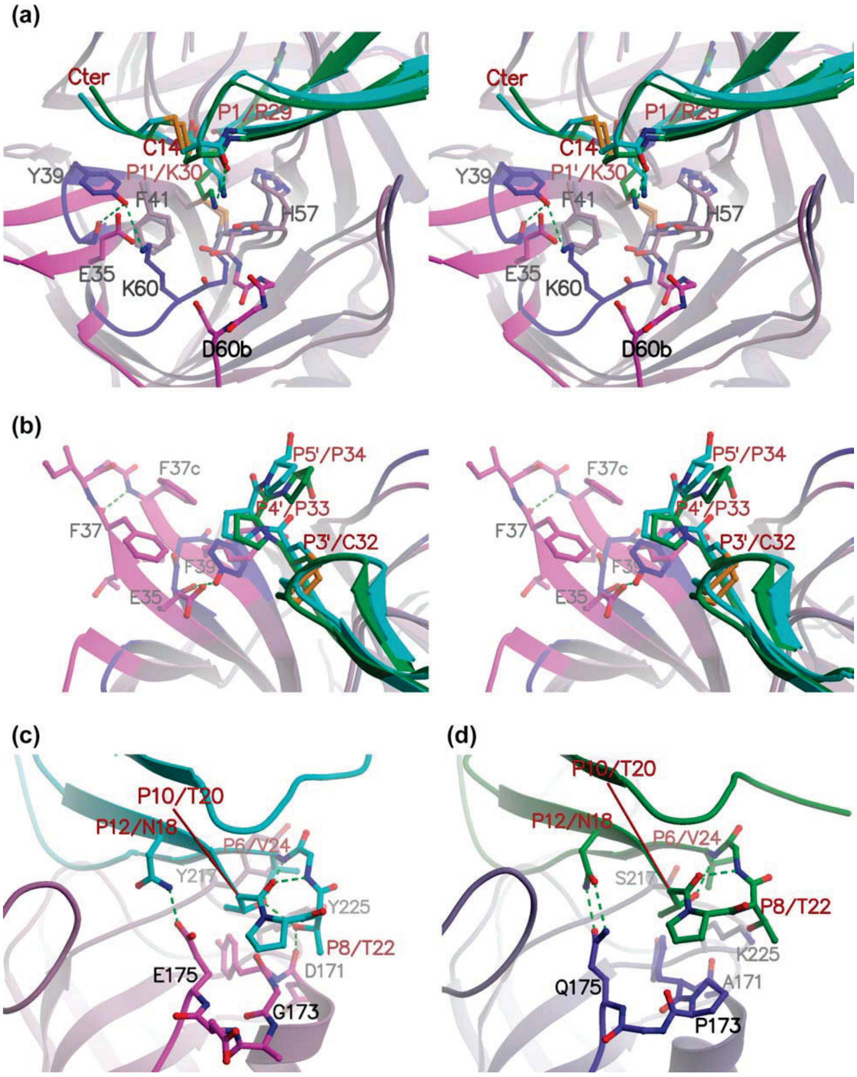


Figure 4. Extended binding region determining taxon specificity of SGTI. The crystal structure of the crayfish trypsin (magenta)–SGTI (light blue) complex superimposed over the representative model (MD, 180 ps step) of bovine trypsin (dark blue)–SGTI (green) complex. Conserved structural motives of trypsin are shown in grey. N, O and S atoms are shown in atomic colors. Hydrogen bonds are shown as green shaded lines. Black and red labels are used for the enzymes and inhibitors, respectively. (a) Stereo view of the P1′ residue accommodated in the S1′ site. Distances between the charged groups of the enzymes and the P1′ lysine amino group are 7.67 Å, 8.22 Å and 6.19 Å for E35 and D60b of crayfish trypsin and K60 of bovine trypsin, respectively. The conformation of the side-chain of the P1′ lysine is stabilized by an intramolecular hydrogen bond. (b) Binding of the P_4'–P_5' region (stereo view) is dominated by hydrophobic contacts. (c) Binding of the P_{12}–P_6 region by the crayfish enzyme is realized by several hydrogen bonds. (d) Binding of the P_{12}–P_6 region by the bovine enzyme. The hydrogen bonds of the P_8 threonine are lost, while stacking interaction is established between the P_9 proline and Pro173. The Figure was generated by MOLSCRIPT.[52]

of this cluster is in a key position, because it forms a stacking interaction with the Pro33 residue (P_4') (light blue). The shorter Loop37 (dark blue) of bovine trypsin is more rigid and contains only one aromatic residue, Tyr39, which forms a stacking interaction with the P_4' proline (light blue). Glu35 of crayfish trypsin, which was mentioned above as one of the electrostatic partners of the P_1' Lys residue of SGTI has another important role that it stabilizes Loop37.

Figure 4(c) illustrates the interaction between the P_{12}–P_6 region (light blue) of the inhibitor with crayfish trypsin (magenta). Val24 (P_6) forms a van der Waals interaction with Tyr217. Thr22 (P_8) forms hydrogen bonds with the 164–173 helix of crayfish trypsin. Thr22, important for the recognition of the enzyme, is stabilized by Thr20 (P_{10}) *via* a hydrogen bond. The hydroxyl group of Thr20 also stabilizes the backbone conformation of the P_{12}–P_6 loop.

In the bovine trypsin–SGTI complex the

interaction pattern is different (Figure 4(d)). At position 217 there is a serine residue instead of tyrosine that forms only a very weak hydrophobic interaction with Val24 of SGTI. In vertebrate trypsins, residue 173 is proline, which makes the 164–173 helix one residue shorter (dark blue). Differences in the backbone conformation of the 172–173 region cause loss of hydrogen bond interactions in the bovine trypsin–SGTI complex; the Thr22–enzyme hydrogen bond is missing, as well as the Thr20–Thr22 intra-molecular hydrogen bond.

Calculation of interaction energies of the proteins in the SGTI–crayfish trypsin and SGTI–bovine trypsin complexes

In order to find the structural basis of different inhibitory efficiencies of SGTI on different trypsins, structure-based calculations on the SGTI–trypsin complexes were performed.

Free energy of binding was calculated with the scoring function of AutoDock 3.0 (Materials and Methods). Scoring of the crystallographic and energy-minimized complexes of SGTI–crayfish trypsin and SGTI–bovine trypsin, respectively, resulted in a $\Delta\Delta G_b = -6.94$ kcal/mol lower binding free energy for the SGTI–crayfish trypsin complex. The difference between the complexes remained significant for conformations of the molecular dynamics trajectory (Supplementary Data; Figure 1). Contribution of each amino acid residue of SGTI to the interaction energy differences (free energy of binding) is depicted in Figure 5.

Extensive interactions facilitate conformational changes in the structure of SGTI

To search for any possible conformational changes in the inhibitor upon its binding to the enzyme we compared the X-ray structure of complexed SGTI with that of the solution structure of the free inhibitor.[13] Superpositions of atoms of SGTI in the complex and the average structure of the free forms yielded a backbone root-mean-square deviation of 1.80 Å in region 4–32 (Figure 6(a)).

Alignment of the NMR structure ensemble with the X-ray structure of complexed SGTI and a careful comparison of the backbone φ, ψ angles were carried out (Supplementary Data; Table 1). Additionally, NOE-derived restraints and corresponding distances in the complex are also compared (Table 2). The first (residues 8–12; see Figure 2(b) for secondary structure) and the second (residues 15–20) β-strands as well as the second loop (labeled as 13/15 in Figure 2(b)) interconnecting strands 1 and 2 have rather similar conformational properties both in the free and the complexed forms of the inhibitor.

The TPT turn (residues 20–22, P_{10}–P_8) shows conformational features resembling a somewhat distorted β-turn (a type II in solution and a type I in the complex) clearly stabilized both in solution and in the complex by the i-($i+2$) backbone hydrogen bond surrounding proline and by the Thr–Thr side-chain interactions. Nevertheless, an important structural change occurs in this part of the inhibitor. Both the above-mentioned 20–22 TPT segment with Gly23 and the Cys4–Thr5 region appear to be relatively stable in solution (with S^2 values among

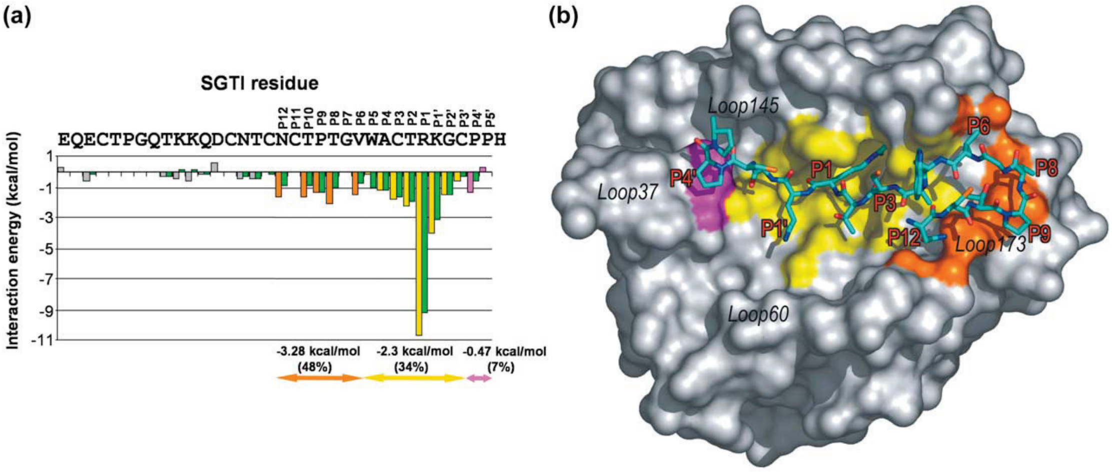


Figure 5. Energetic analysis of binding SGTI by crayfish and bovine trypsin. (a) Intermolecular interaction energy values corresponding to each residue of SGTI in the crystallographic structure of the SGTI–crayfish trypsin complex and energy-minimized structure ($t=0$ ps) of the SGTI–bovine trypsin complex. The energy bars for the SGTI–bovine trypsin complex are colored green, while those of the SGTI–crayfish trypsin complex are colored magenta for P_4'–P_5', yellow for P_5–P_3', orange for P_{12}–P_6 and grey outside these regions. All the three binding regions contribute significantly to the taxon specificity of SGTI featured by the energy difference in interaction energy values shown for these regions. (b) The molecular surface of crayfish trypsin with the binding regions for P_4'–P_5', P_5–P_3' and P_{12}–P_6 of SGTI colored magenta, yellow and orange, respectively. The P_{12}–P_5' segment of SGTI is shown in atomic colors (C atoms in light blue).

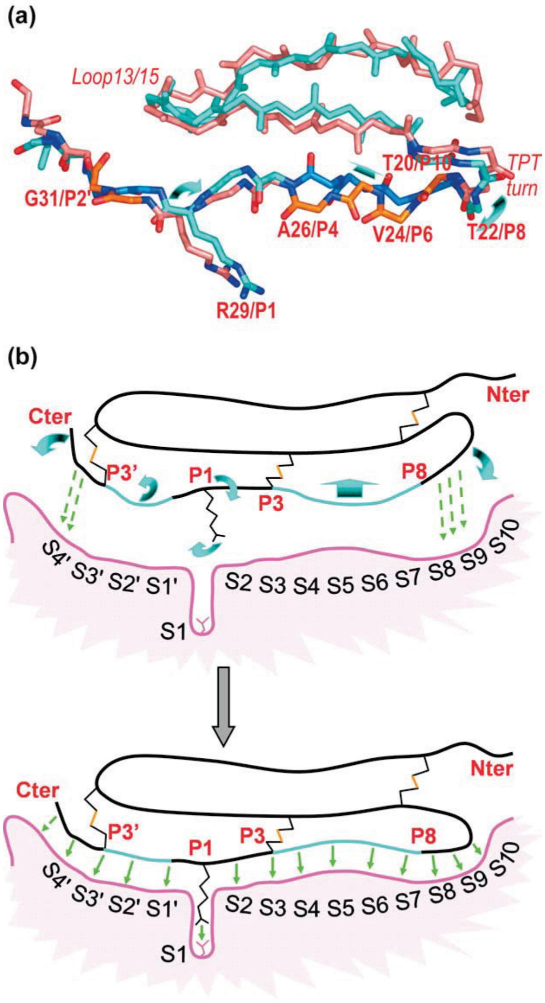


Figure 6. Shape adaptation upon binding of arthropod trypsin inhibitor SGTI to the surface of crayfish trypsin. (a) The structural alignment of the free (rose) and bound (light blue) forms of SGTI reveals three regions of major backbone conformation difference: the N-terminal segment (not shown), residues 20–26 (P_{10}–P_4) and residue 31 (P_2'). The latter two are parts of the binding region (O and N atoms shown in atomic colors). Carbon atoms of residues in the P_7–P_4 and P_2' regions are colored orange and dark blue for the free and bound form of SGTI, respectively. (b) Cartoon of SGTI binding to the enzyme. SGTI is shown in light blue (segments of the binding region with different backbone conformations in the free and bound form) and black (remaining parts). Cysteine and P_1 arginine side-chains of SGTI are shown, while some of its sub-sites are labeled in red. The enzyme surface is shown in magenta with black labels for the substrate binding sub-sites. Upper panel: conformation of the free form is preformed to recognize the S_{12}–S_8 and S_4'–S_5' sub-sites of the enzyme (shown as broken green arrows). In regions P_{10}–P_4 and P_2' conformation changes should occur, causing the rotation of the P_3–P_1' and P_4'–P_5' as well (light blue arrows). Lower panel: these conformational changes facilitate the build-up of an extended interaction network between SGTI and the enzyme (green arrows) in the complex.

the largest ones). However, considering the NOE-derived restraints between these two segments, a number of these are significantly (>0.5 Å) violated in the complex (Table 2). This indicates that these parts exhibit noticeable displacement with respect to each other upon protease binding.

Another important structural difference between the complexed and the free forms of SGTI is found in the 24–27 (P_6–P_4) region. This significant folding alteration is revealed by the values of φ, ψ angles of the complex structure. These angles are outside the entire folding range determined by the NMR ensemble.

Regarding the P_2–P_5' region, the folding similarity of Thr28–Arg29 dipeptide (P_2–P_1), especially the ψ angle of Thr28 and φ angle of Arg29 is very high. Nevertheless, Arg29 (P_1) is positioned farther from loop 13–15 in the complex than it is in the free form (Table 2). The C-terminal region ($31C^{\alpha}$–$34C^{\alpha}$) shows similar local conformations in the free and the bound form as well; however, the orientation of this unit is changed significantly upon complexation as it rotates around Gly31 (P_2'). Investigating the hydrogen bond system of the inhibitor we may conclude that H-bonds between strands 2 and 3 are well formed and shorter, therefore more stable in the complex than they are in the free form.

Discussion

The strength of interaction between crayfish trypsin and SGTI

SGTI, a protease inhibitor isolated from the haemolymph of desert locust, *S. gregaria*, is structurally homologous to the potent chymotrypsin inhibitor, SGCI, isolated from the same source. SGTI, however, contains an arginine residue instead of leucine at its P_1 site.[12] Despite this structural feature, which is favorable for trypsin inhibition, SGTI inhibits bovine trypsin relatively weakly, with an inhibitory constant (K_i) of 2.2×10^{-7} M. This value is five orders of magnitude larger than the equilibrium inhibitory constant of 6.2×10^{-12} M of SGCI determined on bovine chymotrypsin.[12,15] Unexpectedly, SGTI was found to be a potent inhibitor of crayfish trypsin with an eqilibrium K_i value of 0.7×10^{-12} M.[15]

Thermal denaturation experiments with crayfish trypsin–SGTI and bovine trypsin–SGTI complexes presented here have confirmed the results of the inhibition kinetics studies. As seen in Figure 1, the thermal stability of crayfish trypsin is increased dramatically upon binding of SGTI, resulting in a 27 °C higher melting temperature while the addition of SGTI to bovine trypsin led to only a 4.5 °C increase of T_m (Figure 1). The results are in line with the phylum specificity of SGTI and suggest that the observed difference in the K_i values may be realized in a thermodynamic stability difference in the enzyme–inhibitor interactions. We observed an increase in the peak areas, i.e. an

Table 1. Crystallographic data and refinement statistics

Resolution (Å)[a]	32.1–1.20 (1.26–1.20)		
Space group	$P2_12_12_1$		
Cell parameters (Å)	$\alpha=\beta=\gamma=90°$ $a=41.28$ $b=59.67$ $c=97.30$		
Number of observed reflections	878,624		
Number of unique reflections	93,027		
Completeness (%)[a]	91.2 (59.0)		
Mosaicity (°)	0.6		
$<I/\sigma>$[a]	13.1 (2.6)		
R_{merge} (%)[a,b]	5.9 (20.2)		
R_{work} (%)[c]	13.9		
R_{free} (%)[c]	18.2		
r.m.s. bond length (Å)	0.013		
r.m.s. bond angles (°)	2.242		
No. of non-hydrogen atoms			
Protein	1983		
Solvent	379		
Average B-factors (Å^2)	Protein	Water molecules	Overall
Main-chain	12.8±5.6		
Side-chain	15.7±7.7		
All	14.2±6.9	27.8±10.9	16.1±8.9
Anisotropy	0.43±0.14	0.54±0.17	0.44±0.15
Average estimated coordinate errors (esds) (pm)	5.3±3.9	9.7±6.1	5.9±4.6

[a] Values in parentheses indicate statistics for the highest resolution shell.

[b] $R_{\text{merge}}=\Sigma|I_o-\langle I\rangle|/\Sigma I_o\times 100\%$, where I_o is the observed intensity of a reflection and $\langle I\rangle$ is the average intensity obtained from multiple observations of symmetry related reflections.

[c] R factor$=\Sigma||F_{\text{obs}}|-|F_{\text{calc}}||/\Sigma|F_{\text{obs}}|\times 100\%$.

increased calorimetric enthalpy change of unfolding of the complexes compared to that of the single enzymes. This may be an outcome of intermolecular (enzyme–inhibitor) interactions rather than the simultaneous unfolding of SGTI, since the inhibitor alone shows no unfolding transitions up to 120 °C (not shown in Figure 1). In a previous study we pointed out the importance of the interdomain interactions in the function and structural stability of pancreatic serine proteases.[22] The X-ray structure of the complex revealed extensive interactions of SGTI with both β-barrel domains of crayfish trypsin, which may analogously explain the observed dramatic increase of thermal stability.

Based on the 3D structures of the crayfish trypsin–SGTI and the superimposed, energy minimized bovine trypsin–SGTI complexes, intermolecular interaction energies between the enzymes and SGTI were calculated (Figure 5(a)). Affinity of bovine trypsin to SGTI was found to be $\Delta\Delta G_b=-6.94$ kcal/mol smaller than affinity of crayfish trypsin. This structure-based free energy calculation is consistent with experimental stability difference between the investigated SGTI–trypsin complexes (6.6 kcal/mol) as converted from the inhibition constants.[12]

SGTI–crayfish trypsin interactions are extended over region P_{12}–P_5'

The crayfish trypsin–SGTI structure shows that the binding interface extends well beyond the primary binding region on both sides. Moreover, our energetic calculations reveal that all of these three regions are involved in stronger binding of SGTI to crayfish than to bovine trypsin (Figure 5(a) and (b)).

P_3–P_3': more favorable binding by the crayfish enzyme

Molecular dynamics (MD) simulation showed that the P_1–S_1 interaction is weaker in the bovine trypsin–SGTI complex than it is in the crayfish

Table 2. Some NMR restraints of SGTI and their violations in the SGTI–crayfish trypsin complex

Residue 1	Atom name	Residue 2	Atom name	NOE restraint (Å)	Violation (Å)
Cys4	H^{α}	Gly23	$H^{\alpha 1}$	5	1.2
Cys4	H^{α}	Gly23	H_N	5	2.6
Thr5	H^{β}	Thr22	H_N	5	5.6
Thr5	$H^{\gamma 2\#}$	Thr22	H^{α}	5	5.5
Thr5	$H^{\gamma 2\#}$	Thr22	H_N	5	5.3
Thr5	H^{β}	Gly23	H_N	5	3.6
Thr5	H_N	Gly23	$H^{\alpha 2}$	5	0.5
Asn15	$H^{\beta 1}$	Arg29	H^{α}	5	0.5

$H^{\gamma 2\#}$ is the pseudo-atom used for the γ-methyl group of Thr5.

trypsin–SGTI complex. In the former one, the distance between carboxylate 189 and the P_1 guanidino group is significantly longer during MD trajectory (Supplementary Data; Figure 2). A possible reason for the different behavior of the arginine side-chain is the looser fit of some neighboring sub-sites in bovine trypsin.

It was proposed in a previous study of ours that crayfish trypsin prefers positively charged residues at the P_1' position while bovine trypsin requires a neutral side-chain at the corresponding site.[13] Changing the P_1' residue lysine to methionine caused a one order of magnitude decrease in the inhibitory constant of SGTI on bovine trypsin. This result alludes to the importance of the P_1'–S_1' interaction for the determination of phylum specificity of SGTI (Figure 4(a)). Our MD study on the bovine trypsin–SGTI complex confirms this previous hypothesis regarding the conformation of Lys60 that was stabilized at the bottom of the S_1' cavity during the simulation. This residue is surrounded by hydrophobic residues and a positively charged one in the S_1' pocket in bovine trypsin, which is not favorable for adequate binding of SGTI. In contrast, the crayfish trypsin–SGTI structure shows that the broad S_1' cavity is more suitable for the positively charged P_1' lysine because it interacts with two negatively charged residues, Asp60b and Glu35 of the crayfish enzyme. MD shows high flexibility of Asp60b while Glu35 is stabilized at about 5 Å distance from the Lys30 N^{ζ} atom of SGTI. In the crayfish enzyme Glu35 seems to be especially important in respect of inhibitor binding, and it also stabilizes the Loop37 region that extends the primary binding region. The dual role of this residue ensures shaping of the binding surface in the primary binding region.

Hydrophobic binding patch at the S4′–S5′ region

A cluster of aromatic residues became inserted in Loop37 of arthropod trypsins. This cluster forms the binding surface for the Pro33–Pro34 (P_4'–P_5') region of SGTI (Figure 4(b)). MD shows that the main-chain conformation of this extended Loop37 and the three C-terminal residues of SGTI undergo only minor changes, and the movement of these two surface regions defined by Loop37 and the C-terminal residues of SGTI is restricted. The phenylalanine cluster interacts with the proline residues of SGTI; an alternative stacking interaction can be established by Phe37 and Pro34 (P_5') or Phe39 and Pro33 (P_4'). In the bovine trypsin–SGTI complex the weak interaction between Pro33 and Tyr39 is well maintained during simulation. This P_4'–P_5' region has the same conformation in all known complex structures of SGTI related peptides[10] while the mobility of these motifs in solution is relatively high. The phenylalanine cluster of crayfish trypsin might play a role in pre-orienting the inhibitor for the recognition of its C-terminal hydrophobic segment.

P_{12}–P_6 region stabilized by a network of hydrogen bonds

Both the interaction energy calculations and visual analysis of the contact interactions in trypsin–SGTI complexes have shown that the P_{12}–P_6 region forms significantly more favorable interactions with the crayfish enzyme (Figure 5(a) and (b)). The interactions in the crystal structure between P_{12}–P_6 and the 171–175 region of crayfish trypsin (Figure 4(c)) are practically unchanged in the MD trajectory, suggesting that these interactions may also be stable in solution.

Kellenberger and co-workers[16] proposed a hypothesis that the P_{10}–P_6 region of pacifastin-type trypsin inhibitors has an important role in its phylum selectivity. They suggested that the binding of P_{10}–P_6 to vertebrate trypsins is unfavorable because of the steric clash with Pro173 of these enzymes. Our present study confirms that Pro173 is indeed a key determinant for the binding difference, but rather than introducing an unfavorable steric effect it is disrupting the helical conformation at the 172–173 region of the enzyme. The key factor that determines the selective binding to trypsins of inferior or superior species is the molecular recognition of the C-terminal end of the 164–173 helix backbone. Trypsins with one residue insertion in the 164–175 region and glycine in position 173 (Figure 2), such as crayfish or *Fusarium oxysporum* trypsins, are likely to form a helix with a backbone conformation suitable for SGTI binding *via* a hydrogen bond network. Vertebrate trypsins have shorter loops in this region and Pro173 also breaks the helix. Our present molecular dynamics study also supports that Pro173 forms a stacking interaction with the Pro21 (P_9) residue of SGTI. These observations, when taken together, provide a circumstantial explanation of why vertebrate enzymes can form only less favorable interactions with SGTI.

SGTI binding to trypsin; anchor points and conformational adaptation

A comparison of the free and bound forms of SGTI reveals a conformational change upon binding to the enzyme (Figure 6(a)) that facilitates the emergence of an extended and strong interaction network. Local conformation of both P_{12}–P_4 and P_4'–P_5' regions of the inhibitor shows significant changes upon binding, suggesting that either or both of these regions may act as additional molecular recognition sites. The number of NMR distance (NOE) restraints (322 in total, ∼10/residue; 123 long-range) provides adequate information to establish the overall conformation of the inhibitor in solution. In loop regions, however, short-range NOEs dominate and determine the local structure (e.g. 20–22 TPT). This is in agreement with the generalized order parameters (S^2), where those of T20 and T22 are somewhat higher than those belonging to their close vicinity, indicating that the β-turn is likely to be involved in flip-flop type motions although not directly detected on the μs–ms time scale. This is consistent with the scarcity of long-range NOEs in this region despite of the short inter-atom distances in the solution structure. The suggested movement is strongly supported by the increased distance of the T20–G23 part and the N terminus clearly detectable as NOE restraint violations (Table 2). The P_6–P_4 region moves towards β-strand 2 of the inhibitor (the atom–atom distances become shorter and are consistent with the NOE experimental data; thus, this type of motion cannot be detected as restraint violation), influencing also local conformational preference, now forming a strong inter-strand H-bond network and adopting a conformation assuring a perfect match with the enzyme surface. As a consequence, the P_1 arginine residue is forced into the S_1 pocket of the enzyme and the P_1' lysine residue rotates into its binding groove. The C-terminal region preserves its conformation while it rotates around Gly31 (P_1') making a close fit with the enzyme surface (Figure 6(b)).

The extension and plasticity of crayfish trypsin–SGTI interaction offers new avenues for inhibitor specificity engineering

A great wealth of knowledge has been collected on the highly specific functions of serine proteases in the living organism. Development of computer-aided protein engineering opens a new interdisciplinary route for the design of specific inhibitors for therapeutic use. Although there are a large number of efficient small molecule serine protease inhibitors, they are not sufficiently specific and often too toxic for medical use. The most selective and potent inhibitors provided by nature are either oligopeptides or proteins. Canonical or standard mechanism inhibitors represent an important subset of these protein protease inhibitors.[6,7] A common structural feature of this class of inhibitors is that they have a reactive peptide bond in a loop that binds to the protease in a standard manner, and that this loop of six to nine residue long (also called primary binding region) has a more or less similar conformation in inhibitors and also in the enzyme–inhibitor complexes.[7] The uniform structure and homologous binding mode of these loops, however, may not provide these inhibitors with an extreme selectivity of their action. The ovomucoid third domain is a good example of a typical serine protease inhibitor with a relatively broad specificity.[23] In the complex of human leucocyte elastase with ovomucoid third domain the interaction is extended to the P_5–P_3'region of the inhibitor. Regading molecular movements upon binding, only the N-terminal region shows minor movements, the binding loop preserves its conformation.[24] However, some protease inhibitors, in addition to their typical primary binding loops, possess secondary binding sites as well. Hirudin, the most active and specific natural thrombin inhibitor uses an even more sophisticated mode of binding; out of its 65 residues 27 directly interact with thrombin.[25] The examples of hirudin and some recently developed two-binding site protease inhibitors of Factor VII show that new binding sites of an inhibitor tremendously increase its specificity and strength of interaction with the target protease.[26,27] As our present study shows, SGTI uses an inhibition strategy somewhat different from the inhibitors described above. In complex with crayfish trypsin, SGTI exhibits, instead of more than one distinct binding site, more or less continuous contacts in an extended region (through sites P_{12}–P_5') of the molecule (Figure 5). Some of these contacts result from a conformational change of SGTI that was induced by its binding to the enzyme (Figure 6). This is strongly supported by the precise comparison of the atomic resolution crystal structure of the crayfish trypsin–SGTI complex with that of uncomplexed SGTI. In contrast to other serine protease–protein inhibitor complexes where secondary interactions are mostly van der Waals contacts and do not affect specificity,[28] our present study shows that the extension of the binding surface leads to an increased specificity and stabilization of the complex. The comparison of the complexes of bovine and crayfish trypsin with SGTI shows that more than half of the interaction energy difference originates from the differential binding of the extended regions in the two complexes (Figure 5). The high resolution structure presented here provides a good basis for further study of the structural aspects of protease inhibitor specificity and to introduce new interaction sites into the inhibitor to increase its specificity towards proteases of interest.

Materials and Methods

DNA preparation, amplification and sequencing

Crayfish trypsin mRNA was obtained from *A. leptodactylus* hepatopancreas. Tissue (100 mg) was homogenized in 1 ml TRI-REAGENT (Sigma Chemical Co., Hungary) and RNA isolated according to the protocol of the manufacturer. Due to the known complete sequences of species *A. fluviatilis* and *Pacifastacus leniusculus* that are relatives of *A. leptodactylus* we could design oligonucleotide primers for amplifying the coding region of crayfish trypsin. RT-PCR was performed using the following primers: CFT3′: 5′-GGAGCTCAGACTGCATTTGCTTTGAT-3′, CFT5′: CCGAAGCTTTTCCCGTGGATGATGATGACAAGATCGTTGGTG. These primers include a HindIII site at the 5′ end and a SacI site at the 3′ end. Additionally, since the propeptide sequence of crayfish trypsin is unknown we attached a rat trypsin propeptide sequence at the 5′ region. The amplified DNA was cloned into a pET17b vector. The sequence of the chimeric trypsin was determined by automated dideoxy sequencing (ABI Prism) using the Big Dye Terminator Kit (GenBank accession no. AY906961). Since our experiments that were aimed at producing recombinant crayfish trypsin yielded only a low amount of enzyme, for the crystallization procedure natural crayfish trypsin was used.

Isolation and characterization of crayfish trypsin

Narrow-clawed crayfish (*A. leptodactylus*) trypsin was purified by the procedure as described by Zwilling *et al.*,[29] with some modifications. A stock of about 500 crayfish cardia fluid was collected by introducing a teflon capillary tube attached to a syringe into the cardia of the animal. The collected cardia fluid was ultrafiltrated on AMICON (AMICON Corp., Beverly, MA, USA) membranes with 50 kDa and 10 kDa cut-off. The fraction between 10 kDa and 50 kDa was loaded onto a CNBr-Sepharose-4B soybean-trypsin inhibitor column. The column was washed with three volumes of distilled water and then eluted with dilute NH_3 solution (pH 11.0). Fractions containing crayfish trypsin were pooled and loaded on a MONO Q (Pharmacia, Sweden) ion exchange column equilibrated with 10 mM Mes (pH 6.0) and eluted with a linear gradient of 0 M to 1 M NaCl. Fractions containing different forms of crayfish trypsin were collected, and checked by SDS-PAGE, 2D-SDS-gel electrophoresis and activity measurements. All izoenzymes of crayfish trypsin were found to be identical regarding their enzymatic activities and sensitivities to inhibition. For further study the most abundant form was chosen, and concentrated by ultrafiltration using Centricon-10 concentrators (AMICON Corp., Beverly, MA, USA).

Differential scanning calorimetry (DSC)

Calorimetric measurements were performed on a VP-DSC (MicroCal) differential scanning calorimeter. Equimolar mixtures of trypsin and SGTI were used for studying the effect of the inhibitor on bovine and crayfish trypsins. The protein concentration was set to 0.1 mg/ml. Samples were dialyzed against 20 mM sodium phosphate (pH 7.0), 100 mM NaCl, and the dialysis buffer was used as a reference. Denaturation curves were recorded between 10 °C and 120 °C at a pressure of 2.5 atm, using a scanning rate of 1 deg.C/minute. The thermal unfolding curves were analyzed using MicroCal Origin 7.0 software. We note that bovine trypsin and its complex exhibited additional minor components at lower temperatures, which is a consequence of heterogeinity, probably due to an autolysis product of the commercial enzyme sample.

Chemical synthesis of SGTI

The inhibitor was synthesized, oxidized and purified as described.[12]

Preparation of the crayfish trypsin–SGTI complex and its crystallization

A fourfold molar excess of SGTI was added to crayfish trypsin and incubated for 15 minutes at room temperature. The complex was loaded to a HiPrep S-100 gel filtration column (Amersham Biosciences, UK), and eluted with 10 mM Mes (pH 6.0). The pure crayfish trypsin–SGTI complex was collected and concentrated to 11 mg/ml. Crystals of the crayfish trypsin–SGTI complex were grown by the hanging drop method at 20 °C. Equal amount of protein solution (11 mg/ml protein in 10 mM Mes (pH 6.0)) and precipitant solution (30% (w/v) polyethylene glycol (PEG) 400, 0.1 M cadmium chloride, 0.1 M sodium acetate (pH 4.6)) were mixed and equilibrated against 0.5 ml of precipitant solution. Crystals were grown in two days.

X-ray diffraction studies

Two datasets were collected from a single crystal at ESRF on beamline ID 14 EH2 at cryogenic temperature (100 K). Crystallograpic intensities were integrated and scaled to a resolution of 1.2 Å using Mosflm[30] and Scala[31] of the CCP4 package V5.0.[32] Completeness of the data was 91.2% at 1.2 Å resolution. The structure was solved by molecular replacement using the program Molrep[33] from the CCP4 package. A polyalanine search model was used which was derived from the X-ray structure of human trypsin IV (PDB entry 1H4W).[34] The asymmetric unit contains one trypsin–inhibitor 1:1 complex. Automated model building was carried out with Arp/wArp.[35] The model was systematically improved using iterative cycles of manual rebuilding with the program O[36] and restrained least-squares refinement with SHELX.[37] Atomic *B*-factors were refined anisotropically and this step reduced the *R*-factor and R_{free}[38] values by 4.3% and 3.3%, respectively. Finally, all except for hydroxyl and His $N^{\varepsilon 2}$ and $N^{\delta 1}$ riding hydrogen atoms were added to the structure. The geometry of the P_2–P_2' residues of the inhibitor were not restrained in the final rounds of the refinement. The final model contains residues 16–244 (chymotrypsin numbering system[39]) of crayfish trypsin, and residues 2–34 of SGTI. The stereochemistry of the structure was assessed with Whatcheck[40] and PROCHECK.[41] The distribution of anisotropic *B*-factors was monitored with the program Parvati.[42] Data collection and refinement statistics are shown in Table 1.

Calculations

SGTI–bovine trypsin complex was derived from superposition of the structure of bovine trypsin (PDB entry 3PTB)[43] on the crayfish trypsin complex of the present study using the LSQMAN program[44] from Uppsala Software Factory (USF). The GROMACS[45] program package was applied for generation of a simulation box, addition of explicit water molecules and counter ions. Both complexes (SGTI–crayfish trypsin and SGTI–bovine trypsin) and SGTI alone were energy minimized using the GROMOS[46] force field implemented in the program package. The 100 ps long position restrained and 500 ps long unrestrained molecular dynamics simulations (MD) were performed to equilibrate the surrounding molecules and to generate conformations for calculations of intermolecular interaction energy (E_i).

Intermolecular energy terms of scoring function of AutoDock 3.0 program[47] were applied to 50 conformations of the complexes (sampled at every 10 ps of the 500 ps trajectories). Extra penalty constants of H-bonds were not used. Thus, scaled Coulombic, Lennard-Jones terms and the desolvation free energy term[48] were involved in calculation of E_i. Difference of the E_i-s is considered as $\Delta\Delta G_b$ (difference in free energy of binding) of the interactions of SGTI with the two trypsins. Preparation of trypsin and SGTI molecules and grid calculations were done as described in our previous studies[49,50] for each conformation using shell scripts. AMBER charges were applied for all molecules.

Acknowledgements

This work was supported by the Hungarian National Science Foundation (OTKA), grant TS044730, the ICGEB-HUN04-03 fund and an EU Improving Human Potential program, BBSRC grant. We thank Zoltán Gáspári for helpful discussions. We also thank Hanna-Kirsti Schrøder Leiros for the expert assistance at beamline ID14-EH2, ESRF. G.K. thanks Richard Neutze, Nigel S. Scrutton and David Leys for their support. C.H. is a Békésy Fellow of the Hungarian Ministry of Education. This paper is dedicated to Professor András Lipták and his outstanding scientific and professional activities on his 70th birthday.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.04.039

References

1. Kühne, W. (1867). Über die Verdaung der Eiweistoffe durch den Pankreassaft. *Virchows Arch.* **39**, 130–174.
2. Kunitz, M. & Northrop, J. H. (1936). Isolation from beef pancreas of crystalline trypsinogen, trypsin, trypsin inhibitor and an inhibitor trypsin compound. *J. Gen. Physiol.* **19**, 991–1007.
3. Bajusz, S., Széll, E., Bagdy, D., Barabás, E., Horváth, G., Diószegi, M. *et al*. (1990). Highly active and selective anticoagulants: D-Phe-Pro-Arg-H, a free tripeptide aldehyde prone to spontaneous inactivation, and its stable N-methyl derivative, D-MePhe-Pro-Arg-H. *J. Med. Chem.* **33**, 1729–1735.
4. Hirsh, J., O'Donnell, M. & Weitz, J. I. (2005). New anticoagulants. *Blood*, **105**, 453–463.
5. Kaplan, K. L. & Francis, C. W. (2002). Direct thrombin inhibitors. *Semin. Hematol.* **39**, 187–196.
6. Bode, W. & Huber, R. (1991). Proteinase–protein inhibitor interaction. *Biomed. Biochim. Acta*, **50**, 437–446.
7. Laskowski, M. & Qasim, M. A. (2000). What can the structures of enzyme–inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim. Biophys. Acta*, **1477**, 324–337.
8. Schechter, I. & Berger, A. (1968). On the site of the active site in proteases. *Biochem. Biophys. Res. Commun.* **27**, 157–162.
9. Boigegrain, R. A., Mattras, H., Brehelin, M., Paroutaud, P. & Coletti-Preverio, M. A. (1992). Insect immunity: two proteinase inhibitors from hemolymph of *Locusta migratoria*. *Biochem. Biophys. Res. Commun.* **189**, 790–793.
10. Kellenberger, C., Boudier, C., Bermudez, I., Bieth, J. G., Luu, B. & Hietter, H. (1995). Serine protease inhibition by insect peptides containing a cysteine knot and a triple-stranded beta-sheet. *J. Biol. Chem.* **270**, 25514–25519.
11. Hamdaoui, A., Schoofs, L., Wateleb, S., Bosch, L. V., Verhaert, P., Waelkens, E. & De Loof, A. (1997). Purification of a novel, heat-stable serine protease inhibitor protein from ovaries of the desert locust, *Schistocerca gregaria*. *Biochem. Biophys. Res. Commun.* **238**, 357–360.
12. Malik, Z., Amir, S., Pál, G., Buzás, Z., Várallyay, E., Antal, J. *et al*. (1999). Proteinase inhibitors from desert locust, *Schistocerca gregaria*: engineering of both P(1) and P(1)′ residues converts a potent chymotrypsin inhibitor to a potent trypsin inhibitor. *Biochim. Biophys. Acta*, **1434**, 143–150.
13. Gáspári, Z., Patthy, A., Gráf, L. & Perczel, A. (2002). Comparative structure analysis of proteinase inhibitors from the desert locust, *Schistocerca gregaria*. *Eur. J. Biochem.* **269**, 527–537.
14. Szenthe, B., Gáspári, Z., Nagy, A., Perczel, A. & Gráf, L. (2004). Same fold with different mobility: backbone dynamics of small protease inhibitors from the desert locust, *Schistocerca gregaria*. *Biochemistry*, **43**, 3376–3384.
15. Patthy, A., Amir, S., Malik, Z., Bódi, A., Kardos, J., Asbóth, B. & Gráf, L. (2002). Remarkable phylum selectivity of a *Schistocerca gregaria* trypsin inhibitor: the possible role of enzyme-inhibitor flexibility. *Arch. Biochem. Biophys.* **398**, 179–187.
16. Kellenberger, C., Ferrat, G., Leone, P., Darbon, H. & Roussel, A. (2003). Selective inhibition of trypsins by insect peptides: role of P6-P10 loop. *Biochemistry*, **42**, 13605–13612.
17. Roussel, A., Mathieu, M., Dobbs, A., Luu, B., Cambillau, C. & Kellenberger, C. (2001). Complexation of two proteic insect inhibitors to the active site of chymotrypsin suggests decoupled roles for binding and selectivity. *J. Biol. Chem.* **276**, 38893–38898.
18. Helland, R., Otlewski, J., Sundheim, O., Dadlez, M. & Smalas, A. O. (1999). The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI. *J. Mol. Biol.* **287**, 923–942.
19. Katona, G., Wilmouth, R. C., Wright, P. A., Berglund, G. I., Hajdu, J., Neutze, R. & Schofield, C. J. (2002). X-ray structure of a serine protease acyl-enzyme complex at 0.95-Å resolution. *J. Biol. Chem.* **277**, 21962–21970.
20. Bode, W., Turk, D. & Karshikov, A. (1992). The refined 1.9-A X-ray crystal structure of D-Phe-Pro-Arg chloromethylketone-inhibited human alpha-thrombin: structure analysis, overall structure, electrostatic properties, detailed active-site geometry, and structure-function relationships. *Protein Sci.* **1**, 426–471.
21. Roach, J. C., Wang, K., Gan, L. & Hood, L. (1997). The molecular evolution of the vertebrate trypsinogens. *J. Mol. Evol.* **45**, 640–652.
22. Kardos, J., Bódi, A., Závodszky, P., Venekei, I. & Gráf, L. (1999). Disulfide-linked propeptides stabilize the structure of zymogen and mature pancreatic serine proteases. *Biochemistry*, **38**, 12248–12257.
23. Lu, W., Apostol, I., Quasim, M. A., Warne, N., Wynn, R., Zhang, W. L. *et al*. (1997). Binding of amino acid side-chains to S1 cavities of serine proteinases. *J. Mol. Biol.* **266**, 441–461.
24. Bode, W., Wei, A. Z., Huber, R., Meyer, E., Travis, J. & Neumann, S. (1986). X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J.* **5**, 2453–2458.
25. Grutter, M. G., Priestle, J. P., Rahuel, J., Grossenbacher, H., Bode, W., Hofsteenge, J. & Stone, S. R. (1990). Crystal structure of the thrombin-hirudin complex: a novel mode of serine protease inhibition. *EMBO J.* **9**, 2361–2365.

26. Lee, G. F., Lazarus, R. A. & Kelley, R. F. (1997). Potent bifunctional anticoagulants: Kunitz domain-tissue factor fusion proteins. *Biochemistry*, **36**, 5607–5611.
27. Roberge, M., Peek, M., Kirchhofer, D., Dennis, M. S. & Lazarus, R. A. (2002). Fusion of two distinct peptide exosite inhibitors of Factor VIIa. *Biochem. J.* **363**, 387–393.
28. Bode, W. & Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *Eur. J. Biochem.* **204**, 433–451.
29. Zwilling, R., Pfleiderer, G., Sonneborn, H.-H., Kratz, V. & Stucky, I. (1969). The evolution of endopeptidases-V. Common and different traits of bovine and crayfish trypsin. *Comp. Biochem. Physiol.* **28**, 1275–1287.
30. Leslie, A. G. W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EAMCB Newsletter on Protein Crystallography*, **26**, 1–9.
31. Evans, P. R. (1997). SCALA. *Joint CCP4 and ESF-EAMBC Newsletter on Protein Crystallography*, **33**, 22–24.
32. CCP4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallog. sect. D*, **50**, 760–763.
33. Vagin, A. & Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *J. Appl. Crystallog.* **30**, 1022–1025.
34. Katona, G., Berglund, G. I., Hajdu, J., Gráf, L. & Szilágyi, L. (2002). Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. *J. Mol. Biol.* **315**, 1209–1218.
35. Lamzin, V. S. & Wilson, K. S. (1993). Automated refinement of protein models. *Acta Crystallog. sect. D*, **49**, 129–147.
36. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A*, **47**, 110–119.
37. Scheldrick, G. & Schneider, T. (1997). SHELXL: High-resolution refinement. *Methods Enzymol.* **277**, 319–343.
38. Brunger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–475.
39. Hartley, B. S. & Kauffman, D. L. (1966). Corrections to the amino acid sequence of bovine chymotrypsinogen A. *Biochem. J.* **101**, 229–231.
40. Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). Errors in protein structures. *Nature*, **381**, 272.
41. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.
42. Merritt, E. A. (1999). Expanding the model: anisotropic displacement parameters in protein structure refinement. *Acta Crystallog. sect. D*, **55**, 1109–1117.
43. Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallog. sect. B*, **39**, 480.
44. Madsen, D. & Kleywegt, G. J. (2002). Interactive motif and fold recognition in protein structures. *J. Appl. Crystallog.* **35**, 137–139.
45. Lindahl, E., Hess, B. & van der Spoel, D. (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **7**, 306–317.
46. Stocker, U. & van Gunsteren, W. F. (2000). Molecular dynamics simulation of hen egg white lysozyme: a test of the GROMOS96 force field against nuclear magnetic resonance data. *Proteins: Struct. Funct. Genet.* **40**, 145–153.
47. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and and empirical binding free energy function. *J. Comp. Chem.* **19**, 1639–1662.
48. Stouten, P. F. W., Frömmel, C., Nakamura, H. & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.* **10**, 97–120.
49. Hetényi, C. & van der Spoel, D. (2002). Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **11**, 1729–1737.
50. Hetényi, C., Maran, U. & Karelson, M. (2003). A comprehensive docking study on the selectivity of binding of aromatic compounds to proteins. *J. Chem. Inf. Comput. Sci.* **43**, 1576–1583.
51. DeLano, W. L. (2002). *PyMOL Molecular Graphics System*, DeLano Scientific, San Carlos CA, USA.
52. Kraulis, Per J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by R. Huber

(Received 10 February 2005; received in revised form 8 April 2005; accepted 12 April 2005)

D21

J|A|C|S
ARTICLES
Published on Web 01/06/2006

Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins

Csaba Hetényi,*,† Gábor Paragi,‡ Uko Maran,§ Zoltán Timár,|| Mati Karelson,§ and Botond Penke‡,||

Contribution from the Department of Biochemistry, Eötvös Loránd University, 1/C Pázmány P. sétány, H-1117 Budapest, Hungary, Protein Chemistry Research Group, Hungarian Academy of Sciences, 8 Dóm tér, H-6720 Szeged, Hungary, Department of Chemistry, Tartu University, 2 Jakobi Street, EE−51014 Tartu, Estonia, and Department of Medical Chemistry, University of Szeged, 8 Dóm tér, H-6720 Szeged, Hungary

Received September 1, 2005; E-mail: csabahete@yahoo.com

Abstract: Bulky, flexible molecules such as peptides and peptidomimetics are often used as lead compounds during the drug discovery process. Pathophysiological events, e.g., the formation of amyloid fibrils in Alzheimer's disease, the conformational changes of prion proteins, or β-secretase activity, may be successfully hindered by the use of rationally designed peptide sequences. A key step in the molecular engineering of such potent lead compounds is the prediction of the energetics of their binding to the macromolecular targets. Although sophisticated experimental and in silico methods are available to help this issue, the structure-based calculation of the binding free energies of large, flexible ligands to proteins is problematic. In this study, a fast and accurate calculation strategy is presented, following modification of the scoring function of the popular docking program package AutoDock and the involvement of ligand-based two-dimensional descriptors. Quantitative structure−activity relationships with good predictive power were developed. Thorough cross-validation tests and verifications were performed on the basis of experimental binding data of biologically important systems. The capabilities and limitations of the ligand-based descriptors were analyzed. Application of these results in the early phase of lead design will contribute to precise predictions, correct selections, and consequently a higher success rate of rational drug discovery.

Introduction

Flexible, peptidic molecules are often involved in rational drug design. These compounds find various applications for important biochemical problems such as the inhibition of β-secretase,[1] a key enzyme in the pathomechanism of Alzheimer's disease,[1] or the blocking of various types of trypsins.[2] Similarly, the beta sheet breaker peptides have proved useful in hindering self-aggregation of the β-amyloid peptide of Alzheimer's disease and conformational changes in prion proteins of transmissible spongiform encephalophaties.[3] The number of such relevant applications of peptides as potent bioactive partners or lead compounds is still increasing. In rational drug discovery, estimation of the free energies of binding (ΔG_b) of bioactive ligands to their macromolecular targets is an essential step in the molecular engineering process. Although sophisticated methods do exist for the experimental measurement of binding thermodynamics (e.g., isothermal titration calorimetry[4]), they are usually time-consuming and/or require special conditioning for problematic cases such as amyloid aggregation.[5]

Different in silico strategies for the structure-based calculation[6] of ΔG_b have become an alternative to the instrumental techniques. One branch of these computational methods works on a statistical ensemble of structures produced by a molecular dynamics (MD) simulation. The MD-based techniques, e.g., the linear interaction energy method[7] supported by perturbation theory,[8] have been successfully applied to modified peptides,[9]

[†] Eötvös Loránd University.

[‡] Hungarian Academy of Sciences.

[§] Tartu University.

[||] University of Szeged.

(1) (a) Hong, L.; Koelsch, G.; Lin, X.; Wu, S.; Terzyan, S.; Ghosh, A. K.; Zhang, X. C.; Tang, J. *Science* **2000**, *290*, 150−153. (b) Ghosh, A. K.; Shin, D.; Downs, D.; Koelsch, G.; Lin, X.; Ermolieff, J.; Tang, J. *J. Am. Chem. Soc.* **2000**, *122*, 3522−3523. (c) John, V.; Beck, J. P.; Bienkowski, M. J.; Sinha, S.; Heinrikson, R. L. *J. Med. Chem.* **2003**, *46*, 4625−4630.

(2) Fodor, K.; Harmat, V.; Hetényi, C.; Kardos, J.; Antal, J.; Perczel, A.; Patthy, A.; Katona, G.; Gráf, L. *J. Mol. Biol.* **2005**, *350*, 156−169.

(3) (a) Soto, C.; Sigurdsson, E. M.; Morelli, L.; Kumar, R. A.; Castaño, E. M.; Frangione, B. *Nature Med.* **1998**, *4*, 822−826. (b) Soto, C.; Kascsak, R. J.; Saborio, G. P.; Aucouturier, P.; Wisniewski, T.; Prelli, F.; Kascsak, R.; Mendez, E.; Harris, D. A.; Ironside, J.; Tagliavini, F.; Carp, R. I.; Frangione, B. *Lancet* **2000**, *355*, 192−197. (c) Hetényi, C.; Körtvélyesi, T.; Penke, B. *Bioorg. Med. Chem.* **2002**, *10*, 1587−1593. (d) Hetényi, C.; Szabó, Z.; Klement, É.; Datki, Z.; Körtvélyesi, T.; Zarándi, M.; Penke, B. *Biochem. Biophys. Res. Commun.* **2002**, *292*, 931−936. (e) Dobson, C. M. *Nature* **2005**, *435*, 747−749.

(4) (a) Leavitt, S.; Freire, E. *Curr. Opin. Struct. Biol.* **2001**, *11*, 560−566. (b) Campoy, A. V.; Freire, E. *Biophys. Chem.* **2005**, *115*, 115−124.

(5) Kardos J.; Yamamoto, K.; Hasegawa, K.; Naiki, H.; Goto Y. *J. Biol. Chem.* **2004**, *279*, 55308−55314.

(6) (a) Murphy, K. P. *Med. Res. Rev.* **1999**, *19*, 333−339. (b) Lazaridis, T. *Curr. Org. Chem.* **2002**, *6*, 1319−1332.

(7) (a) Åqvist, J. *J. Comput. Chem.* **1996**, *17*, 1587−1597. (b) Marelius, J.; Hansson, T.; Åqvist, J. *Int. J. Quantum Chem.* **1998**, *69*, 77−88.

10.1021/ja055804z CCC: $33.50 © 2006 American Chemical Society

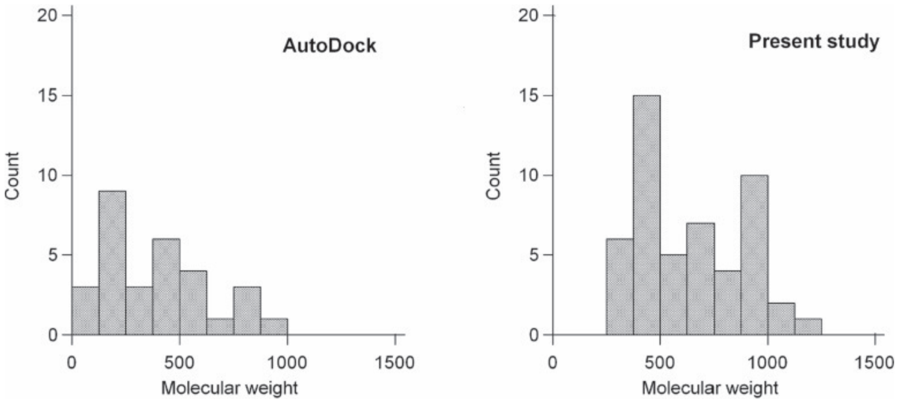


Figure 1. Distribution of molecular weights of the 30 ligands of the AutoDock calibration set[12] and the 50 compounds investigated in the present study. In the case of the present study, the number of compounds with higher molecular weights is significantly larger.

as well. Another strategy for the calculation of ΔG_b is the use of a single protein–ligand complex structure (preferably the crystallographic structure or an energy minimum). This approach requires a scoring function (SF), along with a parameter set appropriate for the type of ligand molecules investigated. The SFs developed for rapid calculation of ΔG_b are primarily implemented to drive the docking simulations.[10] In most of the cases they are parametrized for different types of small, druglike compounds to fit the requirements of the virtual high-throughput screening of compound libraries. It has been demonstrated in a number of studies that the crystallographic ligand positions in the protein–ligand complexes can be calculated precisely by using the appropriate SFs.[11] As SFs have been successfully used in calculations on various small compounds, it is a rational (but not trivial) wish to extend their applicability to larger, flexible ligands.

In the present study, the SF of the popular docking program package AutoDock 3.0[12] is tested and modified by using a set of flexible, peptidic ligands of biologically important complex systems. Predictive quantitative structure–activity relationships (QSARs) are developed for experimental ΔG_b values, using the modified SF of AutoDock and two-dimensional (2D) molecular descriptors of the ligand molecules. Our aim is to extend the capabilities of the SFs by means of easy-to-calculate ligand-based descriptors so as to develop a new, hybrid calculation strategy that combines advantages of the intermolecular terms of the SF and the ligand-based 2D descriptors for the rapid and accurate calculation of ΔG_b data for the problematic, bulky ligand molecules.

Methods

Protein–Ligand Systems. In the present study, 53 different protein–ligand complexes with known experimental values of ΔG_b ($\Delta G_{b(exp)}$) were involved. Complexes having large, peptidic ligands (MW > 350, Figure 1) and physiological importance (e.g., the "om"-series of β-secretase inhibitors; see Introduction for references on pathophysiological role of β-secretase) were prioritized for this study. Systems with di/tripeptide ligands were also selected to balance the structural data set. The atomic coordinates of 41 of the complexes, 1a30, 1abo, 1b05, 1b32, 1b3f, 1b3g,1b3l, 1b46, 1b51, 1b52, 1b58, 1b5i, 1b5j, 1b9j, 1bai, 1cka, 1fkn (om99-2), 1hhi, 1hhh, 1hhj, 1hhk, 1jet, 1jeu, 1jev, 1joj, 1k9r, 1m4h (om00-3), 1mcb, 1mcj, 1ody, 1qkb, 1str, 1vac, 1vwf, 2er9, 2rkm, 2vaa, 2vab, 4sga, 5sga, and 5er1 were obtained from the Protein Databank[13] (PDB). 12 β-secretase-inhibitor systems (om12, om13, om14, om15, om16, om17, om18, om19, om22, om23, om24, and om99-1)[14b] with no PDB structures available were modeled by modification of the 1fkn structure. $\Delta G_{b(exp)}$'s were compiled from previous studies.[14] Detailed data on the protein–ligand complexes and the corresponding codes are listed in the Supporting Information, Table A.

Molecular Modeling. The Babel,[15] Vega,[16] VMD,[17] and PyMol[18] packages were applied for file conversion, visualization, and modeling. Some of the GROMACS[19,20] topology files were generated with the program ProDrg.[21]

Molecular Mechanics Minimization. A standard routine was applied for all complexes to create a uniform set of coordinate files. The GROMACS program package and the force field[19,20] and explicit SPC[22] water model were involved in the calculations. The protein–ligand complexes and surrounding water molecules were placed in a cubic box together with the appropriate amount of neutralizing counterions. Dissociable protons were added by a built-in GROMACS algorithm, except for the β-secretase complexes, where the active site

[8] Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
[9] Hansson, T.; Åqvist, J. *Protein Eng.* **1995**, *8*, 1137–1144.
[10] (a) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409–443. (b) Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373. (c) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. *J. Med. Chem.* **2004**, *47*, 3032–3047.
[11] (a) Hetényi, C.; van der Spoel, D. *Protein Sci.* **2002**, *11*, 1729–1737. (b) Hetényi, C.; Maran, U.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1576–1583.
[12] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
[13] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
[14] (a) Donnini, S.; Juffer, A. H. *J. Comput. Chem.* **2004**, *25*, 393–411. (b) Ghosh, A. K.; Bilcer, G.; Harwood, C.; Kawahama, R.; Shin, D.; Hussain, K. A.; Hong, L.; Loy, J. A.; Nguyen, C.; Koelsch, G.; Ermolieff, J.; Tang, J. *J. Med. Chem.* **2001**, *44*, 2865–2868. (c) Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2977–2980. (d) Turner, R. T.; Koelsch, G.; Hong, L.; Castenheira, P.; Ghosh, A.; Tang, J. *Biochemistry* **2001**, *40*, 10001–10006.
[15] Walters, P.; Dolata, M. S. Babel – A Molecular Structure Information Interchange Hub. Department of Chemistry, University of Arizona, Tucson, AZ 85721.
[16] Pedretti, A.; Villa, L.; Vistoli, G. *J. Mol. Graph.* **2002**, *21*, 47–49.
[17] Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.
[18] DeLano, W. L. *PyMol Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002.
[19] Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model* **2001**, *7*, 306–317.
[20] Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Comm.* **1995**, *91*, 43–56.
[21] Schuttelkopf, A. W.; van Aalten, D. M. F. *Acta Crystallogr. D* **2004**, *60*, 1355–1363.
[22] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*. Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, 1981; pp 331–342.

was protonated according to the results of a recent study.[23] The systems were optimized with steepest descent and conjugate gradient methods at tolerance levels of 1000 and 600 kJ mol^{-1} nm^{-1} and maximum step sizes of 0.05 and 0.001 nm, respectively. The optimum coordinates of the protein and ligand molecules were extracted for the subsequent calculations. Whenever necessary (e.g., 1ody) the crystallographic water molecule was also extracted as an essential part of the active site of the protein.

Scoring. Grid maps of 120 × 120 × 120 grid points at a spacing of 0.375 Å were generated around the center of the ligand binding site by the utility Autogrid of the program package AutoDock 3.0.[12] Heavy atoms and polar H atoms of the protein molecules were supplied with Kollman's partial charges. Atomic solvation parameters and fragmental volumes were inserted via the utility Addsol.[12] Gasteiger charges[24] were assigned to the ligand molecules. Charges of apolar H atoms were merged with charges of the connecting C atoms and aromatic atoms were selected by the utility Autotors.[12] The free energies of binding of the ligands to the proteins were calculated by using the SF implemented in the program package AutoDock[12] (eq 1):

$$\Delta G_{\mathrm{AD}} = f_{\mathrm{elec}}\sum_{i,j}\frac{q_j q_i}{\epsilon(r_{ij})r_{ij}} + f_{\mathrm{vdw}}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + f_{\mathrm{hbond}}\sum_{i,j}\xi(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + f_{\mathrm{sol}}\sum_{i,j}S_i V_j \mathrm{e}^{(-\mathrm{r_{ij}}^2/2\sigma^2)} + T_{\mathrm{HBD}} + T_{\mathrm{TOR}}$$

where

$$T_{\mathrm{HBD}} = \sum_i P_{\mathrm{HBD},i};\ P_{\mathrm{HBD},i} = \begin{cases} 0.118\ \mathrm{kcal/mol} & \text{if atom}_i = \text{polar H (H in a polar covalent bond)} \\ 0.236\ \mathrm{kcal/mol} & \text{if atom}_i = \mathrm{O} \\ 0.000\ \mathrm{kcal/mol} & \text{if atom}_i \neq \text{polar H or O} \end{cases}$$

$$T_{\mathrm{TOR}} = P_{\mathrm{TOR}} N_{\mathrm{TOR}} \qquad (1)$$

ΔG_{AD} (the calculated AutoDock binding free energy) is the sum of three intermolecular interaction energy terms, one desolvational free energy term (these four terms are referred to as "bimolecular" in the next sections) and two "monomolecular" terms describing hydrogen-bonding (T_{HBD}) and torsional penalties (T_{TOR}) of the ligand molecule. It should be noted that the original formula of the AutoDock SF[12] is reorganized in eq 1 to make a distinction between the bimolecular and the ligand-based (monomolecular) terms.

The f coefficients were determined empirically from a multilinear regression (MLR) to a set of 30 protein−ligand complexes (AutoDock calibration set) with known binding constants.[12] The indices i and j correspond to ligand and protein atoms, respectively. The Coulombic term includes the partial charges (q) and a distance-dependent dielectric permittivity value (ϵ).[25] A, B, C, and D are the Lennard−Jones parameters in the dispersion/repulsion (12−6) and H-bonding (12−10) formulas, and r denotes the distance between the atomic pairs. $\xi(t)$ is a directional weight depending on angle t at the H-bonds.[12] T_{HBD} accounts for the broken H-bonds between the ligand and solvent molecules, and it is calculated by summation of the P_{HBD} penalty constants for the polar H or O atoms in the ligand molecule. In practice, these constants are added to the appropriate atomic affinity grid maps during calculation. The value of P_{HBD} for polar H atoms was derived[12] as $P_{\mathrm{HBD}} = 0.0656 \times 0.36 \times 5$ kcal/mol, where 0.0656 is f_{hbond}, the MLR coefficient, 0.36 is the proportion of H-bonding sites utilized on average, and 5 kcal/mol is the maximal well depth of the H-bonding interaction.[12] The constant P_{HBD} (for O atoms) is equal to 2 × P_{HBD} (for polar H's) counting for two possible H-bonds at O atoms. P_{TOR} has a constant (0.3113 kcal/mol) value per torsion. N_{TOR} is the number of free torsions in the ligand. The product (T_{TOR}) of P_{TOR} and N_{TOR} gives an estimate of the unfavorable torsional entropy loss upon ligand binding. S and V denote the solvation parameter and fragmental volume, respectively, in the solvation function of Stouten et al.[26] In the SF of AutoDock 3.0, only the C atoms of the ligand molecules are involved in the solvation model. The exponential term is an envelope function with a constant value[26] of $\sigma = 3.5$ Å. By elimination of T_{HBD}, T_{TOR}, or both terms, new, modified SFs (ΔG_{H}, ΔG_{T}, or ΔG_{TH}) are defined and applied in the present study.

Quantum Mechanics (QM) Calculations. At the ab initio level, the density functional method was used for calculation of the partial charges on the atoms of the ligand molecules.[27] The B3LYP functional and 6-311 basis set augmented with polarization functions were employed in the Gaussian98[28] calculations.

Development of Quantitative Structure−Activity Relationships (QSARs). The development and statistical analysis of the MLRs and the selection of 2D descriptors were achieved with the program package CODESSA (ver. 2.0).[29] The MLRs have the following general formula (eq 2):

$$\Delta G_{\mathrm{b(exp)},j} = \sum_{i=1}^{n} \alpha_i \mathrm{D}_{ji} + \text{constant}; \quad (j = 1, 2, ..., N) \qquad (2)$$

where i and j are the serial numbers of the descriptors and ligands, respectively, N is the total number of ligands (complex systems), n is the total number of descriptors involved in the model, D_{ji} denote the descriptors, and α_i's are the regression coefficients. The mean square errors and t-values of the regression coefficients, the F-values, the standard deviations (s^2), and the squares of the correlation coefficients (R^2) of the regressions were also calculated. The descriptor pool created with CODESSA formed the basis for the selection of ligand-based 2D descriptors (Supporting Information, Table B). The "best multilinear regression (BMLR)" procedure was applied for the development of QSAR models A and B (see Results and Discussion for the naming of QSARs). During the BMLR procedure the pool of descriptors is cleaned from insignificant descriptors ($R^2 < 0.1$) and the descriptors with missing values. In the following steps of BMLR, construction of the best two-parameter regression, the best three-parameter regression, etc. are done based on the statistical significance and noncollinearity criteria ($R^2 < 0.6$) of the descriptors. In BMLR, the descriptor scales are normalized, centered automatically, and the final result is given in natural scales. The final model has the best representation of the property in the given descriptor pool with the given number of parameters. Numerical values of the selected descriptors are tabulated in the Supporting Information, Table C. Having residuals ≥ 2.00 kcal/mol (QSAR B), three (codes 1hhj, om22, and om24) of the 53 systems were outliers and excluded from the final models. Two of them (om24 and 1hhj) were found to be outliers from models in other studies,[30,31] as well. Thus, QSARs with $N = 50$ systems and up to 3 descriptors were developed.

Results and Discussion

Test and Modification of the Scoring Function. For the 50 complexes of the present study the $\Delta G_{\mathrm{b(exp)}}$'s had poor

[23] Park, H.; Lee, S. *J. Am. Chem. Soc.* **2003**, *125*, 16416−16422.

[24] Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219−3228.

[25] Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided. Mol. Des.* **1996**, *10*, 293−304.

[26] Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. *Mol. Simul.* **1993**, *10*, 97−120.

[27] Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864−871.

[28] Frisch, M. J. et al. *GAUSSIAN.98*, revision A.7; Gaussian,Inc.: Pittsburgh, PA, 1998.

[29] (a) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279−287. (b) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual (ver. 2)*; Gainesville, Florida, 1994. (c) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027−1043.

[30] Tounge, B. A.; Reynolds, C. H. *J. Med. Chem.* **2003**, *46*, 2074−2082.

[31] Liu, Z.; Dominy, B. N.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **2004**, *126*, 8515−8528.

Table 1. Correlation of Experimental and Calculated Binding Free Energy Values of the 50 Complexes[a,b]

terms	scoring function (D_1)							
excluded	code	coefficient (α_1)	error of coeff.	*t*-value	R^2	R^2_{cv}	s^2	*F*-value
- - -	ΔG_{AD}	2.5622×10^{-1}	4.8939×10^{-2}	5.2355	0.364	0.323	3.27	27.41
	constant	-5.8868	6.2266×10^{-1}	-9.4542				
T_{HBD}	ΔG_H	2.9877×10^{-1}	4.3668×10^{-2}	6.8419	0.494	0.458	2.60	46.81
	constant	-4.5114	6.7514×10^{-1}	-6.6821				
T_{TOR}	ΔG_T	3.1491×10^{-1}	3.4981×10^{-2}	9.0021	0.628	0.601	1.91	81.04
	constant	-3.1140	6.6747×10^{-1}	-4.6653				
T_{HBD} and T_{TOR}	ΔG_{TH}	3.1686×10^{-1}	2.9505×10^{-2}	10.7392	0.706	0.684	1.51	115.33
	constant	-2.1434	6.4904×10^{-1}	-3.3023				

[a] Linear regressions (eq 2, $n = 1$) were performed using free energies calculated with the (modified) AutoDock SFs as descriptors (D_1) [b] ΔG_{AD} denotes the default AutoDock SF. ΔG_H, ΔG_T, and ΔG_{TH} denote the modified SFs with T_{HBD}, T_{TOR}, and both terms eliminated, respectively. Standard deviations (s^2), squares of the correlation coefficients (R^2), and leave-one-out cross-validated correlation coefficients (R^2_{cv}) of the regressions are tabulated.

correlation with the ΔG_{AD} values calculated with the original SF of eq 1 (squared correlation coefficient, $R^2 = 0.364$; Table 1). However, good correlation ($R^2 = 0.956$) was obtained[12] for the original calibration set of AutoDock 3.0. This apparent contradiction can readily be explained: ΔG_{AD} was originally calibrated on the basis of a diverse set of 30 druglike compounds, and the molecular weight distribution of the 30 ligands of the AutoDock calibration set[12] and that of the 50 ligands in the present study (Figure 1) are significantly different and shifted to larger molecular weights in the latter case. A plausible reason for the low R^2 value for the set of 50 ligands in the present study is the different compound composition from that for the calibration set. Thus, it is reasonable to re-examine the components of the original AutoDock SF using a set of bulky and flexible peptides in order to yield a better fit to the experimental binding free energies for ligands of this problematic type.

In accordance with this finding, eq 1 was inspected to select out terms that depend on the ligand and influence the efficiency of the scoring. One of the two ligand-based terms is T_{HBD}, which represents a penalty, i.e., the loss of free energy due to broken H-bonds between the ligand and water molecules during complex formation with the protein. The exclusion of T_{HBD} alone increases R^2 to 0.494 (ΔG_H). The other simple, ligand-based term in eq 1 is T_{TOR}, which accounts for the change in free energy upon freezing of the torsional degrees of freedom of the ligand. Elimination of this term results in a much better correlation (ΔG_T in Table 1; $R^2 = 0.628$) between the experimental and calculated ΔG_b's in comparison with ΔG_{AD}. Elimination of both terms yields $R^2 = 0.706$ (ΔG_{TH} in Table 1; Figure 2) and an s^2 of 1.51. This model is fairly promising in comparison with other ΔG_b calculators,[32] and therefore, ΔG_{TH} forms a good basis for further, predictive QSARs.

Similarly to the present results, the terms T_{HBD} and T_{TOR} were modified by other authors[33] in order to obtain a good binding free energy model for carbohydrate ligands. In a recent work,[2] the difference between the binding affinities of SGTI (*Schistocerca gregaria* trypsin inhibitor, a 35-amino-acid-long peptide) to two different trypsins was estimated correctly by elimination of these two ligand-based terms. It should be noted that the accumulation of constant penalties from T_{HBD} results in an erroneous positive sum of the free energy of binding for unusually large ligands such as SGTI.

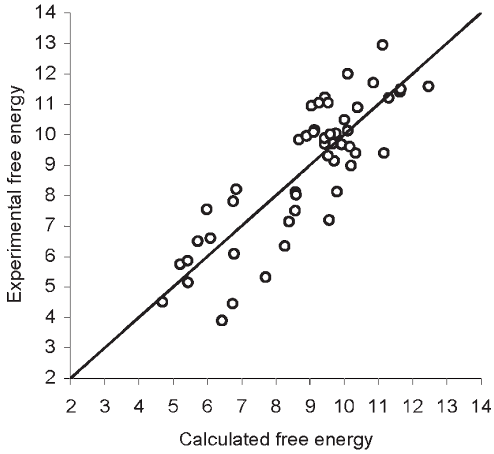


Figure 2. Correlation plot of experimental[14] and calculated binding free energy values (−kcal/mol) of the 50 complexes in the present study. Linear regression (eq 2, $n = 1$) was performed using free energies calculated with the modified SF ΔG_{TH} as a descriptor (D_1).

Development of QSARs Using ΔG_{TH} and Ligand-Based 2D Descriptors. The final correlation ($R^2 = 0.706$) obtained in the previous section is remarkably good showing the usefulness and good predictive power of the remaining bimolecular terms (ΔG_{TH}) having the original AutoDock parameters. Thus, instead of reparametrization of the whole SF, another strategy was followed in the present study. Keeping ΔG_{TH} as a descriptor, which can be reproducibly calculated for any protein−ligand complex structures, new, simple ligand-based descriptors were searched for in order to improve the correlation. Since both T_{HBD} and T_{TOR} can be derived from the 2D molecular graph without inclusion of any 3D information (eq 1),[12,33] the present search for ligand-based descriptors was restricted to 2D ones. A noteworthy advantage of 2D descriptors is that they are easy to calculate and require negligible computational time. Use of the CODESSA descriptor pool complemented with ΔG_{TH} furnishes the QSAR models in Table 2.

The best three-descriptor model (B) in Table 2 includes the bimolecular ΔG_{TH} as a major descriptor and two monomolecular, 2D descriptors, the $RPCG_{EN}$ (relative positive charge based on electronegativity), and the Balaban index (J) (Figure 3).

[32] (a) Böhm, H.-J. *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 309−323. (b) Venkatarangan, P.; Hopfinger, A. J. *J. Med. Chem.* **1999**, *42*, 2169−2179. (c) Marder, M.; Estiú, G.; Blanch, L. B.; Viola, H.; Wasowski, C.; Medina, J. H.; Paladini, A. C. *Bioorg. Med. Chem.* **2001**, *9*, 323−335. (d) Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 11−26. (e) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. *J. Med. Chem.* **2002**, *45*, 2469−2483.

[33] Laederach, A.; Reilly, P. J. *J. Comput. Chem.* **2003**, *24*, 1748−1757.

Table 2. Correlation of Experimental and Calculated Binding Free Energy Values of the 50 Complexes[a,b]

	descriptor (D_i)								
QSAR	*i*	abbreviation	coefficient (α_i)	error of coeff.	*t*-value	R^2	R^2_{cv}	s^2	*F*-value
A	1	ΔG_{TH}	3.1216×10^{-1}	2.4686×10^{-2}	12.6456	0.799	0.774	1.05	93.36
	2	$RPCG_{EN}$	3.2582×10^{1}	6.9963	4.6571				
		constant	−4.1980	6.9930×10^{-1}	−6.0031				
B	1	ΔG_{TH}	2.7077×10^{-1}	2.2926×10^{-2}	11.8105	0.859	0.838	0.76	93.17
	2	$RPCG_{EN}$	5.7129×10^{1}	8.1307	7.0263				
	3	J	-6.2410×10^{-1}	1.4148×10^{-1}	−4.4113				
		constant	−4.6864	6.0281×10^{-1}	−7.7743				

[a] Multilinear regressions (eq 2, $n = 2$ or 3) were performed with ΔG_{TH} and ligand-based 2D descriptors. [b] $RPCG_{EN}$: electronegativity-based relative positive charge (Sanderson's electronegativity scheme). J: Balaban index. For other notes, refer to Table 1.

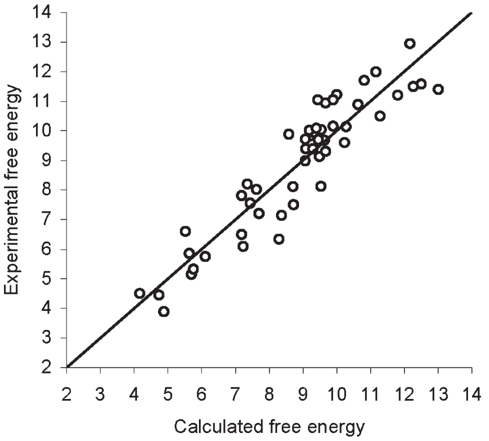


Figure 3. Correlation plot of experimental and calculated binding free energy values (−kcal/mol) of the 50 complexes in the present study in the case of QSAR B. The involvement of $RPCG_{EN}$ and J descriptors significantly improved the correlation as compared with Figure 2.

The $RPCG_{EN}$ values describe the distribution of positive partial charges in a molecule (eq 3):

$$RPCG_{EN} = \frac{\delta_{max}}{\sum_a \delta_a}; \quad a \in \{\delta_a > 0\} \qquad (3)$$

where δ_{max} is the maximum value of the positive partial charges (charge excesses, δ_a) on the atoms (a) of the ligand molecule. In the CODESSA program, the δ values are assigned by a simple method,[34] which uses Sanderson's electronegativities of the atoms. Inspection of the δ_a values in our ligands reveals that most of them are located on H atoms connected to N or O atoms and on the C atoms of the amide bonds.

These H atoms with $\delta > 0$ are the possible H-bonding donor sites on the ligand molecules. Importantly, the regression coefficient of this descriptor is positive (Table 2), which means that it decreases the absolute value of the calculated binding free energy (the $RPCG_{EN}$ values are always positive, eq 3). Similarly, the eliminated T_{HBD} term contributed to ΔG_b with positive penalties, due to vanishing interactions between the ligand and water molecules. The RPCG descriptor was developed and used to account for the effects of polar intermolecular interactions.[35] These results let us conclude that $RPCG_{EN}$ describes (part of) the energy changes due to the altered H-binding system of the ligand during the attachment to a protein. To illustrate the molecular background of the $RPCG_{EN}$ descriptor, the systems 2rkm and 1vwf with ligands having maximum and minimum $RPCG_{EN}$ values (Supporting Information, Table C), respectively, are represented in Figure 4.

It can be seen that the dipeptide ligand (KK) in 2rkm is completely buried inside the protein, while in 1vwf a considerable interaction interface remains between the octapeptide ligand and the surrounding solvent. In 2rkm, the energy contribution of the $RPCG_{EN}$ term to ΔG_b in QSAR B is 6.24, whereas in 1vwf it is only 1.41 kcal/mol. Although the complete burial of a ligand can be considered as an extreme case, the probability of the use of a higher percentage of available H-bonding atoms in the new interactions with the protein is higher for smaller (dipeptide) rather than for larger (octapeptide) ligands. Consequently, the energy penalty corresponding to the loss of ligand-surrounding water interactions should be higher for 2rkm than for 1vwf. The $RPCG_{EN}$ descriptor correctly reflects this observation, as the fewer positively charged H atoms the molecule has, the smaller the denominator and the larger the $RPCG_{EN}$ value, i.e., the penalty (eq 3). If a ligand contains an atom with high δ_{max} (possibly buried into protein), this further increases the penalty. The similar argumentation is also valid for the vanished dipole−dipole interactions between the ligand and the surrounding water, pointing to the generality of $RPCG_{EN}$ descriptor. Besides, $RPCG_{EN}$ contains also indirect information on the size of the molecule via the sum of the partial positive charges (eq 3).

The size of the molecule is directly described by the Balaban index[36] (eq 4) that occurs as the third descriptor in QSAR B:

$$J = \left(\frac{q}{\mu + 1}\right)\sum_{i,j}^{q}(s_i s_j)^{-1/2}; \quad (\mu = q - n + 1) \qquad (4)$$

where q is the number of edges in the molecular graph, n is the number of vertexes in the graph, μ is the cyclometric number, and s_i and s_j are the distance sums obtained by summation of row i and column i or row j and column j, respectively, of the distance matrix between the atoms in the molecule. In J, only the heavy atoms are considered in the molecular graph.

Thus, the J describes not only the size of the molecule but also its internal branching and distances. Interestingly, the number of free torsions (N_{tor}) is a part of the excluded term T_{TOR}, whereas the torsional tree of a ligand is also a type of branching. Considering this and the fact that the change in rotational entropy depends on the moments of inertia, i.e., the

(34) Zefirov, N. S.; Kirpichenok, M. A.; Ismailov, F. F.; Trofimov, M. I. *Dokl. Akad. Nauk.* **1987**, *296*, 883−887.

(35) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323−2329.

(36) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

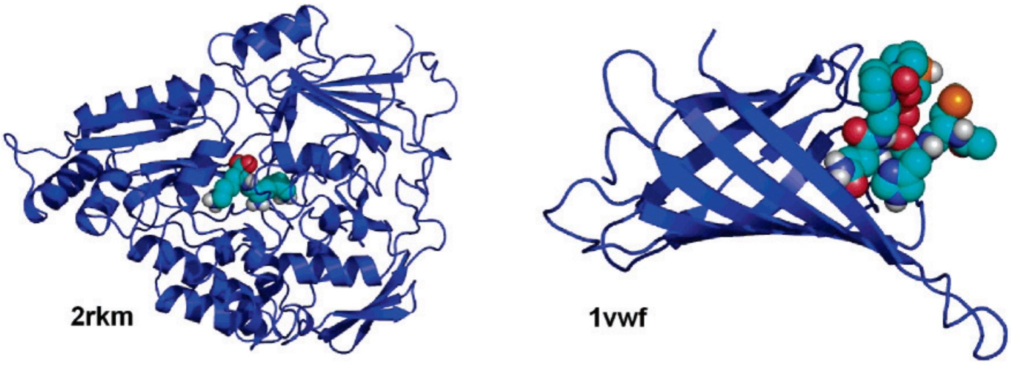


Figure 4. Small dipeptide ligand of the system 2rkm is buried deeply inside the protein, while the octapeptide ligand of 1vwf is sitting on the surface of the protein and its relatively large part can be involved in the ligand−solvent interaction; i.e., a small energy penalty occurs due to deceased H-bonds with the bulk solvent in the case of 1vwf. (Protein molecules and ligands are represented with cartoon and van der Waals surfaces, respectively.)

internal distances of the molecule, J may be descriptive of the change in free energy of binding upon the decrease of rotational and torsional degrees of freedom. In general, the J is based on the molecular structure according to graph theory and the distance matrix and reflects the relative connectivity and effective size of the flexible peptidic molecules. The magnitude of this descriptor increases with (i) an increase in branching and (ii) an increase in the number of atoms in the molecule. However, it would be a much more difficult task to give an analytical explanation for the role of the complex J descriptor than it was for $\mathrm{RPCG_{EN}}$. QSAR B (Table 2) is comparable with other published ΔG_b calculators,[31−33,37] and as it concerns s^2, it is one of the best available calculators for the ΔG_b of large, flexible peptides.

Cross-Validation of the QSARs. The squared correlation coefficients of the leave-one-out cross correlation test (jackknife method) of QSARs are given in Table 2. These coefficients are fairly close to the original R^2's, emphasizing the statistical reliability of the models. The leave-20%-out test provides similarly good R^2 values: 0.780 and 0.848 for QSAR A and B, respectively. As a further test, it can be informative to separate a homogeneous subset of the 50 complexes and use the remaining systems as a training set to check the dependency of the results on this homogeneous part of the data. In our case, there is such a subset of 12 complexes (24%) among the 50, i.e., 12 of the 50 ligands investigated in this study have the same target protein (β-secretase) and are analogous in their structure, and the corresponding experimental inhibition constants used for calculation of the $\Delta G_{b(exp)}$'s were measured in the same laboratory.[14b,d] The results of this test for QSARs A and B are summarized in Table 3. It can be seen that, on the basis of the training set, good correlations are developed for the whole set of 50 points, and therefore, selection of the descriptors for the predictive QSARs is independent of the inclusion of the complexes of the homogeneous subset. Similar R^2 values (0.803 and 0.841 for QSAR-s A and B, respectively) can be calculated if correlating the predicted ΔG_b's of the subset of 12 systems (validation set) with the corresponding $\Delta G_{b(exp)}$'s, using the 38 systems as a training set.

Table 3. Cross-Validation Tests of Descriptor Sets of QSARs A and B Excluding a Homogeneous Subset of 24% of the Data Points[a,b]

	N = 38 (training set)				*N* = 50 (training set + 24% left out)	
QSAR	R^2	R^2_{cv}	s^2	*F*-value	R^2	s^2
A	0.841	0.813	0.85	92.36	0.797	1.10
B	0.893	0.868	0.59	94.26	0.857	0.79

[a] Multilinear regressions were trained for 38 of the 50 systems and tested on all 50 systems. [b] *N* corresponds to the number of systems (data points) used for the correlation. For other notes, refer to Table 1.

Robustness of the Second and Third Descriptors of the Models Obtained. The descriptor J is calculated directly from the molecular graph and is therefore robust, i.e., unambiguously defined by a single chemical formula. The RPCG values are calculated in two steps, as they are derived from the precalculated partial charges (charge excesses, δ, in eq 3) of the atoms of the molecules. It is known that there are several approaches for the assignment of partial charges to the atoms in a molecule. In the case of QSARs A and B, the RPCGs were calculated by using the electronegativity-based charge distribution of the molecules ($\mathrm{RPCG_{EN}}$). However, it may be worthwhile to check whether RPCG remains descriptive on the basis of a different partial charge system. For this reason, QM-based RPCG values ($\mathrm{RPCG_{QM}}$) were calculated and put in the QSARs instead of $\mathrm{RPCG_{EN}}$'s as second descriptors. From among the numerous ways to calculate QM-based partial charges according to different principles (e.g., Mulliken,[38] Hirshfeld[39] charges, etc.), the Breneman and Wiberg approach[40] was selected for the present calculations. This approach reconstitutes the electrostatic potential of a molecule by atomic charges, which is appropriate for this study. It was found that the statistical parameters of the new correlation $\mathrm{A_{QM}}$ ($R^2 = 0.770$; $R^2_{cv} = 0.739$; $s^2 = 1.21$; details of the model are listed in the Supporting Information, Table D) are similar to those of A, with a slight decrease in the R^2 values and that J does not improve the model so effectively in this case ($\mathrm{B_{QM}}$). However, the application of a completely different QM-based partial charge system on the ligand molecules, i.e., a 3D descriptor ($\mathrm{RPCG_{QM}}$) instead of the 2D $\mathrm{RPCG_{EN}}$, does not spoil the descriptive power of RPCG, which

[37] (a) Takamatsu, Y.; Itai, A. *Proteins* **1998**, *33*, 62−73. (b) Huo, S.; Wang, J.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. *J. Med. Chem.* **2002**, *45*, 1412−1419. (c) Vedani, A.; Dobler, M. *J. Med. Chem.* **2002**, *45*, 2139−2149. (d) Ma, X. H.; Wang, C. X.; Li, C. H.; Chen, W. Z. *Protein Eng.* **2002**, *15*, 677−681. (e) Hong, X.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.

[38] Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833−1840, 1841−1846.

[39] Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129−138.

[40] Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361−373.

can therefore be regarded as a robust quantity for the second descriptor in the present QSARs.

Conclusions

The results of the present study indicate that the use of different ΔG_b calculators for ligands of radically different sizes may be considered in future applications and development of docking/scoring methods. A semiempirical SF of a widely used docking method was modified and extended to achieve a precise fit of the structure-based, calculated binding free energy values to the experimental ΔG_b's for bulky, flexible peptidic ligands. The combination of the bimolecular descriptor ΔG_{TH} with additional ligand-based 2D descriptors yielded new, hybrid ΔG_b calculators with good predictive power. The results highlight the possibility of development of such hybrid calculators involving other SF-s in the future. Thorough tests and cross-validations of the QSARs were performed to verify the statistical relevance of the calculators and the descriptors. It was found, that the inclusion of bimolecular terms of the SF is obligatory for a diverse set of protein−ligand systems (ΔG_{TH} is the major descriptor in the QSARs). Both the scoring and the calculation of ligand-based 2D descriptors are rapid processes, even for the large ligands in this study. The precision of their present combination is at least comparable with that of other available calculators of binding thermodynamics. Thus, the proposed strategy is a real alternative for calculation of the binding affinities in the problematic cases of bulky, flexible lead compounds in the early phases of rational drug design. In practice, the docked lead compound−protein complexes can be supplied by AutoDock or other, appropriate automated docking methods and used with the hybrid calculators of the present study to obtain ΔG_b values.

Acknowledgment. This article is dedicated to Docent Dr. István Horváth, Department of Inorganic and Analytical Chemistry, University of Szeged in acknowledgment of his lectures in regression analysis. C.H. is a Békésy Fellow of the Hungarian Ministry of Education. U.M. is grateful to the Estonian Science Foundation for support (Grant No. 5805). This work was supported by OTKA TS 049817 and the Hungarian National Bureau of Research and Development.

Supporting Information Available: Details of the protein−ligand complexes (Table A). The Codessa descriptor pool used for selection of the appropriate 2D descriptors (Table B). Numerical values of the descriptors (Table C). The correlation of experimental and calculated binding free energy values of the 50 complexes. Multilinear regressions (eq 2, $n = 2$ or 3) were performed, using ΔG_{TH}, $RPCG_{QM}$, and ligand-based 2D descriptor J (Table D). Complete ref 28. This material is available free of charge via the Internet at http://pubs.acs.org.

JA055804Z

hetenyi.csaba_83_23

D22

DOI: 10.1002/cmdc.200600251

Structure-Based Calculation of Binding Affinities of α_{2A}-Adrenoceptor Agonists

Balázs Balogh,[a] Csaba Hetényi,*[a, b] Miklós György Keserű,[c] and Péter Mátyus*[a]

Dedicated to Professor E. Sylvester Vizi on the occasion of his 70th birthday.

Adrenergic receptors of the α_2 type (α_2-adrenoceptors) belong to the family of seven transmembrane-spanning G-protein-linked receptors.[1–9] α_2-Adrenoceptors can be grouped into three highly homologous subtypes (α_{2A}, α_{2B}, and α_{2C}) and, because of the difference in pharmacology,[10] a fourth subtype (α_{2D}) can be formally distinguished, though this is rather a species orthologue.

In general, the α_2-adrenoceptors are responsible for the presynaptic feedback of the release of adrenaline and noradrenaline, their physiological agonists. Although numerous findings are available on the receptor subtypes from experiments with knockout mice[11] and these results are of some relevance for human pharmacology, the similar patterns of expression of adrenergic receptors in human and mouse tissues do not guarantee similar functions. Thus, the individual roles of the three α_2-adrenoceptor subtypes in humans have not been completely elucidated. However, the results of the reported studies do indicate (see Supporting Information) that the α_2-adrenoceptor subtypes are involved in various important physiological processes, and further investigations of the differences in their molecular pharmacology are therefore essential.

The identification of subtype-specific functions from pharmacological experiments is currently not possible because of the lack of subtype-specific ligands[3,6–8] and the cross-reactivity with imidazoline receptors.[7] The development of subtype-selective agonists would be useful as it would facilitate further examinations of the molecular pharmacology of the α_2-adrenoceptors. The rational, structure-based design of such agonists requires a precise knowledge of the molecular structure of the binding site. Unfortunately, because of the difficulties inherent in crystallization, atomic-resolution structures of the α_2-adrenoceptors are not available in the Protein Databank.

In the present study, an atomic-resolution model of the α_{2A}-adrenoceptor was constructed through use of its amino acid sequence and the crystallographic bovine rhodopsin structure as a template. Similar homology models were earlier constructed by other researchers[12] and successfully used to provide qualitative explanations. The α_{2A}-adrenoceptor model in the present study is based on a crystallographic template structure with a resolution of 2.2 Å[13] appropriate for quantitative investigations (for details, refer to the Computational Methods below).

In possession of the atomic resolution target structure (α_{2A}-adrenoceptor), 15 known agonist ligands were automatically docked to the presumed binding region of the receptor (Figure 1 a). Inspection of the results revealed that the docked ligand conformations are in physical contact with the key residues D3.32(113), S5.42(200), and S5.46(204), previously identified by site-directed mutagenesis studies.[14–16] As an example, the positively charged amino group of noradrenaline (Figure 1 b) or of methylnoradrenaline forms a salt bridge with the negative side-chain carboxylate of D3.32(113). Similar results involving an interaction between the ionic groups were earlier obtained for noradrenaline.[13] For some other ligands (for example, clonidine, Figure 1 c), interactions can be observed with E4.39(189) instead of D3.32(113) . Additionally, the binding pocket is formed by hydrophobic amino acids such as V5.39-(197), F5.47(205), W6.48(258), F6.49(259), F6.52(262), and the key serine residues.

The qualitative agreement with the site-directed mutagenesis data indicates the usefulness of the homology model and the docking procedure applied. However, a correct (quantitative) estimation of the binding free energy (ΔG_b) is the real challenge in molecular design. Once a ΔG_b calculator has been developed, the screening-out of potent (tight binding) agonists from the candidate compounds becomes possible. To meet this expectation, quantitative structure–activity relationships (QSARs) were developed by using the docked structures and experimental ΔG_b values of the agonists.

As a first attempt, simple linear regression (LR) was performed, involving the modified scoring function values (ΔG_T) of AutoDock 3.0 program package, which includes the intermolecular (enthalpic) terms and a solvation penalty. These values were calculated for the docked agonist–protein complex structures. A detailed discussion on the calculation of ΔG_T is to be found in Ref. [17]. An excellent correlation was obtained for nine ligands not containing chlorine atoms [Eq. (1), Figure 2].

$$\Delta G_b = \underset{\pm 0.1301}{\overset{t=12.3237}{1.6037}}\,\Delta G_T + \underset{\pm 1.0909}{\overset{t=3.9601}{4.3201}}$$

$$\left(r^2 = 0.96; r^2_{cv} = 0.93; F = 151.87; s^2 = 0.09; N = 9\right) \quad (1)$$

An inspection of the t-values indicates that both ΔG_T and the intercept are necessary parameters of the regression equation. The mean square errors of the regression coefficients, the F value, the standard deviation (s^2), the square of the correlation coefficient (r^2), and the leave-one-out cross-validated r^2 (r^2_{cv}) of the regressions reflect the statistical significance of the LR.

[a] B. Balogh, Dr. C. Hetényi, Prof. P. Mátyus
Department of Organic Chemistry, Semmelweis University, Hőgyes E. u. 7.
1095 Budapest, Hungary, and Szentágothai Knowledge Center, Molnár u. 19.
1056 Budapest (Hungary)
Fax: (+36) 1-217-0851
E-mail: peter.matyus@szerves.sote.hu

[b] Dr. C. Hetényi
Institute of Chemical Physics, Tartu University, Tartu (Estonia)
E-mail: csabahete@yahoo.com

[c] Dr. M. G. Keserű
Richter Gedeon Nyrt. Gyömrői út 19-21. 1103 Budapest (Hungary)

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

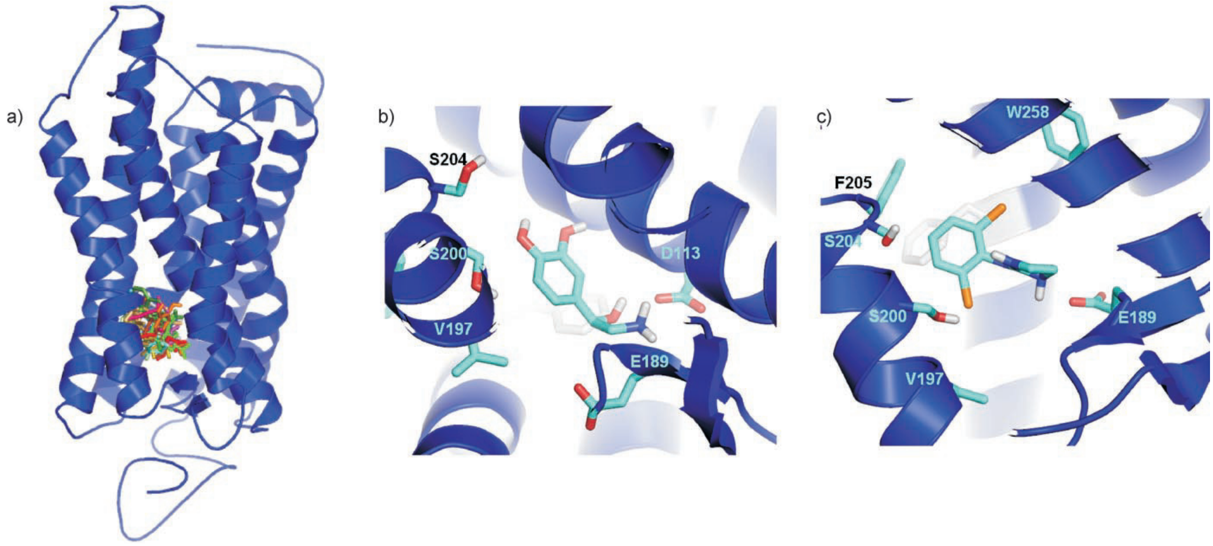


Figure 1. a) Structure of the homology modeled and energy minimized α_{2A}-adrenoceptor. Docked conformations of all 15 ligands are located in the same central binding cavity. b) Noradrenaline binding to the active site of the α_{2A}-adrenoceptor. Key residues of the site are denoted by sticks. A salt bridge is formed between the oppositely charged side-chain of D113 and the amino group of noradrenaline. c) Clonidine binding to the active site of the α_{2A}-adrenoceptor. Experimentally detected key residues of the site are denoted by sticks. A hydrophobic binding pocket is formed by W and F residues.

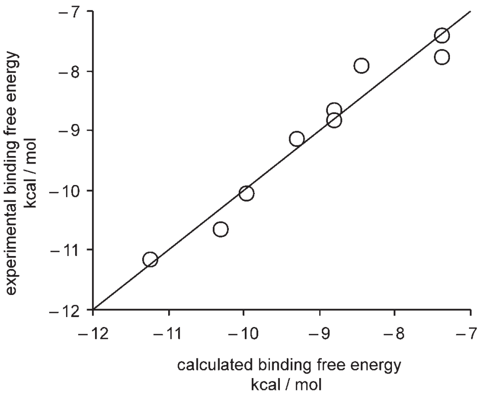


Figure 2. Correlation between the experimental and calculated binding free energies of nine agonists. Small residuals were obtained for nonchlorinated compounds with the use of only one descriptor: the modified AutoDock free energy function, ΔG_T [Eq. (1)].

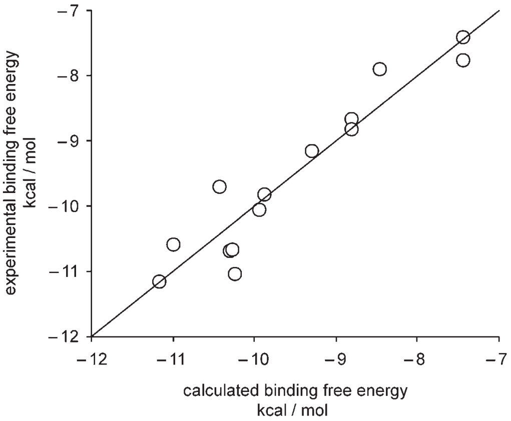


Figure 3. Correlation between the experimental and calculated binding free energies of all 14 agonists. Besides ΔG_T, involvement of a second descriptor resulted in a fair correlation for the chlorinated compounds too [Eq. (2)].

Five chlorinated agonists do not satisfy Equation (1). The common feature of these five molecules is that all of them have a 2,6-dichloro substituted phenyl (2,6-DCP) ring. Thus, it is plausible to involve a binary descriptor of existence (*E*) of the 2,6-DCP ring in the regression, which accounts for the presence or absence of this moiety, that is, *E*=1 (or 0) if there is (or is not) a 2,6-DCP ring in the ligand. Inclusion of this descriptor yields a three-parameter LR [Eq. (2), Figure 3]:

$$\Delta G_b = \underset{\pm 0.1875}{\overset{t=8.2915}{1.5543}}\,\Delta G_T \underset{\pm 0.2524}{\overset{t=-7.3038}{-1.8434}}\,E + \underset{\pm 1.5718}{\overset{t=2.4860}{3.9075}}$$

$$\left(r^2 = 0.90; r^2_{cv} = 0.84; F = 48.62; s^2 = 0.19; N = 14\right) \quad (2)$$

Similarly as for Equation (1), this multiple LR is statistically relevant and only one (dexmedetomidine) of the 15 agonists was an outlier with a residual >1.5 kcal mol^{-1}, and had to be omitted from the final LR. ΔG_T includes mostly intermolecular (enthalpic) contributions to ΔG_b[17] and the constants 4.3201 and 3.9075 kcal mol^{-1} in Equations (1) and (2), respectively, sufficiently represent the entropic loss due to freezing of translational, rotational, and torsional degrees of freedom in the nine ligands. However, descriptor E in Equation (2) requires further discussion. Notably, the sign of the coefficient of E is negative. This means that the presence of a 2,6-DCP ring is favorable for binding, indicating two possibilities. 1) The substituent chlorine atoms are involved in interactions with the protein which are not correctly represented by ΔG_T. Comparison of the atomic contributions of the chlorine to the electrostatic and van der Waals terms of ΔG_T with those of other ligand atoms with a similar character (for example, oxygen) allows the conclusion that the enthalpic contributions are not underrepresented for the chlorine atoms. 2) The presence of the 2,6-DCP ring alters the entropy of binding. Conformational energy diagrams for the phenyl rotation (Supporting Information) show that the energy gap between the stable and the high energy conformation is twice as high for the 2,6-DCP ring as it is for the simple phenyl ring. Besides this intramolecular interaction effect, the heavy chlorine atom may alter the corresponding rotational frequency too. Certainly, movements of the phenyl rotor are restricted following chlorine substitution, and the entropic loss of this freezing rotor is therefore smaller. This decrease in the entropic loss may be a realistic explanation of the negative sign of E in Equation (2). Involvement of other 2,6-DCP ring-containing (at any event not a *para*-chlorophenyl-containing) ligand–protein complexes and quantum chemical calculations would be necessary for a detailed elucidation, but that is beyond the scope of the present study. The structures and conformational degrees of freedom of the ligand molecules within the two, that is, chlorinated and nonchlorinated subsets are similar. Thus, our results agree with the rational assumption that the binding entropy is approximately the same for the ligands within the two subsets.

The experimental ΔG_b values of these 15 agonists were converted from the pK_i (logarithm of inhibition constant) values obtained from radioligand assays. For nine of the 15 compounds, the pK_i values were determined with two different radioligands [^{3}H] MK-912 and [^{3}H] RX821002. A LR using the corresponding two vectors of the experimental ΔG_b data yields valuable information on the interchangeability and reproducibility of the available experimental data. Although the two vectors are correlated ($r^2 = 0.76$) with each other, the statistical parameters (see the Supporting Information for details) of this correlation are not as fascinating as might be hoped. Thus, in the present study, it was a good choice to use experimental data obtained with only one radioligand ([^{3}H] MK-912) for QSAR building.

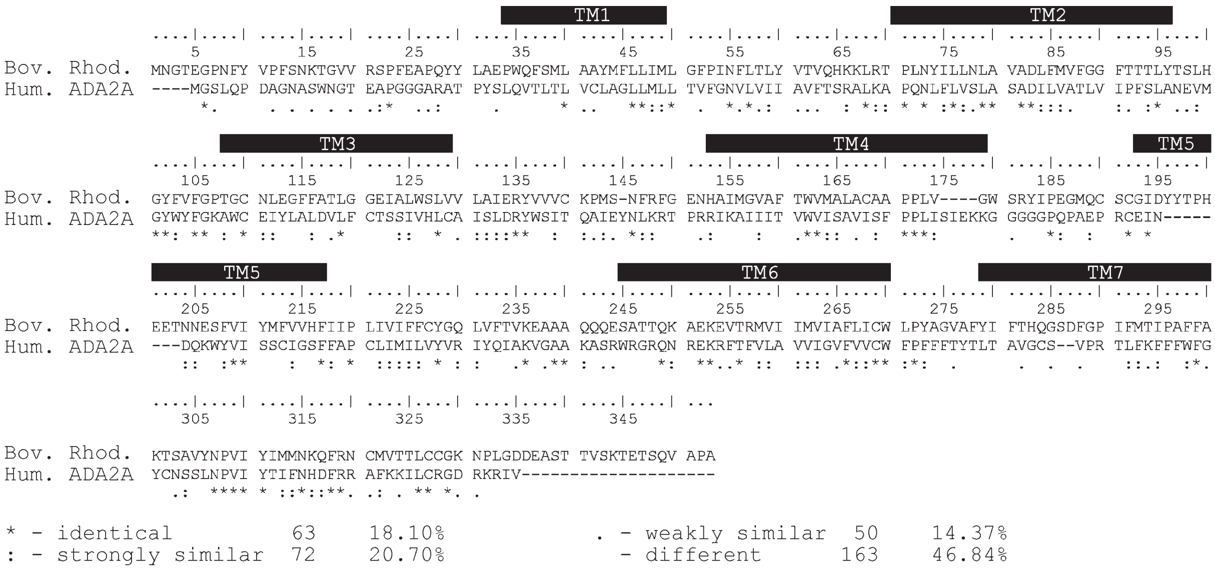
In conclusion, 15 agonists with various structures were docked to an atomic resolution homology model of the human α_{2A}-adrenoceptor. The docked conformations of the compounds are in contact with previously reported key binding site residues emphasizing the good quality of the homology model. QSARs of binding affinity were developed involving structure-based bimolecular terms of the AutoDock scoring function, a simple, ligand-based binary descriptor, and a set of the corresponding experimental ΔG_b values. A good correlation was achieved between the experimental and calculated ΔG_b values. The statistical parameters of the LRs are somewhat better than those of the reproducibility of the experimental data. To the best of our knowledge, this study represents the first verified calculations of binding affinities of agonists to the α_{2A}-adrenoceptor. Thus, our results indicate the direction of precise engineering of agonists, either for the elucidation of open questions of subtype selectivity (see introductory sections) or for the design of drug candidates in α_{2A}-adrenoceptor-related diseases and therapeutic issues such as hypertension,[18–19] glaucoma,[20] acute migraine,[21] analgesia, anesthesia, sedation,[22–24] drug and alcohol withdrawal,[25–27] gastroprotective effects,[28–30] and Parkinson's disease.[31–32]

Computational Methods

Homology modeling and refinement. The amino acid sequences of both bovine rhodopsin (template protein) and the human α_{2A}-adrenoceptor were obtained from the online protein database.[33] Both sequences were loaded in Bioedit 7.0.5.2[34] and were aligned (Figure 4) with ClustalW 1.4. Manual correction of the alignment was performed if needed (see Supporting Information). A bovine rhodopsin coordinate file (Protein Databank code: 1U19) of 2.20 Å[13] was selected as the structural template. Modeller 8v1[35–37] was used for model building. Inputs of Modeller were the protein coordinates of 1U19, the amino acid sequences, and a file containing the options of the calculations. One hundred α_{2A}-adrenoceptor homology models were created and the model with the lowest modeller objective function value was selected. The quality of the model was checked with the web version of the program ProCheck1.5[38–39] (see Supporting Information). The α_{2A}-adrenoceptor homology model with a blind docked[40] noradrenaline ligand conformation sitting at the binding region was refined by GROMACS[41] molecular mechanics minimization, as described previously.[17]

Docking and scoring. The **s**tructures of the 15 agonist molecules (noradrenaline, α-methyl-noradrenaline, B-HT 920, brimonidine, clonidine, dexmedetomidine, guanabenzamidine, guanfacine, levlofexidine, oxymetazoline, *p*-aminoclonidine, rilmenidine, st91, xylometazoline, and a54741) were built, optimized, and supplied with Gasteiger charges by using the SYBYL program package and force field.[42–44] The docking box with $22.5 \times 22.5 \times 22.5$ Å^3 volume was centered at the binding region known from site-directed mutagenesis studies.[6] All docking calculations were performed as in Ref. [45], using the AutoDock 3.0 program package.[46] Ligand molecules with different proton locations were investigated in cases where protonation was not trivial. Protonated forms (and the corresponding ΔG_T) resulting in the smallest residuals were selected for QSAR. Detailed results of docking are tabulated in the Supporting Information. Molecular graphics was prepared with PyMol.[47]

Linear regressions. QSARs were developed with the CODESSA program package[48–50] and its two-dimensional descriptor pool. The binary descriptor E was constructed manually, and included in the pool and selected automatically by the improve correlation module of CODESSA. Experimental ΔG_b values used in the LRs (Supporting Information) were converted from pK_i values ($T = 298$ K) of previous studies of radioligand assays.[51–53] Only pK_i values obtained with radioligand [^{3}H] MK-912 were used to construct Equations (1) and (2). When more than one experimental pK_i value was available, the larger one was selected for correlation.



```
                                                         TM1                                           TM2
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                5          15         25         35         45         55         65         75         85         95
Bov. Rhod.  MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
Hum. ADA2A  ----MGSLQP DAGNASWNGT EAPGGGARAT PYSLQVTLTL VCLAGLLMLL TVFGNVLVII AVFTSRALKA PQNLFLVSLA SADILVATLV IPFSLANEVM
                 *.     . . . . . .:*  ..:     .      * ..   **::*  .  *.*.:  ...  : *:: * * :*:.**  **:::.    :. :* ..:

                               TM3                                                   TM4                               TM5
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               105        115        125        135        145        155        165        175        185        195
Bov. Rhod.  GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMS-NFRFG ENHAIMGVAF TWVMALACAA PPLV----GW SRYIPEGMQC SCGIDYYTPH
Hum. ADA2A  GYWYFGKAWC EIYLALDVLF CTSSIVHLCA ISLDRYWSIT QAIEYNLKRT PRRIKAIIIT VWVISAVISF PPLISIEKKG GGGGPQPAEP RCEIN-----
            **: ** : * ::   : .*     :: * . ::::** :  :.:. *::    .:    :   .**:: . :  ***:       .   *:  :   * *

                     TM5                                                   TM6                             TM7
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               205        215        225        235        245        255        265        275        285        295
Bov. Rhod.  EETNNESFVI YMFVVHFIIP LIVIFFCYGQ LVFTVKEAAA QQQESATTQK AEKEVTRMVI IMVIAFLICW LPYAGVAFYI FTHQGSDFGP IFMTIPAFFA
Hum. ADA2A  ---DQKWYVI SSCIGSFFAP CLIMILVYVR IYQIAKVGAA KASRWRGRQN REKRFTFVLA VVIGVFVVCW FPFFFTYTLT AVGCS--VPR TLFKFFFWFG
               ::  :**    :  *: *  ::::: * : :   .* .** : ..    *:  **..* ::  ::: .*::** :*:  .      .  .  .    ::.:  :*.

            ....|....| ....|....| ....|....| ....|....| ....|....| ...
               305        315        325        335        345

Bov. Rhod.  KTSAVYNPVI YIMMNKQFRN CMVTTLCCGK NPLGDDEAST TVSKTETSQV APA
Hum. ADA2A  YCNSSLNPVI YTIFNHDFRR AFKKILCRGD RKRIV--------------------
              .:  **** * ::*::**. .: . ** *. .

* - identical         63    18.10%            . - weakly similar  50    14.37%
: - strongly similar  72    20.70%              - different      163    46.84%
```

Figure 4. The sequence alignment of bovine rhodopsine (template) and human α_{2A} adrenoceptor (target) proteins. There is approximatately 40 % similarity (identical + strongly similar) between the sequences of the two proteins. Transmembrane regions are marked with TM.

Acknowledgements

We thank the National Office for Research and Technology (NKTH), Hungary for support of this study. C.H. is greatful to Foundation Innove (www.innove.ee) project No. 1.0101-0310 for financial support, and the referee for useful suggestions on the manuscript.

Keywords: alpha$_{2a}$-adrenerg receptor · docking · free energy · homology modeling · QSAR

[1] E. Barclay, M. O'Reilly, G. Milligen, *Biochem. J.* **2005**, *385*, 197 – 206.
[2] C. S. Pao, J. L. Benovic, *J. Biol. Chem.* **2005**, *280*, 11052 – 11058.
[3] M. Philipp, L. Hein, *Pharmacol. Ther.* **2004**, *101*, 65 – 74.
[4] J. P. Vilardaga, M. Bünemann, C. Krasel, M. Castro, M. J. Lohse, *Nat. Biotechnol.* **2003**, *21*, 807 – 812.
[5] M. Scheinin, J. Sallinen, A. Haapalinna, *Life Sci.* **2001**, *68*, 2277 – 2285.
[6] L. Hein, J. D. Altman, B. K. Kobilka, *Nature* **1999**, *402*, 181 – 184.
[7] L. B. MacMillan, L. Hein, M. S. Smith, M. T. Piasick, L. E. Limbrid, *Science* **1996**, *273*, 801 – 803.
[8] R. E. Link, K. Desai, L. Hein, M. E. Stevens, A. Chruscinski, D. Bernstein, G. S. Barsh, B. K. Kobilka, *Science* **1996**, *273*, 803 – 805.
[9] B. K. Kobilka, H. Matsui, T. S. Kobilka, T. L. Yang-Feng, U. Francke, M. G. Caron, R. J. Lefkowitz, J. W. Reagen, *Science* **1987**, *238*, 650 – 656.
[10] G. J. Molderings, K. Schmidt, H. Boenisch, M. Goethert, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **1996**, *353*, 245 – 249.
[11] M. Philipp, L. Hein, *Pharmacol. Ther.* **2004**, *101*, 65 – 74.
[12] H. Xhaard, T. Nyrönen, R. Ville-Veikko, J. O. Ruuskanen, J. Laurila, T. Salaminen, M. Scheinin, M. S. Johnson, *J. Struct. Biol.* **2005**, *150*, 126 – 143.
[13] T. Okada, M. Sugihara, A. N. Bondar, M. Elstner, P. Entel, V. Buss, *J. Mol. Biol.* **2004**, *342*, 571 – 583.
[14] P. J. Pauwels, F. C. Colpaert, *Br. J. Pharmacol.* **2000**, *130*, 1505 – 1512.
[15] J. E. Rudling, K. Kennedy, P. D. Evans, *Br. J. Pharmacol.* **1999**, *127*, 877 – 886.
[16] M. G. Eason, S. B. Liggett, *J. Biol. Chem.* **1995**, *270*, 24 753 – 24 760.
[17] C. Hetényi, G. Paragi, U. Maran, Z. Timár, M. Karelson, B. Penke, *J. Am. Chem. Soc.* **2006**, *128*, 1233 – 1239.
[18] C. Fenton, G. M. Keating, K. Lyseng-Williamson, *Drugs* **2006**, *66*, 477 – 496.
[19] D. W. Blake, J. Ludbrook, A. F. Van Leeuwen, *Clin. Exp. Pharmacol. Physiol.* **2000**, *27*, 801 – 809.
[20] J. Savolainen, J. Rautio, R. Razetti, T. Jaevinen, *J. Pharm. Pharmacol.* **2003**, *55*, 789 – 794.
[21] K. Kapoor, E. W. Willems, A. MaassenVanDenBrink, J. P. C. Hiligers, A. A. Cordi, C. Vayssettes-Courchay, *Cephalalgia* **2004**, *24*, 425 – 438.
[22] A. Romero-Sandoval, J. C. Eisenach, *Anesthesiology* **2006**, *104*, 351 – 355.
[23] S. M. Tham, J. A. Angus, E. M. Tudor, C. E. Wright, *Br. J. Pharmacol.* **2005**, *144*, 875 – 884.
[24] J. Schweimer, M. Fendt, H. U. Schnitzler, *Eur. J. Pharmacol.* **2005**, *507*, 117 – 124.
[25] A. D. Lê, S. Harding, W. Juzytsch, D. Funk, Y. Shaham, *Psychopharmacology* **2005**, *179*, 366 – 373.
[26] F. Georges, S. Caillé, C. Vouillac, C. Le Moine, L. Stinus, *Eur. J. Neurosci.* **2005**, *22*, 1812 – 1816.
[27] G. Francois, A. J. Gary, *Neuropsychopharmacology* **2003**, *28*, 1140 – 1149.
[28] K. Fülöp, Z. Zándori, A. Z. Rónai, K. Gyires, *Eur. J. Pharmacol.* **2005**, *528*, 150 – 157.
[29] N. K. Jain, S. K. Kulkarni, A. Singh, *Life Sci.* **2002**, *70*, 2857 – 2869.
[30] K. Müllner, A. Z. Rónai, K. Fülöp, S. Fürst, K. Gyires, *Eur. J. Pharmacol.* **2002**, *435*, 225 – 229.
[31] A. Haapalinna, T. Leino, E. Heinonen, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2003**, *368*, 342 – 351.
[32] T. Archer, A. Frederiksson, *J. Neural Transm.* **2003**, *110*, 183 – 200.
[33] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek, *Nucleic Acids Res.* **2006**, *34*, D187 – D191.
[34] T. A. Hall, *Nucleic Acids Symp. Ser.* **1999**, 41, 95 – 98.
[35] M. A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali, *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291 – 325.
[36] A. Fiser, R. K. Do, A. Sali, *Protein Sci.* **2000**, *9*, 1753 – 1773.
[37] A. Sali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779 – 815.
[38] R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, *J. Appl. Crystallogr.* **1993**, *26*, 283 – 291.
[39] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, J. M. Thornton, *Proteins* **1992**, *12*, 345 – 364.
[40] C. Hetényi, D. van der Spoel, *FEBS Lett.* **2006**, *580*, 1447 – 1450.

[41] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. Berendsen, *J. Comput. Chem.* **2005**, *26*, 1701–1718.
[42] M. Clark, R. D. Cramer, N. Van Opdenbosch, *J. Comput. Chem.* **1989**, *10*, 982–1012.
[43] J. G. Vinter, A. Davis, M. R. Saunders, *J. Comput.-Aided Mol. Des.* **1987**, *1*, 31–51.
[44] I. Motoc, R. A. Dammkoehler, D. Mayer, I. Labanowski, *Quant. Struct.-Act. Relatsh.* **1986**, *5*, 99–105.
[45] C. Hetényi, U. Maran, M. Karelson, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1576–1583.
[46] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, *J. Comput. Chem.* **1998**, *19*, 1639–1662.
[47] W. L. DeLano, *PyMol Molecular Graphics System*, DeLano Scientific, San Carlos CA, USA, **2002**.
[48] A. R. Katritzky, V. S. Lobanov, M. Karelson, *Chem. Soc. Rev.* **1995**, *24*, 279–287.
[49] A. R. Katritzky, V. S. Lobanov, M. Karelson, *CODESSA: Reference Manual (ver. 2)*, Gainesville, Florida, **1994**.
[50] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Soc. Rev.* **1996**, *25*, 1027–1043.
[51] V. Audinot, N. Fabry, J. P. Nicolas, P. Beauverger, A. Newman-Tancredi, M. J. Millan, A. Try, F. Bornancin, E. Canet, J. A. Boutin, *Cell. Signalling* **2002**, *14*, 829–837.
[52] S. J. MacLennan, L. A. Loung, J. R. Jasper, Z. P. To, R. M. Eglen, *Br. J. Pharmacol.* **1997**, *121*, 1721–1729.
[53] J. R. Jasper, J. D. Lesnick, L. K. Chang, S. S. Yamanishi, T. K. Chang, S. A. O. Hsu, D. A. Daunt, D. W. Bonhaus, R. M. Eglen, *Biochem. Pharmacol.* **1998**, *55*, 1035–1043.

Received: October 25, 2006
Revised: February 5, 2007
Published online on April 16, 2007

D23

A Comprehensive Docking Study on the Selectivity of Binding of Aromatic Compounds to Proteins

Csaba Hetényi,*[,†,‡] Uko Maran,[†] and Mati Karelson[†]

Department of Chemistry, Tartu University, 2 Jakobi Street, 51014 Tartu, Estonia, and Department of Medical Chemistry, University of Szeged, Dóm tér 8, 6720 Szeged, Hungary

Received March 19, 2003

Generally, computer-aided drug design is focused on screening of ligand molecules for a single protein target. The screening of several proteins for a ligand is a relatively new application of molecular docking. In the present study, complexes from the Brookhaven Protein Databank were used to investigate a docking approach of protein screening. Automated molecular docking calculations were applied to reproduce 44 protein−aromatic ligand complexes (31 different proteins and 39 different ligand molecules) of the databank. All ligands were docked to all different protein targets in altogether 12 090 docking runs. Based on the results of the extensive docking simulations, two relative measures, the molecular interaction fingerprint (MIF) and the molecular affinity fingerprint (MAF), were introduced to describe the selectivity of aromatic ligands to different proteins. MIF and MAF patterns are in agreement with fragment and similarity considerations. Limitations and future extension of our approach are discussed.

INTRODUCTION

X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) measurements are undoubtedly the most reliable sources of high-resolution structures of protein−ligand complexes. Despite the increase of companies and research laboratories that carry out experimental elucidation of structures of macromolecular complexes of biological interest (the number of determined protein molecules is rapidly increasing[1]), in silico drug screening and design techniques remain indispensable tools of in vitro or in vivo high throughput screening (HTS) and design methods. Using experimentally determined ligand−protein complexes as references, computational docking is one of the most important in silico HTS methods[2] and can easily be combined with combinatorial chemistry.[3] Moreover, even if knowledge of the binding site of the ligand molecule is missing, recently introduced computational approaches can be applied for scanning entire macromolecular targets with small[4] or larger, flexible ligands[5−7] (*blind docking*). Further information on molecular docking and its applications can be found in refs 8−10.

The most important components of each docking algorithm are the search method and the scoring function. Scoring functions provide fast binding energy calculation during the minimum search of docking simulations. Scoring formulas can be applied independently to estimate binding free energies, if there is not appropriate computational capacity for free energy calculation from e.g. MD trajectories[11] or there are too many systems to be calculated. In the present study, the scoring of AutoDock 3.0[12] was applied. This scoring function is based on the Lennard-Jones and screened Coulombic terms of the 2.4 version. The original terms were scaled, and additional solvation and torsional considerations were introduced to obtain a better fit to binding free energy values. However, it should be remarked, that the absolute values of the estimated free energies are obviously not error-free, e.g. because the experimental binding constants involve some uncertainty and the approximations of solvation effects, etc. has limited power, as well. Thus, in the present study, the use of relative (subtracted) binding free energy values was preferred.

Rapid calculation of correct binding geometries and estimation of the conjugated binding free energies are essential not only solely in “traditional” HTS applications, i.e., if thousands of drug candidates are scanned for the same protein target, but also when several proteins or (more correctly) the different binding pockets are screened for the same ligand molecule.

The latter application of docking was used for prediction of drug side effect and toxicity by Chen et al.[13] In their study, docking simulations of drug molecules were used to select the proteins, which might play important role in the biochemical pathways of side effects. A large set of protein binding pockets was used as a basis of the selection. A successful application of AutoDock was reported[14] for virtual protein screening and drug side effect prediction, as well. However, screening of several proteins (instead of ligands) is a relatively new direction in docking studies. Further applications are required to test the approach and find the solution for the problems outlined in the aforementioned papers.

In the present work, a series of protein−aromatic ligand complexes was selected to investigate the efficiency of docking and scoring for selection of appropriate protein(s) for aromatic ligands. The reliability of the scoring (free energy) function and search method of AutoDock 3.0 was verified by structural match of the lowest energy conformers

* Corresponding author phone: +372-7-375254; fax: +372-7-375264; e-mail: csabahete@yahoo.com. Corresponding author address: Department of Chemistry, Tartu University, 2 Jakobi Street, 51014 Tartu, Estonia.

[†] Tartu University.

[‡] University of Szeged.

10.1021/ci034052u CCC: $25.00 © 2003 American Chemical Society
Published on Web 07/11/2003

of the docking experiments (jobs) to native crystallographic structures of ligand molecules of all complexes.

Most of the ligand molecules are structurally similar benzene derivatives (some naphthalene and indole compounds were also involved in the study) and, therefore, fragment considerations and structural similarities allow qualitative verification of the results. Molecular interaction and molecular affinity fingerprints (MIF and MAF, respectively) were introduced to get a comprehensive picture on the selectivity of binding of the aromatic molecules on proteins.

METHODS

Docking. Forty-four complexes of 39 different, small, and middle-size aromatic ligand molecules and 31 different protein molecules (Table 1) were selected from the Brookhaven Protein Databank[15] (PDB), partially with the aid of PDBsum[16] server. All selected complexes are free of close contacts, and the ligand is not buried inside the protein by the side-chains, i.e., only systems with no or moderate induced fits were considered. The latter criterion is important, as the applied docking method handles only rigid target protein molecules and therefore no induced effects can be modeled. Moreover, systems with binding sites containing several water molecules around the ligand were omitted, as those molecules can be involved in specific binding of some ligands.[5,17,18] Thus, if crystallographic water molecules were constrained at the site and involved in docking calculations, the binding site of a certain protein would be appropriate only for docking of its original ligand molecule or very similar molecules.

Ligand and water molecules and all ions were removed from the original PDB file. If the PDB file contained several identical chains, the one bound to the ligand molecule of the lowest B-factors was selected. Essential hydrogen atoms, charges of the Kollman united atom type, and solvation parameters were added to the residues of the protein chain. Generally, the native ligand molecules were used, as their bond lengths and angles were found adequate for this purpose. In the case of systems **4d**, **7a**, **15**, and **32**, the ligand molecules were model-built in MOLDEN[19] and optimized with the aid of TINKER[20] using a modified MM3 force field. Babel[21] and VEGA[22] programs were used for file manipulations. Ligand molecules were equipped with all hydrogen atoms and Gasteiger−Marsili charges.[23] Autotors (an AutoDock[12] tool) was applied for creation of united atom representation and definition of the torsions of the ligands. A uniform procedure was applied for all the (31 × 39 =) 1209 docking jobs. Affinity (grid) maps of 60 grid points in each Cartesian directions and 0.375 Å spacing were generated with the aid of Autogrid.[12] Maps were centered on the original ligand molecules. AutoDock parameter set and distance dependent dielectric function were used in the calculation of the van der Waals and the electrostatic terms, respectively. Docking simulations were performed using the Lamarckian genetic algorithm and the Solis & Wets local search method of Autodock. All torsion angles of the ligand molecules (except of amide bonds and some conjugated or rigid bonds) were released during docking. The numbers of released torsions (RT) of the ligands are listed in Table 1. Initial position, orientation, and torsions of the ligands were set randomly. During the search, translational step of 0.2 Å, quaternion, and torsion steps of 5° were applied. A population of 50 members and a maximum number of 2.5 million energy evaluations were used. Ten docking runs were performed for each job.

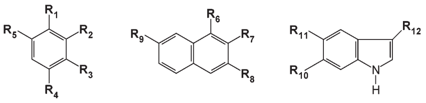
Evaluation of the Results. A C program was used for the evaluation and RMSD (Root Mean Square Deviation) calculation of all data. Binding free energies of each job were collected and the minima were selected. RMSD was calculated for the resulted 10 structures using ligand structures of minimum energies (in Table 2: crystallographic structures) at each job as references. A 2.5 Å tolerance was used to form clusters of the closest structures. The atoms of groups of C_{2v} and C_{3v} symmetries were considered identical during RMSD calculations. Average energies of the clusters of each job were calculated and collected in the $\mathbf{E_1}$ (39 × 31) data matrix. Standard deviations and number of cluster members were calculated and collected too. An additional $\mathbf{E_2}$ matrix was produced using a distance criterion, as follows. The distances between the centers of all resulting structures and that of the crystallographic (reference) ligand were calculated for each run. If the distance between the center of the structure of minimum energy of a job and that of the reference was smaller than the length of the native ligand (limit) of the protein, then the current member of the matrix $\{e_2\}$ was set equal to $\{e_1\}$. Otherwise, the structure with the smallest distance was selected as a reference and a new $\{e_2\}$ was calculated. If all distances of the 10 structures were beyond the aforementioned limit, then the energy value was declared undefined for the actual $\{e_2\}$. (All matrices of this study are available upon request.)

VMD[24] and Raster-3D[25] programs were used for visualization and presentation of the results.

RESULTS AND DISCUSSION

Docking. The results of docking calculations of 44 protein−aromatic ligand complexes (Table 1) is presented in Table 2. Matches of some docked ligands to the crystallographic structure with various RMSD values are depicted on Figure 1. For 40 systems, the energy minimum structure obtained as a result of 10 runs (one docking job) was the closest to the crystallographic position of the ligand. In four cases, the second best ranked structure was the closest to the crystallographic structure. These molecules are marked with letter b in the RMSD(m) column. The RMSD of the energy minimum conformations from the original crystal position was less than 1 Å in 55%, less than 2 Å in 82%, and less than 3 Å in 100% of cases. Summarily, in 82% of the cases good fit was obtained, in the remaining cases the result was acceptable. The distribution of the results is similar (67%; 91%; and 100%) considering the average RMSD values calculated for the groups of structures having RMSD less than 2.5 Å ($\mathrm{RMSD}_{2.5}$ rank). The fraction of dockings of good match increased to 91% after ranking. In 86% of all systems, the number of docked ligand conformations in the $\mathrm{RMSD}_{2.5}$ rank was between 5 and 10, i.e., more than 50% of the runs of a job matched the crystal structure (including the minima of all but four jobs). This result indicates that jobs of 10 runs are adequate to get correct docking results for the systems studied by using the procedure described in the Methods section. The calculated AutoDock minimum

Table 1. Investigated Protein−Ligand Systems Ordered and Numbered According to the Increasing Molecular Weight (MW) of the Ligands[a]



	protein				ligand		
ID	name	PDB code	binding site residues	res. (Å)	-R groups	RT	MW
1	insulin	1mpj	ACHIL	2.30	$\mathbf{R_1}$: $-\mathrm{OH}$	1	94.1
2	insulin	1ev3	ACHIL	1.78	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_3}$: $-\mathrm{CH_3}$	1	108.1
3	insulin	1qiz	ACHIL	2.00	$\mathbf{R_1, R_3}$: $-\mathrm{OH}$	2	110.1
4a	transcription factor malt domain III	1hz4	HML	1.45	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$	0	121.1
4b	chloroperoxidase T	1a8u	FHLMSW	1.60	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$	0	121.1
4c	human peroxiredoxin	1hd2	CFILPRT	1.50	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$	0	121.1
4d	bacterial cocaine esterase	1ju4	FWY	1.63	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$	0	121.1
5a	β-trypsin	3ptb	C*DQSVY	1.70	$\mathbf{R_1}$: $-\mathrm{C(NH_2)_2^{(+)}}$	0	121.2
5b	urokinase-type plasminogen activator	1f5k	C*DSV	1.80	$\mathbf{R_1}$: $-\mathrm{C(NH_2)_2^{(+)}}$	0	121.2
6	bovine trypsin	1tnj	C*DV	1.80	$\mathbf{R_1}$: $-\mathrm{(CH_2)_2{-}NH_3^{(+)}}$	3	122.2
7a	beta-acrosin from RAM spermioza	1fiw	C*DQST	2.10	$\mathbf{R_1}$: $-\mathrm{C(NH_2)_2^{(+)}}$ $\mathbf{R_4}$: $-\mathrm{NH_2}$	0	136.2
7b	human coagulation factor IXA	1rfn	C*DS	2.80	$\mathbf{R_1}$: $-\mathrm{C(NH_2)_2^{(+)}}$ $\mathbf{R_4}$: $-\mathrm{NH_2}$	0	136.2
8	prostaglandin H2 synthase-1	1pth	AILRVY	3.40	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$ $\mathbf{R_2}$: $-\mathrm{OH}$	1	137.1
9	esterolytic and amidolytic 43C9 antibody (immunoglobulin)	43ca	FHQY	2.30	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_4}$: $-\mathrm{NO_2}$	1	139.1
10	insulin	1tym	AC*IL	1.90	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_4}$: $-\mathrm{NH{-}CO{-}CH_3}$	3	151.2
11	poly (ADP-ribose) polymerase	3pax	HSY	2.40	$\mathbf{R_1}$: $-\mathrm{CO{-}NH_2}$ $\mathbf{R_3}$: $-\mathrm{O{-}CH_3}$	3	151.2
12	phenylalanyl-tRNA synthetase	1b70	AEFHQRSVW	2.70	$\mathbf{R_1}$: $-\mathrm{CH_2{-}CH[COO^{(-)}]{-}NH_3^{(+)}}$	4	165.2
13	protocatechuate 3,4-dioxygenase	3pcn	HIPRTWY	2.40	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_3}$: $-\mathrm{CH_2{-}COO^{(-)}}$ $\mathbf{R_5}$: $-\mathrm{OH}$	4	167.1
14	human serum albumin	1e7a	C*FILNV	2.20	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_2, R_5}$: $-\mathrm{C_3H_7}$	3	178.3
15	macrophage migration inhibitory factor (MIF)	1ca7	FIKMNPSVWY	2.50	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_4}$: $-\mathrm{CH_2{-}CO{-}COO^{(-)}}$	3	179.1
16	des-(Ile318-Arg417)-tyrosyl-tRNA synthetase	4ts1	DLQTY	2.50	$\mathbf{R_1}$: $-\mathrm{OH}$ $\mathbf{R_4}$: $-\mathrm{CH_2{-}CH[COO^{(-)}]{-}NH_3^{(+)}}$	5	181.2
17	aromatic amino acid transferase	2ay5	DFILNRSTWY	2.40	$\mathbf{R_{12}}$: $-\mathrm{(CH_2)_2{-}COO^{(-)}}$	3	188.2
18	N1G9 FAB fragment	1ngp	HKRSWY	2.40	$\mathbf{R_1}$: $-\mathrm{NO_2}$ $\mathbf{R_3}$: $-\mathrm{CH_2{-}COO^{(-)}}$ $\mathbf{R_5}$: $-\mathrm{OH}$	4	196.1
19	prostaglandin H2 synthase-1	1eqg	AILRVYW	2.61	$\mathbf{R_1}$: $-\mathrm{CH(CH_3)COO^{(-)}}$ $\mathbf{R_4}$: $-\mathrm{C(CH_3)_3}$	3	205.3
20	protein tyrosine phosphatase 1B	1c85	CDFIKQSVY	2.72	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$ $\mathbf{R_2}$: $-\mathrm{NH{-}CO{-}COO^{(-)}}$	2	207.1
21	carboxypeptidase A	1hdu	AEHIR	1.75	$\mathbf{R_1}$: $-\mathrm{CH_2{-}CH[COO^{(-)}]{-}NH{-}CO{-}NH_2}$	5	207.2
22	tyrosine phosphatase	1d1q	CDFHLRW	1.70	$\mathbf{R_1}$: $-\mathrm{NO_2}$ $\mathbf{R_4}$: $-\mathrm{OPOO_2^{(2-)}}$	2	217.1
23	carboxypeptidase A	3cpa	ADEINRTY	2.00	$\mathbf{R_1}$: $-\mathrm{CH_2{-}CH[COO^{(-)}]{-}NH{-}CO{-}CH_2{-}NH_3^{(+)}}$ $\mathbf{R_4}$: $-\mathrm{OH}$	7	238.2
24	influenza virus B/LEE/40 neuraminidase (sialidase)	1ivb	DERY	2.40	$\mathbf{R_2}$: $-\mathrm{OH}$ $\mathbf{R_3}$: $-\mathrm{NH{-}CO{-}CH_3}$ $\mathbf{R_4}$: $-\mathrm{NO_2}$ $\mathbf{R_5}$: $-\mathrm{COO^{(-)}}$	4	239.2
25	protein tyrosine phosphatase 1B	1c83	ACDFIKQVY	1.80	$\mathbf{R_{10}}$: $-\mathrm{NH{-}CO{-}COO^{(-)}}$ $\mathbf{R_{11}}$: $-\mathrm{COO^{(-)}}$	3	246.2
26	protein tyrosine phosphatase 1B	1c84	ACDFIKQVY	2.35	$\mathbf{R_7}$: $-\mathrm{NH{-}CO{-}COO^{(-)}}$ $\mathbf{R_8}$: $-\mathrm{COO^{(-)}}$	3	257.2
27	cellobiohydrolase I	1dy4	DEHQRSTWY	1.90	$\mathbf{R_6}$: $-\mathrm{O{-}CH_2{-}}$*(S)*$\mathrm{CH(OH){-}CH_2{-}NH{-}CH(CH_3){-}CH_3}$	7	259.3
28	streptavidin	1sri	ALSVWY	1.65	$\mathbf{R_1}$: $-\mathrm{COO^{(-)}}$ $\mathbf{R_2}$: $-\mathrm{N{=}N{-}Ph(4\text{-}OH;\ 3,5\text{-}diMe)}$	4	271.3
29	indole-3-glycerolphosphate synthase	1a53	EFKLNRSW	2.00	$\mathbf{R_{12}}$: $-\mathrm{CH(OH){-}CH(OH){-}CH_2{-}O{-}POO_2^{(2-)}}$	7	285.2
30	protein-tyrosine phosphatase 1B	1bzj	ACFIRVY	2.25	$\mathbf{R_8}$: $-\mathrm{COO^{(-)}}$ $\mathbf{R_9}$: $-\mathrm{CF_2{-}POO_2^{(2-)}}$	2	299.1
31	48G7 hybridoma line FAB	1gaf	HLRWY	1.95	$\mathbf{R_1}$: $-\mathrm{NO_2}$ $\mathbf{R_4}$: $-\mathrm{OPOO^{(-)}{-}(CH_2)_4{-}COO^{(-)}}$	7	301.2
32	calcium binding domain VI of porcine calpain	1alw	FHIKLQRW	2.03	$\mathbf{R_1}$: $-\mathrm{I}$ $\mathbf{R_4}$: $-\mathrm{CH{=}C(SH)}$*(Z)*$\mathrm{\text{-}COO^{(-)}}$	4	305.1
33	protein tyrosine phosphatase 1B	1ecv	DFIKQRVY	1.95	$\mathbf{R_1}$: $-\mathrm{NH{-}CO{-}COO^{(-)}}$ $\mathbf{R_2}$: $-\mathrm{COO^{(-)}}$ $\mathbf{R_4}$: $-\mathrm{I}$	3	333.0
34	29G11 FAB	1a0q	FHKLRWY	2.30	$\mathbf{R_1}$: $-\mathrm{OPOO^{(-)}{-}CH(C_4H_9){-}NH{-}CO{-}(CH_2)_2{-}COO^{(-)}}$	10	341.3
35	catalytic antibody 28B4 fragment	1kel	FKNRWY	1.90	$\mathbf{R_1}$: $-\mathrm{NO_2}$ $\mathbf{R_4}$: $-\mathrm{CH_2{-}N[CH_2{-}POO_2^{(2-)}]{-}(CH_2)_4{-}COO^{(-)}}$	10	343.2
36	bovine trypsin	1az8	C*DQTV	1.80	$\mathbf{R_1}$: $-\mathrm{C(NH_2)_2^{(+)}}$ $\mathbf{R_3}$: $-\mathrm{CH(CH_2{-}COOCH_3){-}(CH_2)_2{-}Ph[4\text{-}C(NH_2)_2^{(+)}]}$	7	354.5
37	human estrogen receptor (α)	3ert	ADEILMTW	1.90	$\mathbf{R_1}$: $-\mathrm{C(C_2H_5){=}CPh(4\text{-}OH){-}Ph[4\text{-}O{-}(CH_2)_2{-}N(CH_3)_2]}$	9	387.5
38	glutathione S-transferase	1guh	DFLQRTVY	2.60	$\mathbf{R_1}$: $-\mathrm{CH_2{-}S{-}CH_2{-}CH[CO{-}NH{-}CH_2{-}COO^{(-)}]{-}NH{-}CO{-}(CH_2)_2{-}CH[COO^{(-)}]{-}NH_3^{(+)}}$	13	396.4
39	dihydrofolate reductase	4dfr	DFILKRT	1.70	$\mathbf{R_1}$: $-\mathrm{CO{-}NH{-}CH[COO^{(-)}]{-}(CH_2)_2{-}COO^{(-)}}$ $\mathbf{R_4}$: $-\mathrm{N(CH_3){-}CH_2{-}(4\text{-}aminopteridin\text{-}6\text{-}yl)}$	7	457.5

[a] The default −R groups are hydrogen atoms. Res: resolution of the protein structure. RT: the number of released torsions. C*: Cys in disulfide bridge.

docking energy and the free energy of binding was always lower than the energies corresponding to the original crystal structures. The decrease (or similarity) of the latter energy values indicates that the docked conformation is in an energy minimum, likewise to the crystal conformation. The free energies of binding of the minimum structures and the average free energy of binding of the members of the RMSD$_{2.5}$ rank were very close to each other, in some cases

Table 2. Verification of the Docking Method for the Investigated Protein−Ligand Complexes of the Present Study[a]

ID	E_d(c)	E_d(m)	ΔG(c)	ΔG(m)	RMSD(m)	N	ΔG(a)	SD	RMSD(a)	SD
1	−2.48	−3.61	−2.17	−3.29	0.64	10	−3.29	0.00	0.65	0.01
2	−3.91	−4.39	−3.60	−4.08	2.44	10	−4.08	0.00	2.44	0.00
3	−4.58	−5.31	−3.97	−4.67	2.53	0[c]				
4a	−3.68	−4.31	−3.68	−4.31	0.87	10	−4.31	0.00	0.87	0.00
4b	−5.80	−6.58	−5.80	−6.58	0.31	8	−6.58	0.01	0.29	0.01
4c	−4.66	−5.36	−4.66	−5.36	0.64	10	−5.36	0.00	0.64	0.00
4d	−4.82	−5.13	−4.78	−5.13	0.63	4	−5.12	0.01	0.64	0.01
5a	−7.89	−8.18	−7.89	−8.18	0.32	10	−8.18	0.00	0.32	0.00
5b	−7.28	−7.93	−7.28	−7.93	1.99[b]	9	−7.63	0.01	1.00	0.75
6	−7.15	−8.63	−6.35	−7.72	2.13	10	−7.70	0.01	2.13	0.01
7a		−9.29		−9.29	0.83	8	−9.29	0.00	0.83	0.00
7b	−8.33	−8.79	−8.33	−8.79	0.61	9	−8.79	0.00	0.61	0.00
8	−3.82	−5.61	−3.50	−4.98	2.70[b]	4	−4.46	0.00	2.36	0.01
9	−5.33	−5.64	−5.01	−5.31	0.66	10	−5.31	0.01	0.67	0.00
10	−1.83	−4.91	−1.39	−3.90	1.74	10	−3.88	0.01	1.72	0.02
11	−5.39	−7.20	−5.73	−6.20	1.30	9	−6.09	0.17	1.57	0.50
12	−7.93	−10.36	−6.83	−8.89	0.84	10	−8.88	0.01	0.91	0.05
13	−4.73	−6.81	−3.48	−5.42	2.64	2	−5.23	0.01	1.07	0.02
14	−5.04	−6.59	−5.28	−6.20	1.05	10	−6.20	0.00	1.05	0.01
15		−8.93		−8.08	0.96	10	−8.01	0.05	0.90	0.07
16	−6.78	−8.72	−5.35	−7.14	0.41	7	−7.05	0.09	0.75	0.77
17	−9.36	−10.77	−8.97	−9.92	2.10	9	−9.51	0.22	1.07	0.58
18	−7.22	−9.32	−6.51	−8.17	0.77[b]	6	−7.68	0.11	0.74	0.03
19	−8.35	−8.81	−7.39	−7.73	0.81	9	−7.72	0.00	0.79	0.02
20	−9.15	−11.31	−8.99	−10.31	1.03	9	−10.30	0.02	0.98	0.06
21	−9.27	−10.49	−8.06	−8.84	0.59	10	−8.80	0.03	0.58	0.02
22	−11.15	−11.79	−10.54	−11.18	0.80	10	−11.17	0.01	0.97	0.39
23	−4.65	−8.45	−6.77	−8.91	1.08	10	−8.70	0.21	1.12	0.15
24	−6.50	−7.60	−6.38	−6.67	0.34	8	−6.65	0.02	0.37	0.06
25	−10.85	−13.09	−10.78	−11.87	0.65	9	−11.81	0.05	0.70	0.04
26	−10.32	−11.92	−10.21	−10.85	1.70	9	−10.79	0.06	0.89	0.33
27	−10.61	−11.26	−8.68	−9.12	0.41[b]	6	−8.70	0.17	0.87	0.49
28	−0.14	−8.17	−2.10	−9.47	1.05	7	−9.33	0.06	0.99	0.05
29	−10.56	−11.77	−9.79	−10.50	1.31	4	−10.21	0.42	1.19	0.23
30	−13.78	−14.25	−12.60	−13.07	0.38	10	−13.06	0.01	0.33	0.03
31	−10.98	−12.80	−8.57	−10.37	0.58[d]	8	−10.10	0.44	0.72	0.32
32		−6.94		−5.66	1.51[e]	9	−5.62	0.03	1.83	0.31
33	−10.72	−12.85	−10.56	−11.59	0.76	8	−11.57	0.02	0.76	0.01
34	−10.72	−13.36	−7.47	−9.64	2.61	3	−9.25	0.25	2.15	0.29
35	−12.73	−15.85	−10.71	−12.59	2.01	5	−12.47	0.15	1.57	0.35
36	−12.75	−14.06	−11.87	−12.30	0.51	10	−12.22	0.06	0.59	0.27
37	−1.43	−7.23	−8.30	−10.21	1.81	7	−9.66	0.36	1.42	0.18
38	−14.88	−16.38	−10.82	−12.14	0.77	5	−11.94	0.18	0.86	0.14
39	−11.71	−14.47	−11.17	−13.77	1.22	5	−13.70	0.13	1.23	0.03

[a] E_d(c): docked energy of the crystallographic ligand (kcal/mol); E_d(m): docked energy of the docked conformation of the lowest free energy of binding of the ligand; ΔG(c): calculated free energy of binding of the crystallographic ligand (kcal/mol); ΔG(m): free energy of binding of the lowest free energy of the ligand; RMSD(m): root-mean-square deviation of the docked conformation of the lowest free energy of binding of the ligand (Å); N: number of conformations in the $\mathrm{RMSD}_{2.5}$ rank (the crystallographic structure was used for reference, see text); ΔG(a): average free energy of binding of the $\mathrm{RMSD}_{2.5}$ rank; RMSD(a): average RMSD of the $\mathrm{RMSD}_{2.5}$ rank; SD: standard deviation. [b] The RMSD value corresponds to the minima of the second best group of docked ligands. [c] The RMSD of docked conformations were only slightly above 2.5 Å (acceptable fit) but $\mathrm{RMSD}_{2.5}$ rank was not defined in this case. [d] Parts of the ligand of high B-factors were omitted during RMSD calculation. [e] The crystallographic ligand had erroneous structure. RMSD was calculated omitting the erroneous part.

they were equal. This similarity between the minimum and average energies corresponds to the structural similarity of the members of $\mathrm{RMSD}_{2.5}$ rank.

In summary, the docking results of this section provide verification for the further investigations on the 44 systems using our standard docking protocol. It should be noted, that at systems **8** and **19**, an octylglucoside molecule covers the binding site. The proteins of these systems were not involved in further studies of the 31 different proteins, as the site seems to be specifically arranged for the original ligands. See also Methods for details on selection of the systems for the study.

As a further test of reproducibility of the docking results on the investigated systems, 31 jobs of different complexes were repeated using new seed parameters for the random number generator at each job. Despite the total randomization of starting positions of the ligand, excellent reproducibility was obtained (Supporting Information, Figure A). This finding strengthens reliability of data in the **E** matrices.

Molecular Interaction Fingerprints (MIFs). The following subtraction was applied to the rows of the **E** matrices defined in the Methods section, to obtain the $\mathbf{MIF_Y}$ matrices

$$\mathbf{MIF_Y} = \mathbf{E_Y} - (\mathbf{REF_{MIF}})^{\mathrm{T}} \qquad (1)$$

where Y = 1 or 2, denoting the two kinds of methods of evaluation, and $\mathbf{REF_{MIF}}$ is the vector of reference energies of each ligand. The reference energies correspond to the binding of the 39 different ligand molecules in complex with their original proteins. Thus, $\mathbf{MIF_Y}$ matrices contain rows, i.e., the molecular interaction fingerprints with values of

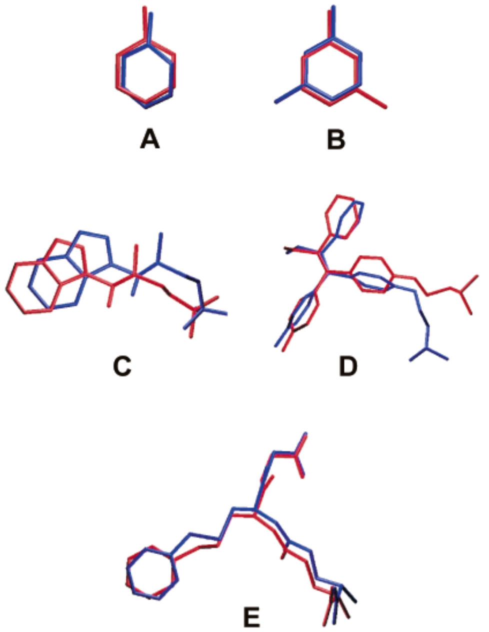


Figure 1. Comparison of docked (energy minimum; blue), and the crystallographic conformations (red) of different RMSD values. A: ligand **1**; B: ligand **2**; C: ligand **29**; D: ligand **37**; E: ligand **38**. Note that in case of small ligands, like **2**, the alteration of positon of only one methyl group causes significantly larger RMSD, due to the small number of atoms, even if the positions of other atoms match with those of the reference ligand.

relative binding free energy of each aromatic compound and the 31 different proteins. The familiar term "fingerprint" was used here to indicate that the interaction profile of each row is specific to only one compound.

A central problem of screening of proteins is the choice of binding energy threshold for selection of the appropriate targets. For this purpose, Chen et al.[13] used an energy value calculated from the correlation with the number of atoms of the ligand as a threshold. However, only an $R^2 \approx 0.5$ can be obtained when correlating binding free energies with the number of heavy atoms of ligand molecules (Supporting Information, Figure B). While the correlation with the number of atoms alone seems to be insufficient for estimation of binding free energies even for a group of similar compounds, the use of a threshold was avoided in our study.

Using the MIF approach, experimental data (binding free energies calculated with the reference structures) form a basis of the relative comparison of the binding affinities of the 31 proteins. For ligands having several crystal complexes, the minima of the individual binding energies could be used as a more precise reference.

The resulting $\mathbf{MIF_1}$ matrix is presented in Figure 2a. According to eq 1, negative or zero values indicate the possibility of interaction between the ligand and the proteins.

In protein screening studies, a further critical point is the verification of the results. It may be rather difficult, considering the expenditure required for determination of the X-ray structures of all possible (in this study: 1209) combinations of ligands and proteins or measurement of experimental binding free energies (or inhibition constants). In some cases, indirect biochemical data are available that relate the activity of selected proteins to drug side effects or toxicities of the compounds.[13,14] However, validation of only the positive results is possible, i.e., when drugs are involved in considerable interaction with the proteins.[14] For biologically inactive compounds (negative result) no or very few data are available.

In the present study, structural considerations were used to perform a check on the aforementioned positive selection. The principle of fragment docking[26-29] and fragment-based drug design[30] asserts that fragments of certain molecules bind to the same sites as the whole molecules. Here, this simple principle was applied to select out the proteins, which are a priori targets of a given compound, having the same (boxes of solid outline on Figure 2a) or similar (boxes of dashed outline) fragments as the native ligand. A definition by Wang et al.[27] was used to classify the "same" fragments of the investigated ligands. For example, salicylic acid (**8**) and 2-(oxalylamino)benzoic acid (**20**) have the same fragment (benzoic acid), the 4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid (**24**) was considered only "similar" to salicylic acid in this study, having more than one group linked to the benzene ring.

Comparison of the location of boxes of Figure 2a, and the corresponding energy values showed good agreement: the proteins (sites) selected by fragment considerations (boxes) have, with few exceptions, zero or negative energy value (Figure 2a). The fragment-based comparison was made for ligands **1**−**14**. As larger compounds have site-specific or bulky groups attached to the aromatic rings, the simple fragment considerations are not as straightforward as they were at the small molecules. For example, ligand **28** and protein sites **4a**−**b** hardly match even if benzoic acid is a common "fragment" of ligands **28** and **4**, because the attached group makes molecule **28** considerably larger than benzoic acid. However, in some cases striking agreement with the rational fragment or similarity considerations can still be found, for larger compounds, as well. One example is the match of tyrosine (**16**) to the site of native ligand phenylalanine (**12;** Figure 2a, box with red outline): the two compounds differ only by a hydroxyl group. A further nice match is that of ibuprofen (**19**) and the site of ligand **17**. Both ligands have carboxyl groups and large hydrophobic parts (isobutylphenyl and indole groups, respectively) arranged at similar distance. There are some other systems, which have negative value in $\mathbf{MIF_1}$ and high degree of similarity between the ligands. One example is ligand **13**, which fits to the binding site of **17**: both ligands have carboxyl groups linked to the benzene ring (the linker differs only in one $-CH_2-$ unit), which play a pivotal role in forming the complexes. Furthermore, the existence of identical binding site residues I, R, T, W and Y (Table 1) at both proteins indicate the similarity of binding pockets (like at the other pairs of systems discussed above). In conclusion, the rows of $\mathbf{MIF_1}$ matrix contain reasonable "interaction fingerprints" of the ligand molecules, which are in good agreement with fragment and similarity considerations.

Interestingly, there is a clear trend in $\mathbf{MIF_1}$ toward the higher selectivity for larger ligands. The only exception is ligand **32**, iodobenzene. This trend seems reasonable as ligands of smaller molecular weight are fragments of the larger molecules in many cases and therefore can fit to the sites of the proteins corresponding to larger ligands. At the same time, larger molecules have more specific groups to

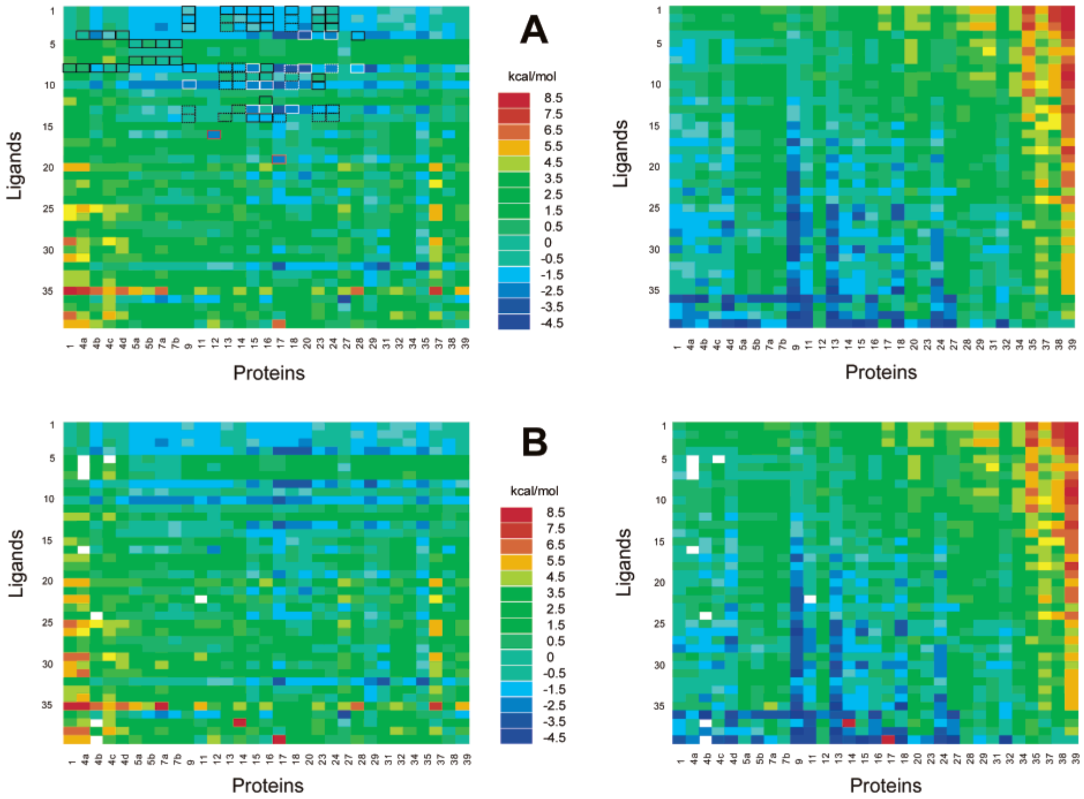


Figure 2. $\mathbf{MIF_1}$ (A), $\mathbf{MIF_2}$ (B), $\mathbf{MAF_1}$ (C), and $\mathbf{MAF_2}$ (D) matrices. The rows are ordered according to the serial number (molecular weight) of the ligand molecules. See text for meaning of other notations. White boxes mean undefined values. Solid (and dashed) outlines mark boxes with same (or similar) fragments of the ligand of the actual protein and the ligand of the actual row. Red outlines correspond to similarities, as well. (See text for details on the latter ones.) The reference crystal structures correspond to boxes of zero energy values and positioned roughly on the diagonal of the matrix. The rows of each matrix are named "fingerprints", including specific data for the ligands.

interact with their specific target and/or there is no space to fit them in the sites of smaller ligands. (In case of system **32**, the heavy iodine atom does not contribute to the specificity relatively as much as it increases the molecular weight. Hence, this compound is an exception to the trend.)

In matrix $\mathbf{MIF_2}$, a similar trend can be observed (Figure 2b). In this case, the docked structures situated far from the center of the subsite of the original crystallographic ligand were filtered out (see Methods for details). In 131 of 1209 cases (11%) the values of $\mathbf{E_1}$ were changed during the generation of $\mathbf{E_2}$, due to the large distance from the reference subsite. In nine cases white boxes depict in the figure, that in all cases the resulting structures were far from the original subsite. In some simple cases a rational explanation can be given for the appearance of such deviations. For example, in case of the benzamidine derivatives and 2-phenylethylamine (**5**, **6**, and **7**), no match was found with the subsite of **4a**, a benzoic acid target (Figure 2b). This subsite, attracting benzoic acid, is repulsive for the positively charged compounds. This repulsion causes a shift of e.g. ligand **5** far away from the original **4a** site (Figure 3) and the appearance of the white box in the $\mathbf{MIF_2}$ matrix. In cases when only the subsite of a protein pocket is of interest, the filtering used to obtain $\mathbf{MIF_2}$ could be essential. For instance, this situation occurs e.g. when the other parts of the pocket are not involved in the key interactions or interact with water molecules. Hence, this filtering helps to avoid overestimating to actual role of the positive interactions with the inactive part of the pocket and thus of small importance. The rows (MIFs) of the $\mathbf{MIF_2}$ matrix reflect relative structural and energy information at the same time and contain therefore more information than $\mathbf{MIF_1}$ (or methods used in refs 13 and 14), which have not applied e.g. a distance criterion to filter ligands bound to nonactive subsites.

Molecular Affinity Fingerprints (MAFs). While the rows of the $\mathbf{E_Y}$ matrices form the basis of the calculation of MIFs, the columns can be considered as "slices" of conventional in silico HTS, which is performed routinely in drug design for a protein site. A subtraction, similar to (1)

$$\mathbf{MAF_Y} = \mathbf{E_Y} - \mathbf{REF_{MAF}} \qquad (2)$$

was performed for the columns to obtain matrices of molecular affinity fingerprints ($\mathbf{MAF_Y}$). ($\mathbf{REF_{MAF}}$ is not completely identical with $\mathbf{REF_{MIF}}$, as the former one has 39 elements instead of the 31.) In this case, the columns of $\mathbf{MAF_Y}$ contain relative energies of ligands for each protein site. However, also the rows of the matrix have valuable meaning, describing the competitive affinity of a compound for the different protein sites relative to the original crystallographic ligand of the site. Therefore, the rows of $\mathbf{MAF_Y}$ were called the molecular affinity fingerprints (MAFs) of the compounds. The diagonal of $\mathbf{MAF_1}$ (Figure 2c) divides the MAF values roughly into two groups: the negative values

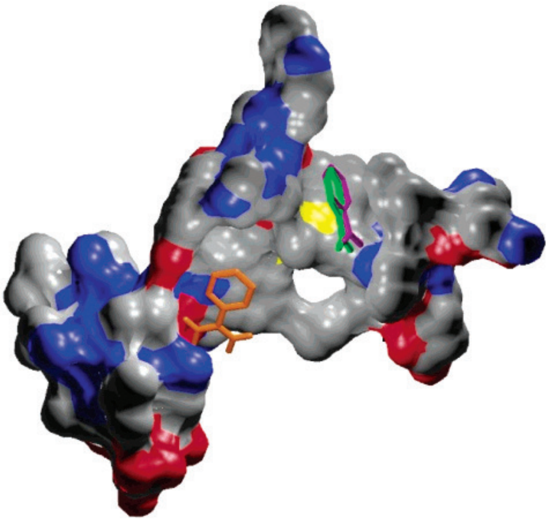


Figure 3. Solvent accessible surface representation of the binding pocket of Transcription factor Malt domain III (protein **4a**). The crystallographic positions of the original ligand benzoic acid (**4**), its docked conformation, and the docked benzamidine (**5**) are marked with green, purple, and orange colors, respectively. The match of the docked and the crystallographic benzoic acid is excellent. The benzamidine molecule is at a distance of ca. 10 Å from the subsite of benzoic acid. Benzamidine is attracted by the negatively charged (red) subsite and repulsed by the H-donor (blue) subsite of benzoic acid. The free energy of binding of benzamidine is smaller (−5.41 kcal/mol) than that of benzoic acid (−4.31 kcal/mol). Thus, the selection of the appropriate binding pocket may be erroneous, using only the free energies as filters. The distance criterion at **MIF$_2$** (see text) decreases the number of the wrongly selected pockets or proteins.

are gathered at the left bottom corner and the positive values are at the right top corner of the matrix. This finding indicates that smaller compounds are weaker competitors at foreign sites, while the relative competitive affinity of the larger compounds is better at foreign sites, as well. However, beyond system **28**, the competitive efficiency seems to decrease for any compound. This limit indicates that if the binding site is large and has a native ligand of more than ca. 300 Da, then the competitive strength of other ligands at the site is smaller. The quantitative results presented in this section are in agreement with the rational considerations: small compounds with fewer interacting points are plausibly weak competitors at sites of larger ligands. The crude trend of decreasing binding free energies with an increasing number of atoms of ligands (see Section “Molecular Interaction Fingerprints” for details) indicates a similar trend, as well. Summarily, the MAFs reflect a realistic picture of the competition.

The **MAF$_2$** matrix (Figure 2d) contains changes at similar positions as **MIF$_2$**. In some cases, a large shift to the positive values can be observed (e.g. at ligand **37** − protein **14** or ligand **39** − protein **17**) after the second filtering.

CONCLUSIONS

In the present study, a docking approach of protein screening was investigated. Based on a validated set of reproduced crystal complexes, relative measures of molecular interaction and affinity fingerprints were introduced and used for the comparison of ligand-protein selectivity. MIF and MAF patterns were found to be in good agreement with rational considerations. As the generation of the corresponding matrices is relatively fast (one docking job took an average of a half an hour of CPU time on a PIV 1.7 GHz processor), the data set can be expanded with further proteins and ligands in an effective way. Furthermore, using the same scoring function, the results on 1209 complexes remain comparable. This statement is highly important, as the number of the available experimental inhibition constants or binding free energy values is very limited, and sometimes the measured values are not comparable with each other due to the different experimental setups.[10] Thus, protein screening would be impossible if to use only the available experimental binding affinity values. An additional advantage of the computational docking approach is that the docked complexes can be precisely compared to the original crystal complex of the site and in the case of the other ligands, structural information of the place of subsite can be involved in the evaluation (filtering) together with the corresponding energy values. Moreover, if no data exist for the reference crystal complex, the binding site can be found by methods based on pocket search[31] or by blind docking.[5]

However, general problems of docking applications, protein flexibility[32] and the role of structural waters,[5,18] should be considered in future protein screening studies. The error of docking caused by structural waters can be reduced with the above-mentioned filtering using subsite information. The elimination of the induced effects is more problematic.[33]

Practically, MIFs can be used for comparison (selection) of protein targets involved in drug side effects and toxicity, while MAFs contain information on the competitive affinity of a compound at a site and therefore may be applied for relative estimation of the possibility of interference of a drug with others at their corresponding binding sites (proteins).

Generally, MIFs and MAFs calculated on the basis of a larger free energy database may aid the exploration of the appropriate biochemical interaction route of any compounds, e.g. by automated comparison of their fingerprints with those of ligands of elucidated biochemistry.

ACKNOWLEDGMENT

This work was supported by a grant from the IMAGETOX EU Research and Training Network: Contract No. HPRN-CT-1999-00015. Support from Estonian Science Foundation is also acknowledged (Grant No. 4571).

Supporting Information Available: Job-by-job reproducibility of 31 docking jobs performed on different protein(-ligand) systems (Figure A) and the trend between the calculated binding free energies and the number of heavy atoms of the ligands (Figure B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nature Rev.* **2002**, *1*, 45−54.

(2) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.

(3) Böhm, H. J.; Stahl, M. Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2000**, *4*, 283−286.

(4) Dennis S.; Körtvélyesi, T.; Vajda, S. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290−4295.

(5) Hetényi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **2002**, *11(7)*, 1729−1737.

(6) Hetényi, C.; Szabó, Z.; Klement, É.; Datki, Z.; Körtvélyesi, T.; Zarándi, M.; Penke, B. Pentapeptide amides interfere with the aggregation of beta amyloid peptide of Alzheimer's disease. *Biochem. Biophys. Res. Comm.* **2002**, *292,* 931−936.

(7) Hetényi, C.; Körtvélyesi, T.; Penke, B. Mapping of the possible binding sequences of two beta-sheet breaker peptides on beta amyloid peptide of Alzheimer's disease. *Bioorg. Med. Chem.* **2002**, *10(5)*, 1587−1593.

(8) Halperlin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.

(9) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Design* **2002**, *16*, 151−166.

(10) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.

(11) Marelius, J.; Hansson, T.; Åqvist, J. Calculation of ligand binding free energies from molecular dynamics simulations. *Int. J. Quantum Chem.* **1998**, *69*, 77−88.

(12) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19(14)*, 1639−1662.

(13) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43*, 217−226.

(14) Rockey, W. M.; Elcock, A. H. Progress toward virtual screening for drug side effects. *Proteins* **2002**, *48*, 664−671.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Databank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(16) Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **2001**, *29*, 221−222.

(17) Minke, W. E.; Diller, D. J.; Hol, W. G. J.; Verlinde, C. L. M. J. The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test-case. *J. Med. Chem.* **1999**, *42*, 1778−1788.

(18) Pang, Y.-P.; Perola, E.; Xu, K.; Prendergast, F. G. EUDOCK: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **2001**, *22(15)*, 1750−1771.

(19) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and postprocessing program for molecular and electronic structures *J. Comput.-Aided Mol. Design* **2000**, *14*, 123−134.

(20) Pappu, R. V.; Hart, R. K.; Ponder, J. W. Analysis and application of potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B* **1998**, *102*, 9725−9742.

(21) Walters, P.; Dolata, M. S. Babel − A Molecular Structure Information Interchange Hub. Department of Chemistry, University of Arizona, Tucson, AZ 85721. (http://smog.com/chem/babel/).

(22) Pedretti, A.; Villa, L.; Vistoli, G. Vega: a versatile program to convert, handle and visualize molecular structure on windows-based PCs *J. Mol. Graph.* **2002**, *21*, 47−49.

(23) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(24) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.

(25) Merritt, E. A.; Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **1997**, *277*, 505−524.

(26) Friedman, A. R.; Roberts, V. A.; Tainer, J. A. Predicting molecular interactions and inducible complementarity: fragment docking of Fab-peptide complexes. *Proteins* **1994**, *20*, 15−24.

(27) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: A multistep strategy approach. *Proteins* **1999**, *36*, 1−19.

(28) Jackson, R. M. Q-fit: A probabilistic method for docking molecular fragments by sampling low energy conformational space. *J. Comput.-Aided Mol. Design* **2002**, *16*, 43−57.

(29) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins* **1999**, *37*, 88−105.

(30) Lesuisse, D.; Lange, G.; Deprez, P.; Benard, D.; Schoot, B.; Delettre, G.; Marquette, J. P.; Broto, P.; Jean-Baptiste, V.; Bichet, P.; Sarubbi, E.; Mandine, E. SAR and X-ray. A new approach combining fragment-based screening and rational drug design: Application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem.* **2002**, *45*, 2379−2387.

(31) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Design* **2000**, *14(4)*, 383−401.

(32) Fradera, X.; de la Cruz, X.; Silva, C. H. T. P.; Gelpi, J. L.; Luque, J. F.; Orozco, M. Ligand-induced changes in the binding sites of proteins. *Bioinformatics* **2002**, *18(7)*, 939−948.

(33) Carlson, H. A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447−452.

CI034052U

hetenyi.csaba_83_23

D24

hetenyi.csaba_83_23

JCTC Journal of Chemical Theory and Computation

ACS AuthorChoice

Article

pubs.acs.org/JCTC

Terms of Use

Quantification of Solvent Contribution to the Stability of Noncovalent Complexes

Haiyang Zhang,[†,‡] Tianwei Tan,[†] Csaba Hetényi,[§] and David van der Spoel[‡,*]

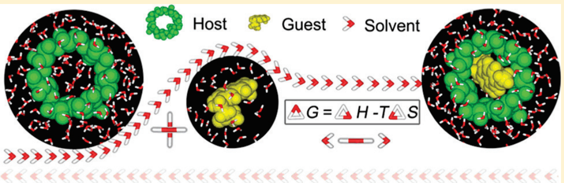
[†]Beijing Key Laboratory of Bioprocess, Department of Biochemical Engineering, Beijing University of Chemical Technology, Box 53, 100029 Beijing, China

[‡]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, Box 596, SE-751 24 Uppsala, Sweden

[§]Molecular Biophysics Research Group, Hungarian Academy of Sciences, Pázmány sétány 1/C, H-1117 Budapest, Hungary

S Supporting Information

ABSTRACT: We introduce an indirect approach to estimate the solvation contributions to the thermodynamics of noncovalent complex formation through molecular dynamics simulation. This estimation is demonstrated by potential of mean force and entropy calculations on the binding process between *β*-cyclodextrin (host) and four drug molecules puerarin, daidzin, daidzein, and nabumetone (guest) in explicit water, followed by a stepwise extraction of individual enthalpy (Δ*H*) and entropy (Δ*S*) terms from the total free energy. Detailed analysis on the energetics of the host–guest complexation demonstrates that flexibility of the binding partners and solvation-related Δ*H* and Δ*S* need to be included explicitly for accurate estimation of the binding thermodynamics. From this, and our previous work on the solvent dependency of binding energies (Zhang et al. *J. Phys. Chem. B* **2012**, *116*, 12684–12693), it follows that calculations neglecting host or guest flexibility, or those employing implicit solvent, will not be able to systematically predict binding free energies. The approach presented here can be readily adopted for obtaining a deeper understanding of the mechanisms governing noncovalent associations in solution.



INTRODUCTION

Correct estimation of thermodynamic parameters governing supra-molecular complexation from empirical calculations is of crucial importance for a better understanding of processes in biomolecules and for virtual screening in structure–function analysis and molecular design. For a host–guest complex both enthalpic and entropic contributions from the binding partners and their environment determine the overall binding free energy. Solvent acts not solely as an inert, bulk medium but also as an active partner during the noncovalent complexation. Various methods have been published to evaluate binding free energy profiles, such as molecular mechanics–Poisson–Boltzmann surface area (MM–PBSA),[1] thermodynamic integration (TI),[2] free energy perturbation (FEP),[3] and potential of mean force (PMF) calculations.[4] However, evaluation of solvation enthalpy as well as configurational entropy contributions still remains a challenge, in particular for large biomolecules. Simplified treatments, such as using implicit solvent models based on, for example, atomic fragmental volumes and solvation parameters[5] or treating the receptor as a rigid body in whole or in part, have been proposed to enable high-throughput virtual screening with the aid of docking techniques.[6,7] Efforts to improve the accuracy of scoring functions by including the effects of solvation and receptor flexibility continue as well.[8]

Cyclodextrins (CDs) are ideal candidates for host (or target) molecules, and they have attracted much attention over the years, particularly because of their pharmaceutical applications in drug delivery.[9] The lipophilic cavity and hydrophilic surface of CDs also provide an enzyme-like environment allowing to mimic protein–ligand interactions.[10] Between natural CDs and guest molecules van der Waals, hydrophobic, and hydrogen bond interactions are major driving forces responsible for the host–guest complexation.[11] Release of strain energy in the CD macrocycle and of "high-energy" (also known as enthalpy-rich) water from the CD cavity upon complexation has been suggested to contribute to the binding as well.[12,13] Induced conformational changes of CDs upon binding to a guest have been proposed and detected by experimental and theoretical studies.[14−19] Inoue et al. reported a compensatory enthalpy–entropy relationship in [CD:guest] complexes, based on thermodynamic measurements of CDs with a series of guest molecules via calorimetric titration.[20,21] They stated that steric hindrance in the complex formation may lead to an entropy loss and cancel out the enthalpy gain in part.[21] These observations indicate that solvation-related changes such as desolvation and/or configurational fit play a role in [CD:guest] associations and must be taken into account during calculation of the complexation thermodynamics. Although a number of studies involving free energy calculations of CD-containing

Received: May 16, 2013
Published: September 3, 2013

complexes have been published,[22−27] few reports focus on the solvation problem mentioned above.

Here, we introduce an indirect approach for quantification of solvent contribution to the energetics of noncovalent complexation by molecular dynamics (MD) simulation. This approach is demonstrated on the complex formation between β-CD and four drug molecules (puerarin, daidzin, daidzein, and nabumetone) as host and guest molecules, respectively, using water as explicit solvent. These four drug molecules have been reported to possess potential medicinal values.[28−30] Steered molecular dynamics (SMD)[31] was used to generate a formation process of the [CD:guest] complex, along which potentials of mean force (PMFs, i.e., free energy profiles) were computed with umbrella sampling.[32] More details on the SMD and PMF techniques are given in refs 24, 31, and 33−37. On the basis of PMF calculations, the total enthalpy and entropy change are evaluated and further decomposed into individual items in order to quantify the energetics of binding in detail. The results assist in understanding thermodynamic properties of biological processes such as drug encapsulation and release from CDs. Implications for prediction of receptor−ligand binding affinities in general are discussed at the end of this paper.

■ METHODS

Simulation Setup. The initial coordinates of the β-CD (host) were extracted from the RCSB protein data bank (PDB code: 1DMB). Drug molecules of puerarin, daidzin, daidzein, and nabumetone (guests) were constructed using the Chem3D software. Structures of the host and guest molecules are shown in Figure 1. The q4md-CD force field was used to model β-CD; this force field has been validated for CD-based systems[38] and for use[39] in the GROMACS suite.[40,41] The generalized Amber force field (GAFF)[42] was chosen to parametrize the guest molecules. Restrained electrostatic potential (RESP)[43] charges of guest molecules were derived by fitting partial charges to electrostatic potentials calculated using Gaussian 03[44] at the HF/6-31G* level of theory. Puerarin, daidzin, and daidzein complexes were simulated at 300 K and nabumetone at 293 K to allow direct comparison with experimental data. Constraints were applied for bond lengths of host and guest molecules with the LINCS algorithm,[45] and for bond lengths and angles of water molecules with SETTLE,[46] allowing a time step of 2 fs. All the simulations were performed with the TIP3P water model,[47] using the GROMACS package (version 4.5.5).[40,41] Long-range electrostatic interactions were treated using the particle mesh Eward (PME) approach[48,49] with a switching distance of 1.0 nm. Further details of the simulation protocol have been presented in ref 16.

Each system contained one host, one guest, and approximately 3300 water molecules in a cubic box of $5 \times 5 \times 4$ nm^3. The host molecule was centered in the box with the Z-coordinate of its seven glycosidic oxygen atoms approximately located at $Z = 2$ nm with the cavity axis of β-CD parallel to the Z-axis. The distance between the center of mass (COM) of the B-ring of the guest and that of the seven glycosidic oxygens of β-CD along the Z-axis was defined as the reaction coordinate ξ (Figure 2). The initial (i) and final (f) values of the reaction coordinate were set to $\xi_i = -2$ nm and $\xi_f = 2$ nm, respectively. Prior to each production we performed an equilibration simulation of 200 ps in which the pressure was maintained at 1 bar with the semi-isotropic Parrinello−Rahman barostat,[50] scaling the box in the X−Y plane only but keeping the box size in the Z-direction fixed. During production simulations the box size was kept unchanged with no pressure coupling. A periodic pulling simulation was carried out in GROMACS,[40,41] allowing the distance to be larger than half the box size, to obtain a formation event of 1:1 [β-CD:guest] complexes. The seven glycosidic oxygen atoms of β-CD were harmonically restrained with an isotropic force constant of 1000 kJ mol^{-1} nm^{-2} and used as an immobile reference for pulling simulations. The B-ring of the guest was pulled through β-CD cavity from the primary or secondary rim, corresponding to the BP or BS arrangement in Figure 1c, respectively, along the Z-axis over 800 ps with a harmonic force constant of 2000 kJ mol^{-1} nm^{-2} and a pulling rate of 0.005 nm ps^{-1}. In some cases the guest did not go inside but rather outside the cavity, giving a BO arrangement (Figure 1c). The COM distance and reaction coordinate as a function of the simulation time for these three arrangements of [β-CD:puerarin] complexes are shown in Figure S1 in the Supporting Information. Finally the guest sampled 4 nm covering the entire [ξ_i, ξ_f] interval. In the [ξ_i, ξ_f]

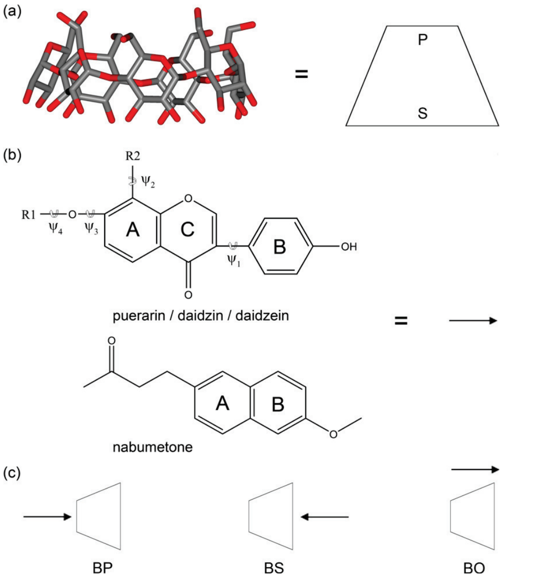


Figure 1. (a) Stick model of β-CD. Hydrogen atoms are omitted for clarity. Primary and secondary hydroxyls are situated at the primary (P) and secondary (S) rim, respectively. (b) Molecular structure of puerarin (R1 = H, R2 = Glucose), daidzin (R1 = Glucose, R2 = H), daidzein (R1 = H, R2 = H), and nabumetone. A, B, and C denote relevant ring groups. Four dihedral angles (ψ_i, i = 1...4) involving non-hydrogen atoms are defined here to describe guest rotations around corresponding bonds. (c) Structural arrangement of the [β-CD:guest] complex formation. BP indicates B-ring of guest inserting into β-CD cavity from the P rim; BS from the S rim. BO means the B-ring locating outside the cavity.

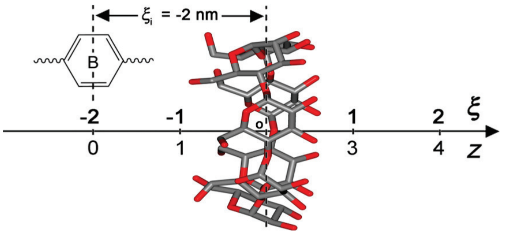


Figure 2. Definition of the reaction coordinate ξ.

reaction coordinate interval we selected 81 windows with a distance of 0.05 nm between adjacent positions and these windows were then used for umbrella sampling simulations. Following the same scheme, we simulated four guest molecules with three different arrangements and therefore obtained 12 potential of mean force (PMF) profiles in total. In order to detect the ultimate entropy loss of a guest inside a rigid cavity, one more PMF for the [β-CD:nabumetone] complex with the BS arrangement was computed with position restraints of all the non-hydrogen atoms of β-CD. The total simulation time for a single PMF profile was 810 ns (10 ns for each window).

Thermodynamic Calculation. After removing the first 2 ns for equilibration, we constructed the PMFs with a periodic version of the weighted histogram analysis method (WHAM).[51,52] As noted by Kumar and co-workers,[51] the integrated autocorrelation times of the umbrella windows were incorporated into the WHAM iteration procedure to yield a more accurate estimate for the PMF, in particular for a periodic PMF in nonhomogeneous systems.[52] Statistical uncertainties of the PMFs were estimated using the Bayesian bootstrap of complete histograms.[52] All the PMFs were defined to zero at ξ_i and ξ_f where host−guest interactions vanish, and thus, we can quantify the free energy difference (ΔG) with respect to the separated state of the binding partners.

The simulated system was first equilibrated at 1 bar and then the volume was kept constant, so the enthalpy of the system roughly amounts to its internal energy. The temperature is controlled throughout our simulations and thus the kinetic energy has a constant contribution to the internal energy. The enthalpy change (ΔH) therefore reasonably equals the potential energy difference with respect to a completely separated state between host and guest (eq 1).[37] Note that all

$$\Delta H(\xi) = V(\xi) - V(\xi_i) \tag{1}$$

thermodynamic variables are functions of ξ. For simplicity, we omit this functional dependence in the forthcoming text. The entropy change (ΔS) of the system was then computed by subtracting the ΔH part from ΔG (eq 2).

$$-T\Delta S = \Delta G - \Delta H \tag{2}$$

An enthalpic profile of the system was further decomposed into eight terms (eq 3) where the

$$\begin{aligned}\Delta H = {} & \Delta H_{\text{host}} + \Delta H_{\text{guest}} + \Delta H_{\text{host-host}} + \Delta H_{\text{guest-guest}} \\ & + \Delta H_{\text{sol-sol}} + \Delta H_{\text{host-guest}} + \Delta H_{\text{host-sol}} \\ & + \Delta H_{\text{guest-sol}}\end{aligned} \tag{3}$$

first two terms contain bonded interactions (bond angle and dihedral angle) and the rest are intra- and intermolecular nonbonded interactions. The bond stretching terms of host and guest molecules amount to zero since all the bond lengths were constrained during the simulation. For the rigid water model TIP3P,[47] bond lengths and angles are fixed and there are no bonded interactions. The nonbonded interaction energy is defined as the sum of respective Lennard-Jones and Coulomb interactions. Decomposition of electrostatic interactions in the reciprocal space when using the PME approach[48,49] is given in the Supporting Information. Error estimates of enthalpy were calculated using a binning analysis.[53]

The configurational entropies of host and guest molecules were computed from the covariance matrices of their atomic fluctuations using the quasiharmonic approximation.[54] We first calculated entropy changes of host and guest with respect to the unbound state separately and then subtracted them from ΔS to obtain the solvent entropy change involved with, for instance, solvent rearrangements during desolvation of host and guest molecules upon binding (eq 4).

$$\Delta S_{\text{sol}} = \Delta S - \Delta S_{\text{host}} - \Delta S_{\text{guest}} \tag{4}$$

Since the error in ΔH would propagate to ΔS, all entropy terms here were assumed to have the same errors as ΔH.

RESULTS

Complex Arrangement. Potential of mean force (PMF) profiles for the formation process of 1:1 [β-CD:puerarin] complexes with BP, BS, and BO arrangements and representative states (A...G) in the reaction coordinate ξ are presented in Figure 3. The three structural arrangements refer to Figure 1c. BP and BS in our simulations indicate that the B-ring of guest inserts into CD cavity along the +ξ and −ξ direction, respectively.

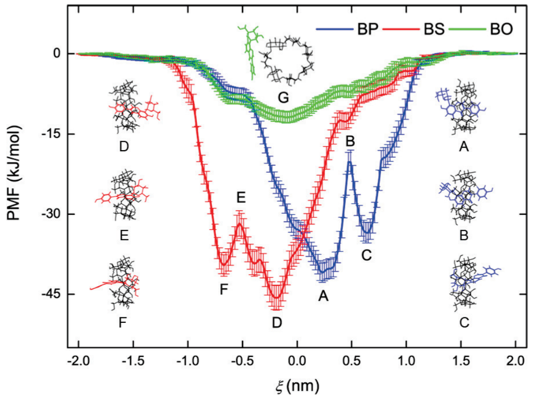


Figure 3. Potential of mean force (PMF) profiles for the [β-CD:puerarin] complex formation in three structural arrangements (BP, BS, and BO). Representative configurations along ξ are shown using line model. β-CD is colored in black and puerarin in the same color as the arrangement.

As shown in Figure 3, periodic PMFs ensure equality of the guest located at ξ = −2 and 2 nm. All the PMFs approach to zero and level off on both sides of the reaction coordinate where there is no interaction between β-CD and puerarin. The A- and D-states with B- and C-rings of puerarin inside the β-CD cavity give the most stable inclusion configuration for BP and BS, respectively. When the A-ring of puerarin approaches the cavity, such as in the B- and E-states, an energy barrier is observed and this barrier might prevent puerarin from further penetrating into the CD cavity. The C- and F-states with the glucose unit of puerarin inside the cavity form local minima in the PMFs. As expected, there is no obvious barrier and a weaker binding is observed for the BO arrangement, as in the G-state, due to a less efficient contact of hydrophobic moieties between host and guest, compared to the inclusion complexes such as BP and BS. For the G-state, puerarin binds to the outer surface of β-CD with its isoflavone skeleton (i.e., the A, B, and C rings in Figure 1b) perpendicular to the glucopyranose residue of β-CD; this way the hydrophobic contact area appears to be maximized. The D-state is more energetically favorable than the A-state and therefore is the most probable configuration, in good agreement with the experiment where a very similar [β-CD:puerarin] inclusion complex (such as the D-state) was detected in aqueous solution by NMR spectroscopy.[55]

PMF profiles (ΔG) for puerarin and daidzin are presented in Figure 4. Daidzin behaves similar as puerarin, whereas it can insert into β-CD cavity more deeply than puerarin in the BS arrangement (Figure 4, panels b and e). PMF profiles (ΔG) for daidzein and nabumetone are given in Figure S2 in the Supporting Information. For daidzein and nabumetone no pronounced energy barriers are observed, and both BP and BS are thermodynamically stable although BS is preferred slightly over BP. NMR experiments have identified these two possible [β-CD:nabumetone] inclusion complexes.[56] When hydrophobic moieties of the guest (such as daidzin, daidzein, and nabumetone) stay inside β-CD cavity, the PMF profiles display a flat landscape (Figure 4 and Supporting Information Figure S2), implying that there is almost no energy barrier and the guest can shuttle freely inside the cavity to some extent. A shuttling motion of puerarin and daidzin inside β-CD cavity in the BS pattern has indeed been detected by MD simulations.[57]

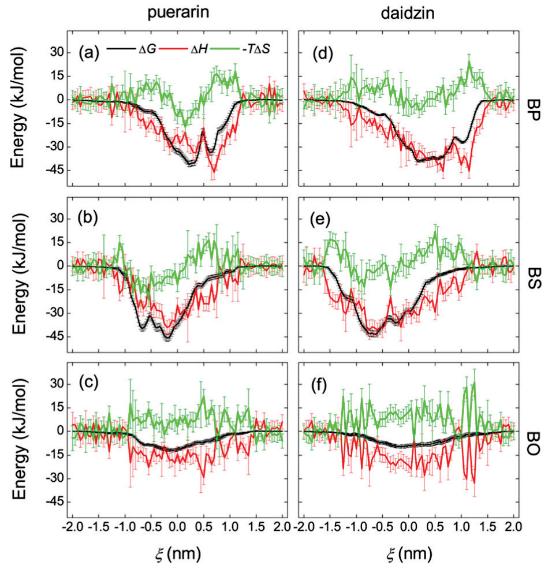


Figure 4. Thermodynamic profiles (ΔG, ΔH, and $-T\Delta S$) of the system for the complex formation of β-CD with puerarin and daidzin in three patterns BP, BS, and BO.

System Thermodynamics. Thermodynamic profiles (ΔG, ΔH, and ΔS) of the system along ξ for the four guests with BP, BS, and BO arrangements are shown in Figure 4 and Supporting Information Figure S2. Here entropy is presented as $-T\Delta S$. From these profiles we can derive contributions of enthalpy and entropy to ΔG. A reduced enthalpy (more favorable) is observed for all the guests upon complexation, while entropy increases in some cases and decreases in other. The thermodynamic stability of these complexes can be therefore attributed to a combination of both ΔH and ΔS. As shown in Figure 4a, for instance, both enthalpy and entropy gains favor a stable complex (i.e., the A-state in Figure 3), which corresponds to the global minimum of the PMF. When puerarin enters the β-CD cavity more deeply with its glucose unit inside the cavity (such as the C-state in Figure 3), ΔH reaches a maximum, whereas an entropy loss cancels out this enthalpy gain, giving a moderate ΔG (Figure 4a). Unlike the C-state, the D-state in Figure 3 is a maximum of enthalpy gain and has an entropy gain, forming a global minimum of ΔG (Figure 4b). The other three guest molecules display similar enthalpy–entropy relationships to puerarin for BP and BS (Figure 4 and Supporting Information Figure S2). For BO, enthalpy gain and entropy loss are detected for puerarin and daidzin (Figure 4, panels c and f), whereas for daidzein and nabumetone the complex stability seems to result exclusively from the enthalpy (Supporting Information Figure S2, panels c and f). Interestingly, puerarin, daidzin, and daidzein share the same isoflavone skeleton and have similar enthalpy gains upon binding to the outer surface of β-CD, but an entropy loss decreases the binding of puerarin and daidzin. This entropy loss may be due to the limited movement of the glucose unit when interacting with the β-CD surface. Daidzein does not have such glucose group (Figure 1b), and there is no significant change in entropy, leading to a relatively stronger binding (Supporting Information Figure S2c).

Now, we turn to the standard thermodynamics of the entire binding reactions for [β-CD:guest] associations. A cylinder approximation[22,58−60] was used to evaluate the standard binding free energies. When a guest enters the β-CD cavity, the sampled volume for the guest is restrained to a small cylinder defined by the area accessible for guest movement in the X−Y plane. The average radius of that cylinder, $r(\xi)$, was obtained from COM positions of the guest at each window. The association equilibrium constant K_a is written as

$$K_a = \pi N_A \int r(\xi)^2 \exp[-\Delta G(\xi)/RT]\mathrm{d}\xi \tag{5}$$

where N_A is Avogadro constant and R the ideal gas constant.[58,59] The thermodynamics of binding can therefore be calculated using

$$\Delta G^0 = -RT\ln(K_a C^0) \tag{6}$$

$$\begin{aligned}\Delta H^0 &= RT^2 \frac{d}{\mathrm{d}T}\ln(K_a C^0)\\ &= \frac{\int r(\xi)^2 \Delta G(\xi) \exp[-\Delta G(\xi)/RT]\mathrm{d}\xi}{\int r(\xi)^2 \exp[-\Delta G(\xi)/RT]\mathrm{d}\xi}\end{aligned} \tag{7}$$

$$-T\Delta S^\circ = \Delta G^\circ - \Delta H^\circ \tag{8}$$

where C° is the standard concentration of 1 mol/L.[61] Note that ΔG° here is the standard free energy of the binding process, while $\Delta G(\xi)$ denotes free energy profiles obtained from PMF calculations. The integration is limited to the interval over which host and guest molecules associate. As noted by Bonal and co-workers,[60] the integration was computed from each side of the PMF profile (where host and guest have no interaction) to the central maximum and they averaged over these two reaction pathways to obtain the thermodynamic parameters. A similar treatment is adopted in our calculation to define the integration interval in eqs 6 and 7. For the cases where there is no obvious central maximum in the PMF, such as daidzein (Supporting Information Figure S2a) and nabumetone (Figure S2e), we perform the integral over the whole PMF.

Table 1 lists the calculated ΔG° for the four drugs studied. For daidzein and nabumetone, ΔG° compares well with the experiment, while the calculation overestimates the binding strength between β-CD and puerarin. The results depend on

Table 1. Comparison of Calculated Binding Free Energy (kJ/mol) with Experimental Determinations

guest	T (K)	$-\Delta G_{exp}$	$-\Delta G^0_{cal}$ BP	$-\Delta G^0_{cal}$ BS
puerarin	300	19.0[a]	26	32
daidzin	300		24	29
daidzein	300	16.6[b]	19	22
nabumetone	293	19.2[c] /19.7[d] /18.7[e]	18	21

[a] Taken from ref 55. [b] Ref 62. [c] Ref 56. [d] Ref 63. [e] Ref 64.

the interval used for integration for sure; a shorter interval gives a weaker binding. If β-CD in the simulation is more rigid that in the experiment, there would exist energy barriers preventing the guest from further accessing some part of the binding site. That is, a more rigid host would lead to a shorter integration interval. If so, we can get much closer to the experiment by adjusting the host flexibility artificially. Another factor responsible for the source of error could probably be the force field used. Data for $\Delta H°$ and $\Delta S°$ are given in Tables S2 and S3 in the Supporting Information. For nabumetone, there is some discrepancy between calculated and observed $\Delta H°$ and $\Delta S°$ (in exp. 2 and 3, but not 1, Table S3).

Enthalpy Decomposition. For a better understanding of the distinct shape of an enthalpic profile, we decomposed it into eight terms including bonded and nonbonded interactions (eq 3). Figure 5 shows the ΔH decomposition for the [β-CD:puerarin] complex formation. For BP and BS, changes in ΔH_{host} and ΔH_{guest} upon binding are positive, which means that the bonded term of binding partners tends to disfavor host–guest inclusion complexations, indicated by black and red lines (Figure 5, panels a and b). For BO (Figure 5c), no significant changes are observed for ΔH_{host} and ΔH_{guest}. $\Delta H_{host-host}$ and $\Delta H_{guest-guest}$ (green and blue lines, panels a and b in Figure 5) tend to favor host–guest complexations (negative values). There are significant enthalpy changes in intramolecular interactions of the host ($\Delta H_{host-host}$) for BP and BS; no obvious changes for BO.

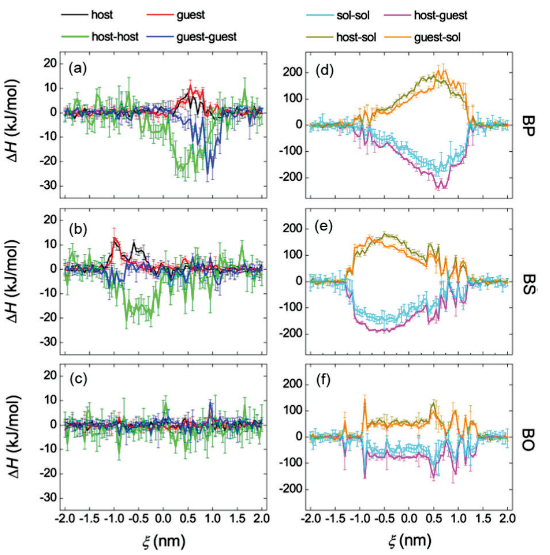


Figure 5. Enthalpy decomposition for the complex formation of β-CD with puerarin in three patterns BP, BS, and BO.

When a guest travels from the bulk into the CD cavity, the solvent molecules entrapped inside the cavity will be expelled, a process such as the release of "high-energy" water. Another contribution to the energetics is due to release of water molecules that participate in host and guest solvation. As a result, the water–water enthalpy $\Delta H_{sol-sol}$ becomes more negative (the cyan line in Figure 5, panels d–f). Unsurprisingly, the strength of the intermolecular interaction between host and guest ($\Delta H_{host-guest}$) increases when forming a complex, as indicated by the magenta line. Accompanied by desolvation, the strength of the interaction between host (or guest) and solvents decreases (positive ΔH, dark yellow and orange lines in Figure 5). When accommodating puerarin as a guest, β-CD reaches a desolvation maximum ($\Delta H_{host-sol}$, panels d and e in Figure 5) where the A- and C-rings of the guest are inserted into the cavity and the glucose unit stays very close to the cavity, such as in the B- and E-states in Figure 3. When the glucose unit of puerarin goes further and stays inside the cavity (C- and F-states in Figure 3), host desolvation gets weakened and guest desolvation maximized (Figure 5, panels d and e). The guest bound to β-CD outer surface also affects (de)solvation of host and guest molecules, but to a lesser degree (Figure 5, panels d–f). Similar observations are detected as well for β-CD complexes with daidzin, daidzein, and nabumetone, as shown in Figures S3–S5, respectively, in the Supporting Information. Since daidzein and nabumetone do not possess a glucose unit, they give more symmetrical profiles of the ΔH decomposition (Figures S4 and S5).

Entropy Decomposition. In order to distinguish individual entropy contributions clearly, the total entropy was decomposed into three single terms corresponding to host, guest, and solvent molecules (eq 4) and presented as $-T\Delta S$. Figure 6 shows the entropy decomposition for β-CD complexes with puerarin and daidzin; data for daidzein and nabumetone are given in Figure S6 in the Supporting Information. An

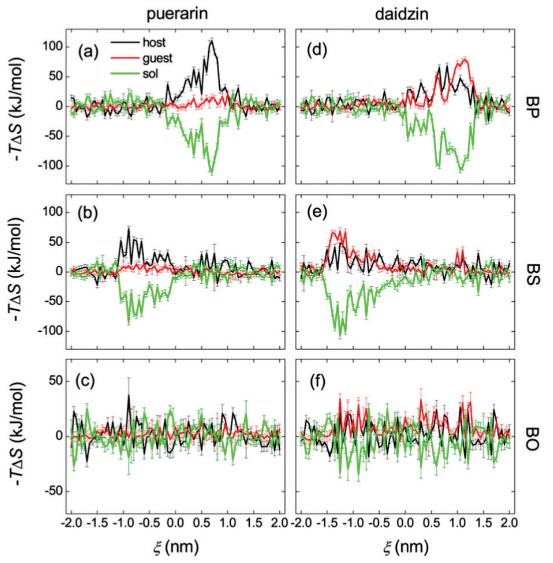


Figure 6. Entropy decomposition for the complex formation of β-CD with puerarin and daidzin in three patterns BP, BS, and BO.

Table 2. Individual Contribution (kJ/mol) of ΔH and ΔS Weighted by Boltzmann Factors for the Actual Binding Reactions between β-CD and Guest Molecules with BP and BS Arrangements (Standard Deviations in Parentheses)

	puerarin		daidzin		daidzein		nabumetone	
$\langle \Delta E \rangle$	BP	BS	BP	BS	BP	BS	BP	BS
ΔH_{host}	2(1)	1(1)	1(1)	−1(1)	1(1)	1(1)	−1(1)	0(1)
ΔH_{guest}	3(1)	1(1)	0(1)	−1(1)	0(1)	0(1)	0(1)	0(1)
$\Delta H_{host-host}$	−17(3)	−13(3)	−9(2)	−8(2)	−9(2)	−10(3)	−12(3)	−10(2)
$\Delta H_{guest-guest}$	−1(1)	−1(1)	−4(2)	0(1)	0(1)	1(1)	−2(1)	−3(1)
$\Delta H_{sol-sol}$	−136(8)	−130(7)	−123(7)	−123(6)	−98(7)	−106(6)	−96(7)	−111(8)
$\Delta H_{host-guest}$	−176(8)	−170(7)	−163(4)	−164(5)	−129(9)	−128(7)	−123(8)	−126(7)
$\Delta H_{host-sol}$	168(9)	155(8)	142(6)	145(6)	119(7)	119(8)	131(8)	132(8)
$\Delta H_{guest-sol}$	127(7)	123(7)	119(5)	117(5)	86(6)	85(6)	80(6)	85(6)
$-T\Delta S_{host}$	42(4)	24(3)	24(3)	14(3)	14(3)	9(2)	21(3)	18(3)
$-T\Delta S_{guest}$	4(2)	7(2)	13(3)	23(4)	0(1)	0(1)	11(2)	14(2)
$-T\Delta S_{sol}$	−55(4)	−38(4)	−38(3)	−41(4)	−12(3)	−14(3)	−33(5)	−33(4)

obvious entropy loss of the host and a slight loss of the guest are observed for puerarin with BP and BS arrangements (Figure 6, panels a and b). For daidzin there are pronounced entropy losses for both the host and guest (Figure 6, panels d and e), due to loss of flexibility in host and guest molecules upon complexation. For puerarin, daidzin, and nabumetone in BP and BS patterns, the solvent in contrast tends to gain entropy (positive ΔS), favoring the complexation. No obvious changes in ΔS are detected for the [β-CD:daidzein] inclusion (Figure S6, panels a and b). Binding of a guest to the outer surface of β-CD may also result in an entropy change to a certain extent (Figure 6 and Supporting Information Figure S6).

Figure S7 in the Supporting Information presents thermodynamic profiles for the BS [β-CD:nabumetone] inclusion with a flexible or rigid host. Compared to the flexible host, the rigid one gives more minima in the PMF and has a weaker binding to nabumetone due to a less favorable enthalpy gain (Supporting Information, Figure S7, panels a and b), as indicated by the ΔH decomposition (Supporting Information, Figure S7, panels c−f). As expected, the rigid host displays little entropy loss upon binding to the guest since all the non-hydrogen atoms are harmonically fixed. When entrapped inside a rigid cavity, nabumetone shows larger entropy loss and solvent molecules give a larger entropy gain (Supporting Information, Figure S7, panels g and h).

For a quantitative determination of the energetics, individual contributions of $\Delta H(\xi)$ and $-T\Delta S(\xi)$ are weighted by their Boltzmann factors (eq 9)

$$\langle \Delta E \rangle = \frac{\int \Delta E(\xi) \exp[-\Delta G(\xi)/RT] \mathrm{d}\xi}{\int \exp[-\Delta G(\xi)/RT] \mathrm{d}\xi} \tag{9}$$

where ΔE represents ΔH or $-T\Delta S$. Weighted values for the actual binding reactions are tabulated in Table 2, showing similar observations to what was mentioned above.

Guest Rotation. The configurational entropy here was determined from covariance matrices of atomic fluctuations.[54] A guest entrapped inside the CD cavity probably cannot rotate as freely as it is in the bulk, which may limit structural fluctuations of the guest and hence cause an entropy loss. To detect guest rotations in the free and complex state, four dihedral angles were defined in Figure 1b. Dihedral potentials of the four angles taken from the GAFF parameters[42] are given in Figure S8 in the Supporting Information. A large energy barrier exists for ψ_1, meaning that it is not easy for ψ_1 to rotate. There are smaller barriers for ψ_3 and ψ_4; no barrier for ψ_2. It should be noted that Supporting Information Figure S8 shows the intrinsic barrier only and dihedral rotations also depend on the environment of the molecule.

Distributions of these dihedrals (ψ_i, i = 1...4) during the formation process of β-CD with puerarin and daidzin in the BP pattern are presented in Figure 7. Free states for puerarin and daidzin locate at $\xi = -2.0$ nm. The B-ring of guest inserted into β-CD cavity at $\xi = 0.5$ nm; the glucose unit of guest stays inside the cavity at $\xi = 1.0$ nm. For puerarin and daidzin in the free and complex state, there is no significant difference for ψ_1, and the same goes for daidzein (not shown here). The glucose rotation (ψ_2) for puerarin is almost not affected when entrapped inside the cavity, and it is similar to the free state, which may explain the small entropy change of guest in Figure 6a. Hydrogen-bonding interactions between the hydroxyl group connected to A-ring and the glucose unit of puerarin are

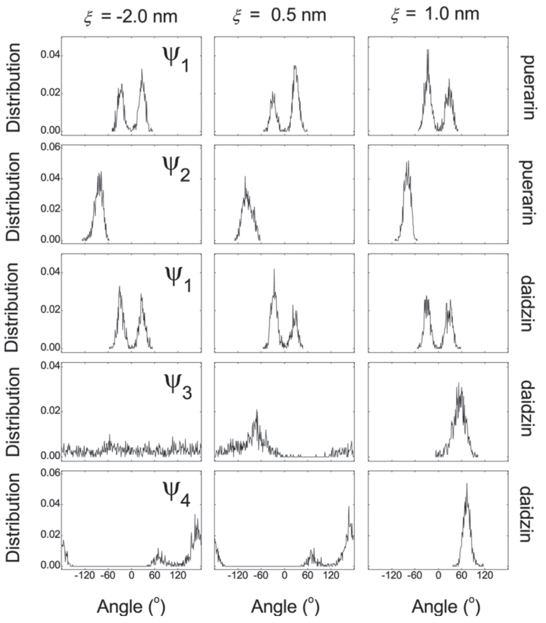


Figure 7. Distribution of dihedral angles for puerarin (ψ_1 and ψ_2) and daidzin (ψ_1, ψ_3, and ψ_4) with the BP arrangement along ξ. Dihedral distribution for daidzein (ψ_1) is similar to that for puerarin and daidzin.

observed in the simulation, which may limit the rotation of ψ_2. For daidzin, the glucose unit rotates freely in the free state (ξ = −2.0 nm), as indicated by ψ_3 and ψ_4, whereas their rotations are evidently limited when the glucose unit stays inside the cavity at ξ = 1.0 nm (Figure 7). This finding explains the large entropy loss of daidzin upon binding to β-CD in Figure 6d.

DISCUSSION

When typical cyclodextrins (α-, β-, and γ-CD containing six, seven, and eight glucopyranose residues in a ring, respectively) form 1:1 complexes with asymmetrical guest molecules, three possible arrangements (BP, BS, and BO in Figure 1c) may be adopted and they are expected to be different in energy due to the guest orientation. Much attention has been paid to BP and BS inclusion patterns both in academic research and industrial applications, since the CD cavity is a specific binding site and the outer surface is not. The asymmetric free energy profiles for the BP and BS arrangements in our simulations (Figures 3 and 4) indicate the difference in the specific binding and in two types of inclusion complexes. Many compounds have been reported to show two possible inclusion models with CDs, such as surfactants with typical CDs[18,65] and steroid drugs,[22] flavanols,[66,67] and aziadamantane derivatives[68] with β-CD.

Host−guest complexes are often used as models to gain general insights on the thermodynamics of binding, due to their small size and simplicity compared to protein−ligand systems. Many projects have been devoted to studying the thermodynamics of cyclodextrin complexation using PMF calculations.[39,60,69−72] Cai and co-workers reported a decomposition of the PMF profile into van der Waals (host−guest), electrostatic (host−guest), and host−solvent interactions.[69−71] In addition, Kovalenko et al. proposed a spatial decomposition analysis for the cyclodextrin complexation and decomposed the thermodynamics into the excluded volume and solvation shell terms.[73] In a very recent study, Wickstrom et al. indicated that the binding free energy can be decomposed into the reorganization free energy and the average binding energy.[74] In this work, we introduce another decomposition to characterize the total and individual contributions (ΔH and ΔS) from the binding partners as well as their solvation environment, as described in detail in the Methods section.

Thermodynamic profiles of the system (Figure 4 and Supporting Information Figure S2) show that both enthalpy and entropy contribute to the binding between the model host β-CD and the guests studied. Such binding reactions are predominantly enthalpy-driven and in some cases an entropy loss weakens the binding. Decomposition of the total enthalpy (ΔH) provides more information on individual contributions from intra- and intermolecular interactions, as shown in Figure 5 and Supporting Information Figures S3−S5. As expected, desolvation of host and guest molecules gives an unfavorable ΔH and the complex formation produces a favorable ΔH. The solvent favors the complex stability as well with enthalpy gains. Surprisingly, the numerical values of these four terms ($\Delta H_{\text{host−sol}}$, $\Delta H_{\text{guest−sol}}$, $\Delta H_{\text{host−guest}}$, and $\Delta H_{\text{sol−sol}}$) are an order of magnitude larger than that for the thermodynamic parameters of the entire binding reactions, as shown in Tables 1 and 2 and Supporting Information Tables S2−S3, which implies that these contributions to the binding need to be considered with care. Moreover, changes in intramolecular energies of the binding partners (ΔH_{host}, ΔH_{guest}, $\Delta H_{\text{host−host}}$, and $\Delta H_{\text{guest−guest}}$), in particular for nonbonded interactions of the host ($\Delta H_{\text{host−host}}$), indicate that host molecules adjust their configurations to the binding environment, and so do guest molecules. This adjustment (configurational fit) reflects fluctuations in atomic positions, known as guest-induced effects,[14] leading to changes in the potential energy and thus to ΔH between 8 and 17 kJ/mol. Dolenc et al. investigated the effect of receptor flexibility on the binding affinity and reported that neglecting the receptor flexibility affected the model structures of the complex and enthalpy contributions to the binding, in particular for a flexible receptor such as DNA.[23] Since the contributions from conformational changes are on the same order of magnitude as the standard thermodynamics of binding (Tables 1 and 2 and Supporting Infomation Tables S2−S3), these need to be considered explicitly when computing binding energies.

For the entropy (ΔS) decomposition in Figure 6, most of entropy changes take place when flexible moieties of the guest are included inside the CD cavity or interact with the CD surface. Both host and guest may lose entropy upon binding, depending on the guest and orientation. The solvent, however, tends to have an entropy gain, favoring the complex formation. Desolvation of the binding partners liberates solvent molecules participating in the solvation, allowing a greater degree of freedom for motion of these water molecules and hence increased entropy, in line with common perception of the hydrophobic effect.[75,76] Daidzein, the most hydrophobic and rigid molecule among the tested guests, has weaker interactions with water molecules and undergoes smaller fluctuations in structure. Inclusion of daidzein to the CD cavity should perturb the binding-site waters and displace them from the cavity. For the solvent, this process ought to give favorable ΔS. However, no significant ΔS for β-CD, daidzein, and water molecules is observed upon complexation in the simulation (Supporting Information Figure S6). This finding could be ascribed to the fact that rigidity of daidzein leads to a weaker (de)solvation and does not affect the surrounding environment too much.

As ideal host−guest models, the [CD:drug] complexes studied in this work hold valuable implications for the receptor−ligand binding. It has been realized for a long time that for truly predictive estimates of ligand−binding energies free energy methods are crucial.[77] However the (high-throughput) virtual screening concept has remained popular, despite suggestions that it may not live up to the hype.[78] Docking and binding-site predictions can yield good candidates for binding sites,[79] but the built-in scoring functions are not necessarily predictive of binding strength,[80] and docking codes are therefore regarded with some skepticism.[81] For high-throughput virtual screening to work, an accurate estimation of the contribution from solvation and from conformational changes would be needed, without the computational cost associated with free energy calculations. Based on our calculations, changes in intramolecular interactions due to configurational fit contribute significantly to the complexation thermodynamics (Figure 5 and Supporting Information Figures S3−S5). Moreover, we observe a large enthalpy gain of the solvent environment for both flexible and rigid hosts, and this contribution most likely cannot be evaluated accurately when neglecting the solvent or using implicit solvent.

Entropy estimation, especially for the solvent environment, is another difficulty faced by high-throughput virtual screening. A detailed review for theory of free energy and entropy in noncovalent binding has been presented by Zhou and Gilson.[82] In this work, we used a quasiharmonic approximation[54] to calculate the configurational entropy; the Schlitter formula was also tested.[83] We note that the quasiharmonic method is an approximation, which does not approach the true entropy even in the limit of infinite sampling, but these two methods were reported to be useful for ΔS estimation of ligands in binding processes.[84] For our cases, the two methods yield very similar results for the relative ΔS (which is very important in this analysis presented), although the absolute values differ (not shown here). The entropy of the host and guest can be computed readily, and this has been incorporated in methods for estimating the binding energy of complexes.[6,7,85] In our calculations, the solvent entropy is computed indirectly using eq 4, and its accuracy depends on the estimation of ΔG and ΔH and the values of $-T\Delta S$ depend heavily on the guest, varying between −12 to −55 kJ/mol (Table 2). The results show that the solvent tends to gain entropy and cancel out most of the entropy losses of the binding partners. Neglecting either of the flexibility or the entropy items would yield an error with the same order of magnitude as the entire binding energy. It is therefore difficult to imagine that the accuracy of scoring functions for use in virtual screening (e.g., pharmaceutical design and biotechnology projects) can be increased sufficiently to systematically reach an accuracy comparable to free energy calculations.[77]

■ CONCLUSION

In this work, all possible complex arrangements between a model host (β-CD) and four drug guests (puerarin, daidzin, daidzein, and nabumetone) were evaluated through steered molecular dynamics and potential of mean force calculations. The total and individual contribution of enthalpy and entropy to the stability of such noncovalent complexes were analyzed in terms of binding mode, solvation, and structural flexibility. Our results show that host flexibility, solvent enthalpy, and solvent entropy play important roles in host−guest complexation, and these items need to be included explicitly for accurate calculation of the binding thermodynamics. We have previously demonstrated that the binding energy of [host:guest] complexes in different organic solvents is only weakly correlated to solvent properties such as the dielectric constants or Log P.[39] An implicit solvent model can provide a useful estimate of solvation free energy only if used under the conditions it was parametrized for (temperature, solvent) and if there are no very specific hydrogen bonds. Implicit models are not suitable to provide detailed information on how that free energy is partitioned into enthalpy and entropy. Full molecular dynamics (MD) simulations using explicit solvents are therefore required for precise estimation of thermodynamic parameters of molecular complexation.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Decomposition of reciprocal Ewald sum (Table S1), calculated $\Delta H°$ and $\Delta S°$ (Tables S2 and S3), COM pulling for puerarin (Figure S1), thermodynamic profiles of the system for daidzein and nabumetone (Figure S2), ΔH decomposition for daidzin, daidzein, and nabumetone (Figure S3−S5), ΔS decomposition for daidzein and nabumetone (Figure S6), thermodynamic profiles for nabumetone with a flexible or rigid host (Figure S7), and dihedral potential (Figure S8). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: spoel@xray.bmc.uu.se.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Daniel Larsson and Dr. Jochen Hub for assistance. The Swedish research council is acknowledged for a grant of computer time (SNIC020-11-15) through HPC2N. H.Z. thanks the China scholarship council for support. This work was supported by the National Basic Research Program of China (973 program: 2013CB733600, 2012CB725200, and 2011CB710800), the National Nature Science Foundation of China (21106005, 21076017, and 21306006), and National High-Tech R&D Program of China (863 program: 2012AA022205D, 2012AA021404, and 2011AA02A205).

■ REFERENCES

(1) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.

(2) Straatsma, T. P.; Berendsen, H. J. C. Free Energy of Ionic Hydration: Analysis of a Thermodynamic Integration Technique to Evaluate Free Energy Differences by Molecular Dynamics Simulations. *J. Chem. Phys.* **1988**, *89*, 5876−5886.

(3) Pearlman, D. A.; Kollman, P. A. A New Method for Carrying out Free Energy Perturbation Calculations: Dynamically Modified Windows. *J. Chem. Phys.* **1989**, *90*, 2460−2470.

(4) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.

(5) Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simulat.* **1993**, *10*, 97−120.

(6) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. Dock 6: Combining Techniques to Model RNA−Small Molecule Complexes. *RNA* **2009**, *15*, 1219−1230.

(7) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodock4 and Autodocktools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(8) Gohlke, H., Ed. Protein−Ligand Interactions. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R.; Kubinyi, H.; Folkers, G., Eds. Wiley-VCH: Weinheim, 2012; Vol. *53*.

(9) van de Manakker, F.; Vermonden, T.; van Nostrum, C. F.; Hennink, W. E. Cyclodextrin-Based Polymeric Materials: Synthesis, Properties, and Pharmaceutical/Biomedical Applications. *Biomacromolecules* **2009**, *10*, 3157−3175.

(10) Fan, Z.; Diao, C.-H.; Song, H.-B.; Jing, Z.-L.; Yu, M.; Chen, X.; Guo, M.-J. Encapsulation of Quinine by β-Cyclodextrin: Excellent Model for Mimicking Enzyme−Substrate Interactions. *J. Org. Chem.* **2006**, *71*, 1244−1246.

(11) Shin, K.-m.; Dong, T.; He, Y.; Taguchi, Y.; Oishi, A.; Nishida, H.; Inoue, Y. Inclusion Complex Formation between α-Cyclodextrin and Biodegradable Aliphatic Polyesters. *Macromol. Biosci.* **2004**, *4*, 1075−1083.

(12) Ellis, A. V.; Chong, S.; Jansen, M. Formation of an α-Cyclodextrin/16-Mercaptohexadecanoic Acid Complex and Its Deposition on Gold Surfaces. *J. Inclusion Phenom. Macrocyclic Chem.* **2009**, *63*, 267−272.

(13) Raffaini, G.; Ganazzoli, F.; Malpezzi, L.; Fuganti, C.; Fronza, G.; Panzeri, W.; Mele, A. Validating a Strategy for Molecular Dynamics Simulations of Cyclodextrin Inclusion Complexes through Single-

Crystal X-Ray and NMR Experimental Data: A Case Study. *J. Phys. Chem. B* **2009**, *113*, 9110−9122.

(14) Saenger, W.; Noltemeyer, M.; Manor, P. C.; Hingerty, B.; Klar, B. "Induced-Fit"-Type Complex Formation of the Model Enzyme α-Cyclodextrin. *Bioorg. Chem.* **1976**, *5*, 187−195.

(15) Dodziuk, H. Rigidity Versus Flexibility. A Review of Experimental and Theoretical Studies Pertaining to the Cyclodextrin Nonrigidity. *J. Mol. Struct.* **2002**, *614*, 33−45.

(16) Zhang, H.; Ge, C.; van der Spoel, D.; Feng, W.; Tan, T. Insight into the Structural Deformations of β-Cyclodextrin Caused by Alcohol Cosolvents and Guest Molecules. *J. Phys. Chem. B* **2012**, *116*, 3880−3889.

(17) Zhang, H.; Feng, W.; Li, C.; Tan, T. Investigation of the Inclusions of Puerarin and Daidzin with β-Cyclodextrin by Molecular Dynamics Simulation. *J. Phys. Chem. B* **2010**, *114*, 4876−4883.

(18) Zheng, X.; Wang, D.; Shuai, Z.; Zhang, X. Molecular Dynamics Simulations of the Supramolecular Assembly between an Azobenzene-Containing Surfactant and α-Cyclodextrin: Role of Photoisomerization. *J. Phys. Chem. B* **2011**, *116*, 823−832.

(19) Pan, W.; Zhang, D.; Zhan, J. Theoretical Investigation on the Inclusion of Tcdd with β-Cyclodextrin by Performing Qm Calculations and Md Simulations. *J. Hazard. Mater.* **2011**, *192*, 1780−1786.

(20) Inoue, Y.; Hakushi, T.; Liu, Y.; Tong, L.; Shen, B.; Jin, D. Thermodynamics of Molecular Recognition by Cyclodextrins. 1. Calorimetric Titration of Inclusion Complexation of Naphthalenesulfonates with α-, β-, and γ-Cyclodextrins: Enthalpy−Entropy Compensation. *J. Am. Chem. Soc.* **1993**, *115*, 475−481.

(21) Rekharsky, M.; Inoue, Y. Chiral Recognition Thermodynamics of β-Cyclodextrin:The Thermodynamic Origin of Enantioselectivity and the Enthalpy−Entropy Compensation Effect. *J. Am. Chem. Soc.* **2000**, *122*, 4418−4435.

(22) Cai, W.; Sun, T.; Liu, P.; Chipot, C.; Shao, X. Inclusion Mechanism of Steroid Drugs into β-Cyclodextrins. Insights from Free Energy Calculations. *J. Phys. Chem. B* **2009**, *113*, 7836−7843.

(23) Dolenc, J.; Riniker, S.; Gaspari, R.; Daura, X.; van Gunsteren, W. Free Energy Calculations Offer Insights into the Influence of Receptor Flexibility on Ligand−Receptor Binding Affinities. *J. Comput. Aided Mol. Des.* **2011**, *25*, 709−716.

(24) Zhang, Q.; Tu, Y.; Tian, H.; Zhao, Y.-L.; Stoddart, J. F.; Ågren, H. Working Mechanism for a Redox Switchable Molecular Machine Based on Cyclodextrin: A Free Energy Profile Approach. *J. Phys. Chem. B* **2010**, *114*, 6561−6566.

(25) Sun, T.; Shao, X.; Cai, W. Self-Assembly Behavior of β-Cyclodextrin and Imipramine. A Free Energy Perturbation Study. *Chem. Phys.* **2010**, *371*, 84−90.

(26) Liu, P.; Cai, W.; Chipot, C.; Shao, X. Thermodynamic Insights into the Dynamic Switching of a Cyclodextrin in a Bistable Molecular Shuttle. *J. Phys. Chem. Lett.* **2010**, *1*, 1776−1780.

(27) El-Barghouthi, M. I.; Jaime, C.; Akielah, R. E.; Al-Sakhen, N. A.; Masoud, N. A.; Issa, A. A.; Badwan, A. A.; Zughul, M. B. Free Energy Perturbation and MM/PBSA Studies on Inclusion Complexes of Some Structurally Related Compounds with β-Cyclodextrin. *Supramol. Chem.* **2009**, *21*, 603−610.

(28) Keung, W. M.; Vallee, B. L. Kudzu Root: An Ancient Chinese Source of Modern Antidipsotropic Agents. *Phytochemistry* **1998**, *47*, 499−506.

(29) Lowe, E. D.; Gao, G.-Y.; Johnson, L. N.; Keung, W. M. Structure of Daidzin, a Naturally Occurring Anti-Alcohol-Addiction Agent, in Complex with Human Mitochondrial Aldehyde Dehydrogenase. *J. Med. Chem.* **2008**, *51*, 4482−4487.

(30) Moorwood, C.; Lozynska, O.; Suri, N.; Napper, A. D.; Diamond, S. L.; Khurana, T. S. Drug Discovery for Duchenne Muscular Dystrophy Via Utrophin Promoter Activation Screening. *Plos One* **2011**, *6*, e26169.

(31) Lemkul, J. A.; Bevan, D. R. Assessing the Stability of Alzheimer's Amyloid Protofibrils Using Molecular Dynamics. *J. Phys. Chem. B* **2010**, *114*, 1652−1660.

(32) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(33) Rashid, M. H.; Kuyucak, S. Affinity and Selectivity of Shk Toxin for the Kv1 Potassium Channels from Free Energy Simulations. *J. Phys. Chem. B* **2012**, *116*, 4812−4822.

(34) Hub, J. S.; Winkler, F. K.; Merrick, M.; de Groot, B. L. Potentials of Mean Force and Permeabilities for Carbon Dioxide, Ammonia, and Water Flux across a Rhesus Protein Channel and Lipid Membranes. *J. Am. Chem. Soc.* **2010**, *132*, 13251−13263.

(35) Wennberg, C. L.; van der Spoel, D.; Hub, J. S. Large Influence of Cholesterol on Solute Partitioning into Lipid Membranes. *J. Am. Chem. Soc.* **2012**, *134*, 5351−5361.

(36) Caleman, C.; Hub, J. S.; van Maaren, P. J.; van der Spoel, D. Atomistic Simulation of Ion Solvation in Water Explains Surface Preference of Halides. *Proc. Natl. Acad. Sci.* **2011**, *108*, 6838−6842.

(37) Hub, J. S.; Caleman, C.; van der Spoel, D. Organic Molecules on the Surface of Water Droplets—An Energetic Perspective. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9537−9545.

(38) Cezard, C.; Trivelli, X.; Aubry, F.; Djedaini-Pilard, F.; Dupradeau, F. Y. Molecular Dynamics Studies of Native and Substituted Cyclodextrins in Different Media: 1. Charge Derivation and Force Field Performances. *Phys. Chem. Chem. Phys.* **2011**, *13*, 15103−15121.

(39) Zhang, H.; Tan, T.; Feng, W.; van der Spoel, D. Molecular Recognition in Different Environments: β-Cyclodextrin Dimer Formation in Organic Solvents. *J. Phys. Chem. B* **2012**, *116*, 12684−12693.

(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(41) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(42) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(43) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(44) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G., Scuseria, E., Robb, M. A., Cheeseman, J. R., Montgomery, J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(45) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. Lincs: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(46) Miyamoto, S.; Kollman, P. A. Settle—An Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952−962.

(47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(48) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(49) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald—An N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(50) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(51) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(52) Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713−3720.

(53) Hess, B. Determining the Shear Viscosity of Model Liquids from Molecular Dynamics Simulations. *J. Chem. Phys.* **2002**, *116*, 209−217.

(54) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289−6292.

(55) Zhao, R.; Tan, T.; Sandström, C. NMR Studies on Puerarin and Its Interaction with β-Cyclodextrin. *J. Biol. Phys.* **2011**, *37*, 387−400.

(56) Valero, M.; Costa, S. M. B.; Ascenso, J. R.; Mercedes Velázquez, M.; Rodríguez, L. J. Complexation of the Non-Steroidal Anti-Inflammatory Drug Nabumetone with Modified and Unmodified Cyclodextrins. *J. Inclusion Phenom. Macrocyclic Chem.* **1999**, *35*, 663−677.

(57) Zhang, H.; Feng, W.; Li, C.; Lv, Y.; Tan, T. A Model for the Shuttle Motions of Puerarin and Daidzin inside the Cavity of β-Cyclodextrin in Aqueous Acetic Acid: Insights from Molecular Dynamics Simulations. *J. Mol. Model.* **2012**, *18*, 221−227.

(58) Auletta, T.; de Jong, M. R.; Mulder, A.; van Veggel, F. C. J. M.; Huskens, J.; Reinhoudt, D. N.; Zou, S.; Zapotoczny, S.; Schönherr, H.; Vancso, G. J.; Kuipers, L. β-Cyclodextrin Host−Guest Complexes Probed under Thermodynamic Equilibrium: Thermodynamics and AFM Force Spectroscopy. *J. Am. Chem. Soc.* **2004**, *126*, 1577−1584.

(59) Yu, Y.; Chipot, C.; Cai, W.; Shao, X. Molecular Dynamics Study of the Inclusion of Cholesterol into Cyclodextrins. *J. Phys. Chem. B* **2006**, *110*, 6372−6378.

(60) Filippini, G.; Goujon, F.; Bonal, C.; Malfreyt, P. Energetic Competition Effects on Thermodynamic Properties of Association between β-CD and Fc Group: A Potential of Mean Force Approach. *J. Phys. Chem. C* **2012**, *116*, 22350−22358.

(61) Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99a Mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255−1273.

(62) Borghetti, G. S.; Pinto, A. P.; Lula, I. S.; Sinisterra, R. D.; Teixeira, H. F.; Bassani, V. L. Daidzein/Cyclodextrin/Hydrophilic Polymer Ternary Systems. *Drug Dev. Ind. Pharm.* **2011**, *37*, 886−893.

(63) Todorova, N. A.; Schwarz, F. P. The Role of Water in the Thermodynamics of Drug Binding to Cyclodextrin. *J. Chem. Thermodyn.* **2007**, *39*, 1038−1048.

(64) Goyenechea, N.; Sánchez, M.; Vélaz, I.; Martín, C.; Martínez-Ohárriz, M. C.; González-Gaitano, G. Inclusion Complexes of Nabumetone with β-Cyclodextrins: Thermodynamics and Molecular Modelling Studies. Influence of Sodium Perchlorate. *Luminescence* **2001**, *16*, 117−127.

(65) Brocos, P.; Díaz-Vergara, N.; Banquy, X.; Pérez-Casas, S.; Costas, M.; Piñeiro, A. n. Similarities and Differences between Cyclodextrin−Sodium Dodecyl Sulfate Host−Guest Complexes of Different Stoichiometries: Molecular Dynamics Simulations at Several Temperatures. *J. Phys. Chem. B* **2010**, *114*, 12455−12467.

(66) Ishizu, T.; Kintsu, K.; Yamamoto, H. NMR Study of the Solution Structures of the Inclusion Complexes of β-Cyclodextrin with (+)-Catechin and (−)-Epicatechin. *J. Phys. Chem. B* **1999**, *103*, 8992−8997.

(67) Yan, C.; Xiu, Z.; Li, X.; Hao, C. Molecular Modeling Study of β-Cyclodextrin Complexes with (+)-Catechin and (−)-Epicatechin. *J. Mol. Graphics Modell.* **2007**, *26*, 420−428.

(68) Zifferer, G.; Sellner, B.; Kornherr, A.; Krois, D.; Brinker, U. H. Molecular Dynamics Simulations of β-Cyclodextrin-Aziadamantane Complexes in Water. *J. Phys. Chem. B* **2008**, *112*, 710−714.

(69) He, J.; Chipot, C.; Shao, X.; Cai, W. Cyclodextrin-Mediated Recruitment and Delivery of Amphotericin B. *J. Phys. Chem. C* **2013**, *117*, 11750−11756.

(70) Liu, P.; Chipot, C.; Shao, X.; Cai, W. How Do α-Cyclodextrins Self-Organize on a Polymer Chain? *J. Phys. Chem. C* **2012**, *116*, 17913−17918.

(71) Liu, P.; Chipot, C.; Shao, X.; Cai, W. Solvent-Controlled Shuttling in a Molecular Switch. *J. Phys. Chem. C* **2012**, *116*, 4471−4476.

(72) Lopez, C. A.; de Vries, A. H.; Marrink, S. J. Computational Microscopy of Cyclodextrin Mediated Cholesterol Extraction from Lipid Model Membranes. *Sci. Rep.* **2013**, *3*, 2071.

(73) Yamazaki, T.; Kovalenko, A. Spatial Decomposition Analysis of the Thermodynamics of Cyclodextrin Complexation. *J. Chem. Theory Comput.* **2009**, *5*, 1723−1730.

(74) Wickstrom, L.; He, P.; Gallicchio, E.; Levy, R. M. Large Scale Affinity Calculations of Cyclodextrin Host−Guest Complexes: Understanding the Role of Reorganization in the Molecular Recognition Process. *J. Chem. Theory Comput.* **2013**, *9*, 3136−3150.

(75) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2001**, *106*, 521−533.

(76) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640−647.

(77) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(78) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273−276.

(79) Hetényi, C.; van der Spoel, D. Toward Prediction of Functional Protein Pockets Using Blind Docking and Pocket Search Algorithms. *Protein Sci.* **2011**, *20*, 880−893.

(80) Carlsson, J.; Boukharta, L.; Åqvist, J. Combining Docking, Molecular Dynamics and the Linear Interaction Energy Method to Predict Binding Modes and Affinities for Non-Nucleoside Inhibitors to HIV-1 Reverse Transcriptase. *J. Med. Chem.* **2008**, *51*, 2648−2656.

(81) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2005**, *49*, 5912−5931.

(82) Zhou, H. X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092−4107.

(83) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.

(84) Carlsson, J.; Åqvist, J. Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding. *J. Phys. Chem. B* **2005**, *109*, 6448−6456.

(85) Chen, W.; Chang, C.-E.; Gilson, M. K. Calculation of Cyclodextrin Binding Affinities: Energy, Entropy, and Implications for Drug Design. *Biophys. J.* **2004**, *87*, 3035−3049.

D25

hetenyi.csaba_83_23

THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

pubs.acs.org/JPCC

Terms of Use

Cooperative Binding of Cyclodextrin Dimers to Isoflavone Analogues Elucidated by Free Energy Calculations

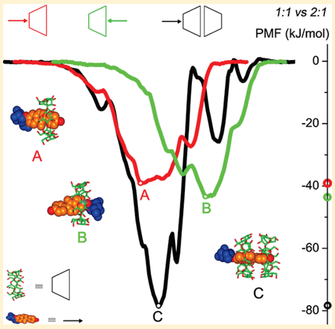
Haiyang Zhang,[†,‡] Tianwei Tan,*[,†] Csaba Hetényi,[§] Yongqin Lv,[†] and David van der Spoel*[,‡]

[†]Beijing Key Laboratory of Bioprocess, Department of Biochemical Engineering, Beijing University of Chemical Technology, Box 53, 100029 Beijing, China

[‡]Uppsala Center for Computational Chemistry, Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, Box 596, SE-75124 Uppsala, Sweden

[§]Molecular Biophysics Research Group, Hungarian Academy of Sciences, Pázmány sétány 1/C, H-1117 Budapest, Hungary

S Supporting Information



ABSTRACT: Dimerization of cyclodextrin (CD) molecules is an elementary step in the construction of CD-based nanostructured materials. Cooperative binding of CD cavities to guest molecules facilitates the dimerization process and, consequently, the overall stability and assembly of CD nanostructures. In the present study, all three dimerization modes (head-to-head, head-to-tail, and tail-to-tail) of β-CD molecules and their binding to three isoflavone drug analogues (puerarin, daidzin, and daidzein) were investigated in explicit water surrounding using molecular dynamics simulations. Total and individual contributions from the binding partners and solvent environment to the thermodynamics of these binding reactions are quantified in detail using free energy calculations. Cooperative drug binding to two CD cavities gives an enhanced binding strength for daidzin and daidzein, whereas for puerarin no obvious enhancement is observed. Head-to-head dimerization yields the most stable complexes for inclusion of the tested isoflavones (templates) and may be a promising building block for construction of template-stabilized CD nanostructures. Compared to the case of CD monomers, the desolvation of CD dimers and entropy changes upon complexation prove to be influential factors of cooperative binding. Our results shed light on key points of the design of CD-based supramolecular assemblies. We also show that structure-based calculation of binding thermodynamics can quantify stabilization caused by cooperative effects in building blocks of nanostructured materials.

INTRODUCTION

Cyclodextrins (CDs) are promising building blocks extensively used in the construction of nanostructured materials with sophisticated structures and functions.[1,2] CDs belong to a class of cyclic oligosaccharides with more than six D-glucopyranose residues linked together via α-1,4 glycosidic bonds and arrangement of these residues in a ring endows CDs with a somewhat hydrophobic cavity and a hydrophilic surface.[3,4] This property permits association of varied guest molecules with suitable size to form stable host−guest complexes or supramolecular assemblies, which leads to a variety of fascinating applications in many fields like pharmaceutical research.[5−7] In recent years, construction of one- and multidimensional nanoarchitectures using CDs as building blocks has attracted much attention, particularly due to their alluring potential in molecular machines[8−10] and functional materials.[11−13] CD-based nanostructures integrate together a number of functional groups that have been already captured by CD cavities. These functional groups along with CD cavities provide multiple binding sites for substrates, allowing one to mimic the cooperative multimode complexation existing in biological systems widely. CD-based nanoarchitectures are therefore acknowledged to be ideal candidates for drug or gene carriers[14−16] and artificial enzyme models.[17]

Cooperative binding of at least two CD monomers to a template (also known as guest) molecule is the driving force responsible for self-assembly processes in the construction of CD-based nanoarchitectures. For the case without template, the assembly is usually driven by hydrophobic interactions between substituent arms of CD derivatives with the neighboring cavities of other CDs.[1] The following text will focus on the former case with template. Polymer chains such as poly(ethylene glycol) (PEG) and poly(propylene glycol) (PPG) are often used as a template to thread several CD cavities for the construction of one-dimensional nanoarchitectures like (pseudo)polyrotaxanes.[2,14,18] Furthermore, CDs can be grafted covalently to the polymer chain as a bulky stopper for polyrotaxanes, and cooperative binding of two bulky CD cavities to one template molecule (like C_{60})[19] allows construction of long nanowires based on the polyrotaxanes. Starting from CD-based polyrotaxanes, one can prepare nanotubes by covalent reactions of neighboring CD units with short cross-linking agents such as epichlorohydrin, followed by the cutoff of bulky ends and removal of the

Received: December 9, 2013
Revised: March 11, 2014
Published: March 14, 2014

dx.doi.org/10.1021/jp412041d | J. Phys. Chem. C 2014, 118, 7163−7173

polymer thread.[20] These tubular polymers are capable of including long guest molecules like 1,6-dimethylhexatriene inside the molecular tube efficiently. Randomly cross-linked CDs without preassembly by a polymer chain cannot form a tube easily and hence do not possess such an inclusion ability.[2]

Typical CDs used as the building blocks contain 6, 7, and 8 glucopyranose residues, denoted as α-, β-, and γ-CD, respectively. Harada and co-workers characterized topology structures of α-CD/PEG, β-CD/PEG, and β-CD/PPG (pseudo)polyrotaxanes using X-ray crystallography and reported that all CD monomers are oriented as head-to-head (HH) and tail-to-tail (TT) dimers through threading onto the polymer chain.[21−23] They indicated that secondary hydroxyl groups of CDs hydrogen-bond to each other forming a tight hydrogen-bonding network and that the interactions between primary hydroxyls are weak. Mavridis et al. observed an unusual crystal of β-CD trimers in HH and head-to-tail (HT) fashions which cooperatively bind to two guest molecules.[24] HT orientations were found in the crystal packing of γ-CDs as well.[25] Figure 1a depicts β-CD dimers in the three orientations of HH, HT, and TT taken from Mavridis's work;[24] head indicates the wide (secondary) rim of β-CD and tail the narrow (primary) rim.

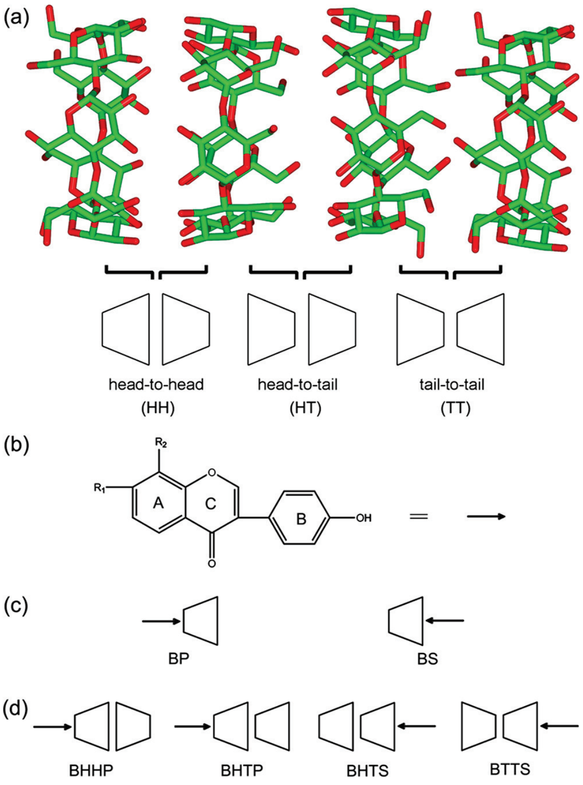


Figure 1. Molecular structure of (a) β-CD dimers and (b) isoflavone guests and possible [host:guest] binding modes with stoichiometric ratios of (c) 1:1 and (d) 2:1. Head means the secondary rim of β-CD and tail the primary rim. The guest molecules include puerarin (R1 = H, R2 = glucose), daidzin (R1 = glucose, R2 = H), and daidzein (R1 = H, R2 = H). A, B, and C denote relevant isoflavone rings. The arrow indicates the guest molecule and the orientation that the guest penetrates into β-CD cavity.

Because of the outstanding performance of CD-based nanoarchitectures, it is highly desirable to find out the mechanism underlying cooperative effects of CD-based assemblies. Molecular dynamics simulation serves as a powerful tool for exploring the mechanism associated with CD-based systems and has contributed valuable explanations for experimental observations.[26−30] Association of two CD monomers (i.e., dimer) is an essence for cooperative binding of CD cavities. Many theoretical reports focused on the relative stability of noncovalent CD dimers in HH, HT, and TT fashions (Figure 1a) and revealed that hydrogen-bonding (HB) interactions between hydroxyl groups of adjacent CD monomers are a key factor determining the dimer stability.[31−36] Cai and co-workers recently examined dimerization of α-CDs onto a PEG chain using free energy calculations and Monte Carlo simulations.[37] They indicated that the dimerization is driven primarily by HB interactions between two α-CDs and that HH is preferred over HT and TT. Pineiro et al. evaluated α-, β-, and γ-CD complexes with sodium dodecyl sulfate (SDS) in ratios of 1:2 and 2:1 through MD simulations and reported that [CD_2:SDS] in the HH orientation seems a potential building block for nanotubular polymers.[27] Marrink and co-workers performed potential of mean force (PMF) calculations to investigate the mechanism of cyclodextrin-mediated extraction of cholesterol from model membranes.[38,39] In previous work, we investigated the dissociation of β-CD HH dimer through PMF calculations and concluded that the dimer binding depends on the guest and solvent properties.[40]

Here we present an extensive free energy examination on cooperative binding of β-CD dimers (HH, HT, and TT) to three isoflavone analogues (puerarin, daidzin, and daidzein) through molecular dynamics (MD) simulation. The three isoflavone components (guest molecules) have potential use in medicinal therapies,[41,42] and a more efficient encapsulation of these drugs by β-CD dimers promotes their practical applications. Also, structural properties of the isoflavone skeleton with/without glucose motivated us to choose them as template molecules to examine hydrophobic and hydrophilic interactions that constitute the main driving forces responsible for the construction of CD-based nanostructures. A number of free energy calculations have been implemented to evaluate 1:1 and 2:1 [CD:guest] complexes,[27,37−40,43−47] while few reports consider all possible cooperative binding of CD cavities. In this work free energy profiles governing all possible formation processes of [β-CD_2:guest] complexes were calculated with umbrella sampling.[48] Center of mass (COM) pulling[49] was employed to generate configuration sequences for umbrella sampling simulations. Details on COM pulling and PMF techniques have been presented in refs 46 and 49−54. From PMF and entropy calculations, total and individual contributions from enthalpy and entropy were quantified in detail using a recently proposed method for 1:1 binding.[55] The results exhibit a comprehensive thermodynamic and energetic characterization for cooperative effects of CD dimers toward guest molecules. As a fundamental step in the construction of nanostructures with cooperatively bound units (like CDs), dimerization of CD molecules studied here offers a generalized picture on molecular assemblies of CDs by cooperative binding to a template. Implications for design of template molecules and CD assembly models in building blocks of nanoarchitectures are discussed at the end of this work.

METHODS

The initial coordinates of β-CD dimers (HH, HT, and TT) were taken from the Cambridge Crystallographic Data Center (CCDC no. 648855)[24] where β-CD trimers formed a channel-like structure (Figure 1a). Molecular structures of isoflavone guests (puerarin, daidzin, and daidzein) are shown in Figure 1b. All the binding modes of 1:1 and 2:1 [CD:guest] complexes are given in Figures 1c and 1d, respectively. The q4md-CD force field[56] was used to model β-CD and the generalized Amber force field (GAFF)[57] for the guests. The rigid model TIP3P[58] was used for water molecules. All the simulations were carried out at 300 K with GROMACS (version 4.5.5).[59−61] System equilibrations were performed in the *NPT* ensemble (P = 1 bar) and production simulations in the *NVT* ensemble. Other simulation protocols were the same as in the refs 55 and 62.

Each system contained one β-CD dimer, one guest, and approximately 4100 water molecules in a simulation cell of 5 × 5 × 5 nm^3. The dimer was centered in the box with Z-coordinates of its glycosidic oxygen atoms approximately located at Z = −0.3 or +0.3 nm for the two monomers, respectively, making the cavity axis of β-CD dimer parallel to the Z-axis. The distance between the center of mass (COM) of the B-ring of the guest and that of 14 glycosidic oxygens of the dimer along the Z-axis was defined as the reaction coordinate ξ. Figure 2 shows the definition of ξ for the HT dimer with daidzein in the BHTS mode (see Figure 1d for nomenclature). Glycosidic oxygen atoms of β-CD dimers were harmonically restrained and used as an immobile reference for pulling simulations. The B-ring of the guest was pulled through the dimer cavity from the primary (P) or secondary (S) rim along the Z-axis over 1 ns with a pulling rate of 0.005 nm ps^{-1}. All the pulling parameters were the same as in the ref 55 where the 1:1 binding modes (BP and BS, Figure 1c) have been evaluated. In this work, the guest sampled 5 nm covering the entire ξ of [−2.5, 2.5], and a formation process for 2:1 inclusion complexes was detected during the pulling simulation. We then selected 101 windows in the [−2.5, 2.5] interval with a distance equal to 0.05 nm between adjacent positions and these windows were used for umbrella sampling simulations. Following the same procedure, we simulated three guest molecules in the four binding modes of BHHP, BHTP, BHTS, and BTTS (Figure 1d) and therefore obtained 12 PMF profiles in total. The total simulation time for a single PMF was 1.01 μs. For each window the first 2 ns was removed for equilibration, and the rest (2−10 ns) was used for all the data analysis. Details on the calculation of thermodynamic parameters (ΔG, ΔH, or $-T\Delta S$) are given in the Supporting Information.

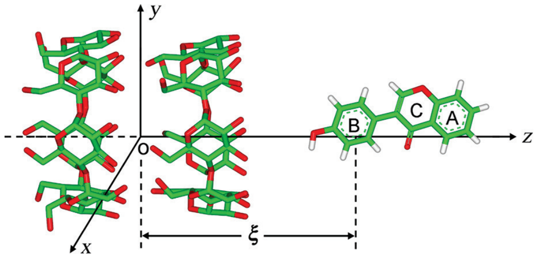


Figure 2. Definition of the reaction coordinate ξ for the BHTS mode with daidzein.

RESULTS

Binding Modes. PMF profiles for the formation process of [β-CD$_2$:guest] inclusion complexes with the three isoflavone guests along ξ in the BHHP mode are shown in Figure 3a. The guest approaches the dimer from the primary rim of β-CD, penetrates into the channel-like cavity, and then gets out of the cavity along the +ξ direction. All the PMFs on both sides of ξ amount to zero and level off corresponding to the completely separated state of the binding partners.

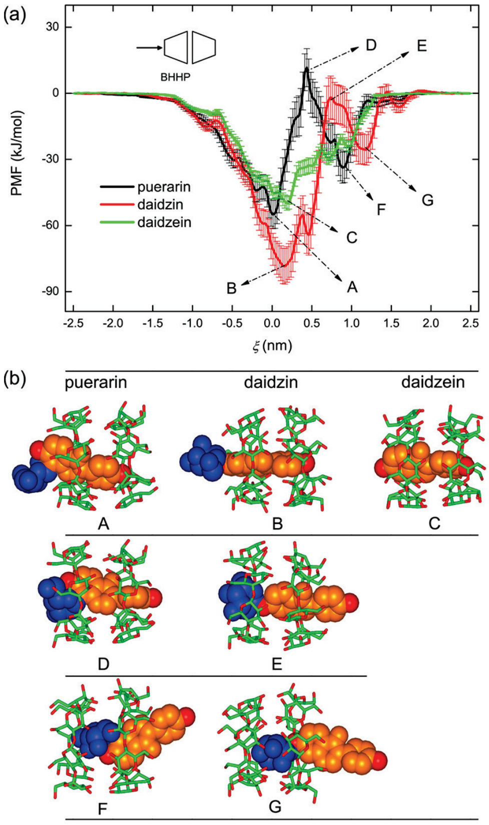


Figure 3. (a) Potential of mean force (PMF) profiles for the [β-CD$_2$:guest] complex formation in the BHHP binding mode and (b) representative inclusion configurations along ξ. β-CD dimer and guest molecules are shown with stick and space-filling models, respectively. The glucose group of guest is colored in blue and isoflavone skeletons in orange.

Representative configuration states (A−G) in the PMFs are marked in Figure 3a and given in Figure 3b. As the isoflavone skeletons (hydrophobic moieties) of the guests get close to the β-CD cavity, the PMF curves drop and become negative (i.e., thermodynamically favorable). When the isoflavone skeleton is located inside the channel-like cavity, leaving the glucose group (hydrophilic) outside, the most stable inclusion configurations of [β-CD$_2$:puerarin] and [β-CD$_2$:daidzin] complexes are sampled, namely the A- and B-states (Figure 3b), respectively. Approaching the cavity for the hydrophilic glucose further results in an upward trend for the PMFs (Figure 3a), revealing a thermodynamically unfavorable state. The most unfavorable states (central maxima in the PMFs) is that with the glucose group being entrapped inside the cavity of one β-CD monomer,

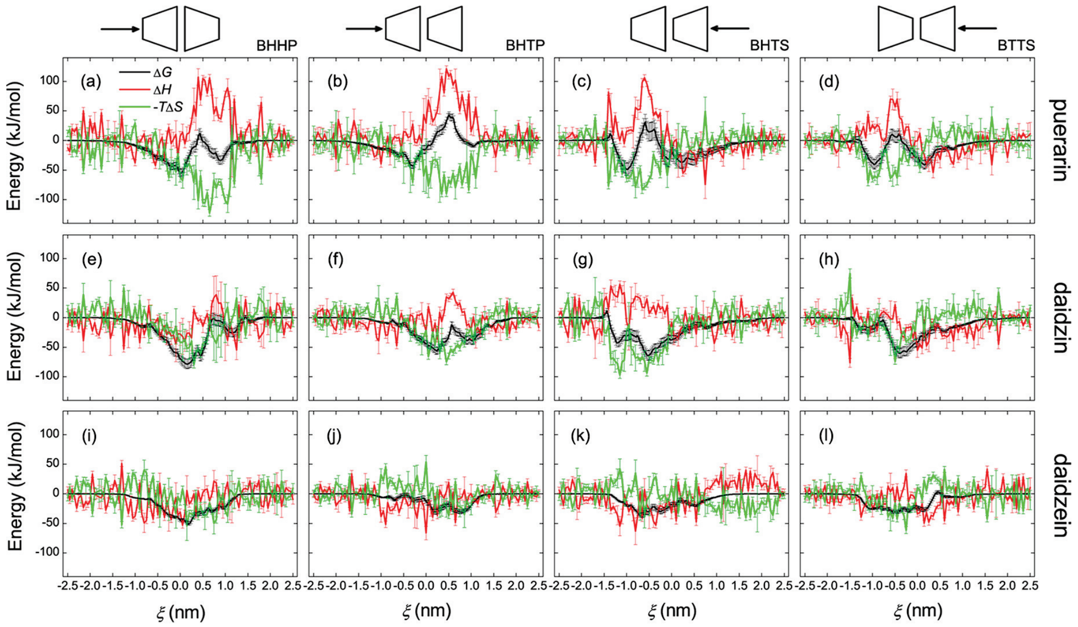


Figure 4. Thermodynamic profiles (ΔG, ΔH, and $-T\Delta S$) of the system for the complex formation of β-CD dimers with puerarin, daidzin, and daidzein in the binding modes of BHHP, BHTP, BHTS, and BTTS.

Table 1. Thermodynamic Parameters (kJ/mol) Calculated at 300 K for the Guests Studied

		dimer				monomer[a]		⟨ΔE⟩[b]	
guest	energy	BHHP	BHTP	BHTS	BTTS	BP	BS	dimer	monomer
puerarin	ΔG^0	−30	−19	−31	−27	−26	−32	−30	−32
	ΔH^0	−43	−21	−41	−39	−36	−41	−41	−41
	$-T\Delta S^0$	13	2	10	12	10	9	11	9
daidzin	ΔG^0	−38	−31	−36	−29	−24	−29	−37	−28
	ΔH^0	−49	−45	−52	−38	−32	−38	−50	−37
	$-T\Delta S^0$	11	14	16	9	8	9	13	9
daidzein	ΔG^0	−35	−18	−20	−18	−19	−22	−35	−21
	ΔH^0	−48	−28	−33	−26	−28	−29	−48	−29
	$-T\Delta S^0$	13	10	13	8	9	7	13	8

[a]Taken from ref 55. [b]Weighted on all binding modes using eq 1.

while the hydrophobic isoflavone skeleton of the guest still interacts with the β-CD cavity, as in the D- and E-states (Figure 3b). When the hydrophilic glucose stays approximately in the center of mass (COM) of β-CD dimer, such as the F- and G-states (Figure 3b), local favorable minima in the PMFs are observed, indicating that the COM region of the dimer is somewhat hydrophilic. The most stable configuration of [β-CD$_2$:daidzein] is similar to puerain and daidzein; see the C-state (Figure 3b) where daidzein is almost completely encapsulated by the β-CD head-to-head dimer. Unlike puerarin and daidzin, the PMF for daidzein however does not display an obvious central maximum (Figure 3a). For convenience, the inclusion models similar to A- and B-states are shortened for *GO* (glucose outside), to D- and E-states for *GIM* (glucose inside monomer), and to F- and G-states for *GID* (glucose inside dimer) in the forthcoming text.

PMF profiles (ΔG) for all three guests in the four binding modes of BHHP, BHTP, BHTS, and BTTS are presented in Figure 4. In our simulations all the guests are inserted into the dimer cavity from the primary or secondary rim of β-CD along the $+\xi$ or $-\xi$ direction, respectively, unless stated otherwise. In the PMFs a central maximum resulted from inclusion of the glucose unit inside the β-CD cavity (model *GIM*) and a local minimum from inclusion of the glucose unit in the dimer center (model *GID*) are observed for all binding modes of puerarin (Figure 4, panels a–d). Inclusion models for *GIM* are located approximately at ξ = +0.5 nm or −0.5 nm for BHHP and BHTP or BHTS and BTTS, respectively; models for *GID* at ξ = ~1.0 nm or −1.0 nm. For BHHP and BHTP the most stable states are similar to model *GO* (Figure 4, panels a and b), while for BHTS and BTTS the glucose unit positioned in the COM region of the dimer (model *GIM*) form the most stable states (Figure 4, panels c and d). Daidzin behaves similar to puerarin, while the most stable states adopt a *GO* model for all binding modes (Figure 4, panels e–h). The PMFs of daidzein for all binding modes do not show clear central maxima, and the inclusion modes with the isoflavone skeleton completely enclosed in the dimer are the most stable (Figure 4, panels i–l).

Binding Energetics. Enthalpy (ΔH) and entropy (ΔS) profiles of the system for the [β-CD$_2$:guest] complex formation along ξ are also given in Figure 4. Here entropy is given as

$-T\Delta S$. These profiles depict how enthalpy and entropy changes contribute to the binding energy and assist in understanding the thermodynamics of binding. For puerarin and daidzin clear enthalpy loss (positive ΔH) and entropy gain (positive ΔS) are observed in most cases, particularly when the glucose unit of the guest stays inside the β-CD cavity (Figure 4, panels a–h). In some cases the host–guest complexation is enthalpy-driven (negative ΔH), like *GO* models of puerarin in BHTS and BTTS modes (Figure 4, panels c and d). For daidzein, no obvious entropy changes (ΔS) are observed, and the binding seems to be exclusively driven by ΔH. Notice that the calculations force the guest artificially to access some region of the binding sites that are thermodynamically unstable, like the *GIM* model of puerarin (Figure 4, panels a–d), which allows us to sample the configuration space as much as possible.

Standard thermodynamic parameters (ΔG^0, ΔH^0, and ΔS^0) for all binding modes of dimer and monomer are given in Table 1 (see eqs S1–S4 in the Suppporting Information for calculation of ΔG^0, ΔH^0, and ΔS^0). For a quantitative evaluation these parameters are weighted by their Boltzmann factors using eq 1

$$\langle \Delta E \rangle = \frac{\sum_i \Delta E_i \exp[-\Delta G_i/RT]}{\sum_i \exp[-\Delta G_i/RT]} \tag{1}$$

where ΔE can be ΔG, ΔH or, $-T\Delta S$. The weighted values are listed in Table 1 as well. A good agreement between calculated and experimental ΔG^0 for 1:1 associations have been shown in the ref 55, which validates the veracity of our calculations. No experimental data for 2:1 associations are available for direct comparison with the calculations yet. Similar to the monomer, the dimer binding to its guest is predominantly enthalpy-driven, and entropy loss cancels out about one-quarter of enthalpy gain. Cooperative binding of the dimer to puerarin does not result in an obvious increase in the binding strength and even in a decrease for the BHTP mode. However, cooperative effects of the two monomers give a clear increase by ~30% (for daidzin) and 60% (daidzein) in binding free energies (ΔG^0). BHHP seems the best mode for such guest binding, followed by BHTS and by BHTP and BTTS.

Decomposition of Energy Terms. For a deeper insight into the enthalpy and entropy profiles, ΔH and ΔS are decomposed into individual contributions from the binding partners and solvent environment. The decomposition refers to eqs S6 and S7 in the Supporting Information. Figure 5 shows the ΔH decomposition for [β-CD$_2$:guest] complexes with puerarin, daidzin, and daidzein along ξ in the BHHP mode. ΔH_{host} and ΔH_{guest} are bonded interactions (torsion energies of bond angle and dihedral angle) of host and guest molecules (indicated by black and red lines), respectively. $\Delta H_{\text{host–host}}$ and $\Delta H_{\text{guest–guest}}$ (green and blue) belong to intramolecular nonbonded interactions of the binding partners. These four items quantify the changes in potential energies of host and guest resulting from the fluctuations in atomic positions. Bonded interactions of the rigid TIP3P[58] water amount to zero and $\Delta H_{\text{sol-sol}}$ (cyan) thus contains nonbonded intra- and intermolecular interactions between water molecules only. The other three terms, $\Delta H_{\text{host–guest}}$, $\Delta H_{\text{host–sol}}$, and $\Delta H_{\text{guest–sol}}$ (magenta, dark yellow, and orange), describe nonbonded intermolecular interactions between different kinds of molecules.

As shown in Figure 5a, the bonded items (ΔH_{host} and ΔH_{guest}) tend to disfavor [β-CD$_2$:puerarin] complexation (positive values), while the nonbonded ones of the binding partners ($\Delta H_{\text{host–host}}$ and $\Delta H_{\text{guest–guest}}$) in contrast favor the complexation (negative). By inclusion of puerarin, the interaction between host and guest is strengthened (negative $\Delta H_{\text{host–guest}}$), and both host and guest molecules are desolvated, as indicated by more positive $\Delta H_{\text{host–sol}}$ and $\Delta H_{\text{guest–sol}}$. Water molecules are shown to gain enthalpy (negative $\Delta H_{\text{sol–sol}}$) favoring the binding. Similar observations were found for daidzin (Figure 5b) and daidzein (Figure 5c). Without the glucose unit (Figure 1b), daidzein shows more symmetric profiles and a relatively small change in $\Delta H_{\text{guest–sol}}$, $\Delta H_{\text{host–guest}}$, and $\Delta H_{\text{sol–sol}}$ (Figure 5c).

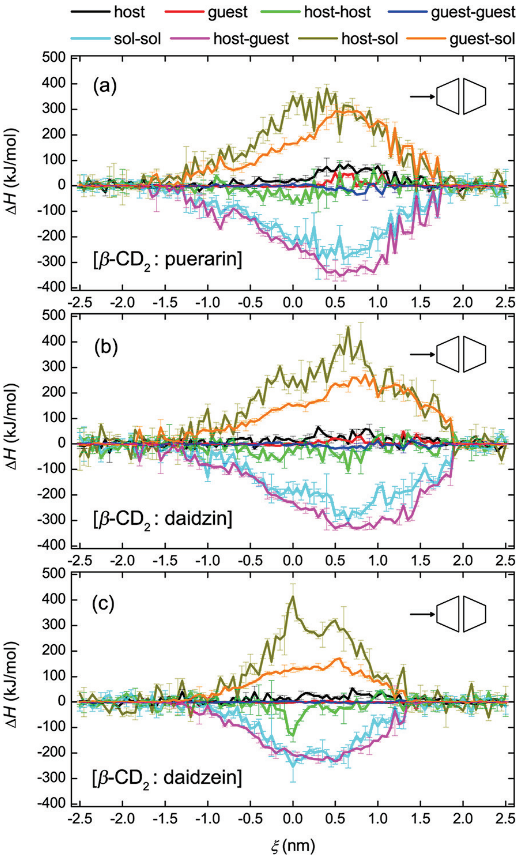


Figure 5. Enthalpy decomposition for the complex formation of β-CD dimers with (a) puerarin, (b) daidzin, and (c) daidzein in the BHHP mode.

Considering Figures 3 and 5, the most stable [β-CD$_2$:guest] complexes with puerarin or daidzin (model *GO*, ξ = 0.0–0.2 nm) do not correspond to the states where the global minima of $\Delta H_{\text{host–guest}}$ and $\Delta H_{\text{sol–sol}}$ are achieved (ξ = 0.5–1.0 nm). At ξ = 0.5–1.0 nm, the contributions from the unfavorable enthalpy items of ΔH_{host}, ΔH_{guest}, $\Delta H_{\text{host–sol}}$, and $\Delta H_{\text{guest–sol}}$ seem to be maximized and reduce the enthalpy gain (Figure 5, panels a and b). The most unstable complexes at ξ = 0.4 nm (puerarin) or 0.76 nm (daidzin), such as the D- and E-states (model *GIM*) in Figure 3, occur in coincidence with the states in which the host desolvation ($\Delta H_{\text{host–sol}}$) reaches its maximum. For the most stable [β-CD$_2$:daidzein] complex all enthalpy components get very close to their maximum or minimum values, either favoring the complexation or not.

Figure 6 presents the ΔS decomposition for puerarin (Figure 6a), daidzin (Figure 6b), and daidzein (Figure 6c) in the BHHP binding mode. Here configurational entropies of host and guest molecules were calculated from the covariance matrices of atomic fluctuations using the quasi-harmonic approximation.[63] A length of at least 8 ns is needed for our simulations to ensure the convergence of such entropy calculations, as shown in Figures S1 and S2 of the Supporting Information. An obvious entropy loss of the host (positive $-T\Delta S$) and compensating entropy gain of the solvent (negative $-T\Delta S$) is observed. When included inside the CD cavity, daidzin shows a significant entropy loss (Figure 6b), but this is not observed for puerarin (Figure 6a) and daidzein (Figure 6c). As discussed in previous work on the monomer binding,[55] the glucose rotation of daidzin was affected much more than that of puerarin when entrapped inside the CD cavity, and daidzein did not display any obvious entropy change due to its structural rigidity. In this work, similar entropy changes for these three guests inside the dimer were found (Figure 6). Daidzein gives more symmetric ΔS profiles than either of puerarin or daidzin. As can be seen from Figures 3 and 6, configurational changes of β-CD are greatly affected by guest inclusion (in particular, when the glucose unit of the guest stays inside the cavity), and the most stable complexes (those with the lowest ΔG) do not correspond to the states with maximum or minimum values of the three ΔS components.

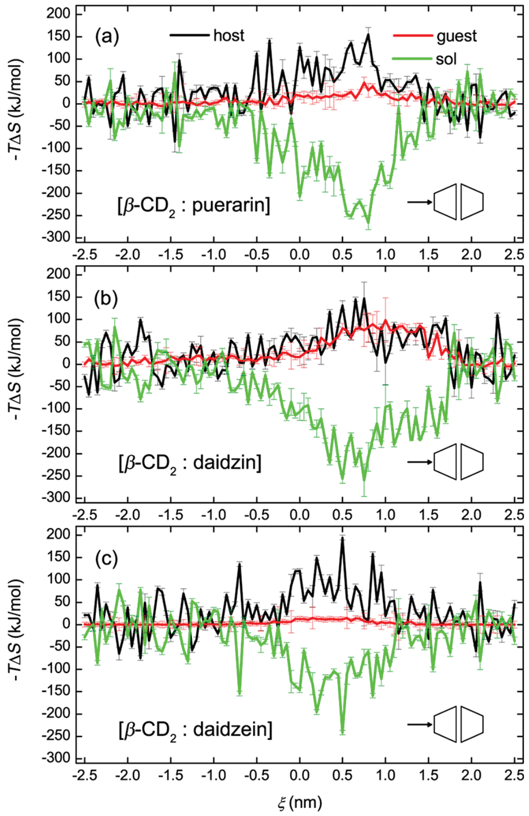


Figure 6. Entropy decomposition for the complex formation of β-CD dimers with (a) puerarin, (b) daidzin, and (c) daidzein in the BHHP mode.

Hydrogen Bonding. Polar moieties of the guests like the glucose units are observed to hydrogen bond to β-CD dimers. The number of hydrogen bonds (HBs) between the binding partners along ξ was analyzed to explore the role of HBs in the complex formation (Figure 7).

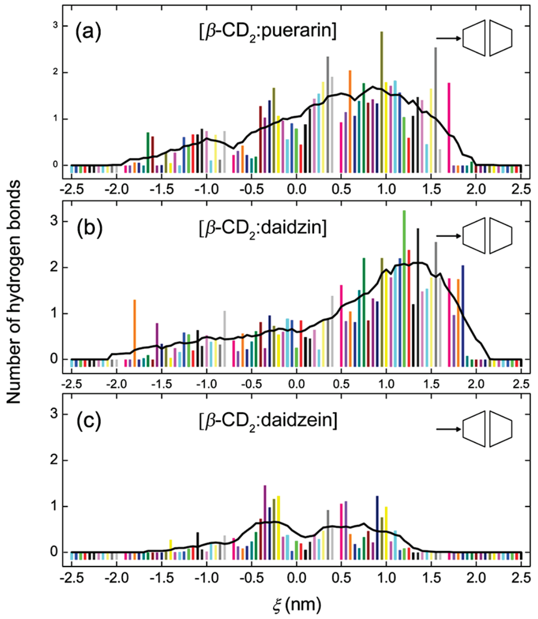


Figure 7. Hydrogen bonding strength during the complex formation of β-CD dimers with (a) puerarin, (b) daidzin, and (c) daidzein in the BHHP mode. The bold black lines represent trend curves smoothed by 10-point-window adjacent averaging.

Here we use a geometrical criterion for HB definition, based on distance and angle cutoffs of 0.35 nm and 30°.[64] No obvious HB interactions are observed at $\xi < -0.5$ nm (Figure 7) where the B-ring of the guest approaches the dimer cavity step by step, implying that the hydroxyl group connected to the B-ring contributes little to the binding. For puerarin and daidzin, the most stable states (Figure 3a) have just one HB (ξ = 0.0−0.5 nm in Figure 7, panels a and b). This HB is formed between the glucose unit of guest and the primary rim of β-CD. At the central region of the dimer (ξ = 1.0−1.5 nm), a stronger HB interaction (about two HBs) is observed for puerarin and daidzin (Figure 7, panels a and b) due to efficient contacts between polar moieties of the binding partners. These HB interactions are expected to contribute to the overall stabilization process by lowering down the PMF curves somewhat.[65] Few influences of HBs on the binding are detected for daidzein (Figure 7c). It should be noted, however, that the stability of HBs, which is the activation energy needed to break HBs, is virtually independent of the environment.[66]

Comparison of Monomer with Dimer. PMF profiles for inclusion complexes of the β-CD monomer in the BP and BS modes and of β-CD dimer in the BHHP mode with the studied guests along ξ are shown in Figure 8. Here the guest passes through the host cavity along $+\xi$ from the primary rim of β-CD for BP and BHHP modes and from the secondary rim for BS. Because of differences in the definition of ξ between the monomer and dimer systems, the PMFs for BP and BS are shifted by 0.3 nm along $-\xi$ and $+\xi$, respectively, allowing for

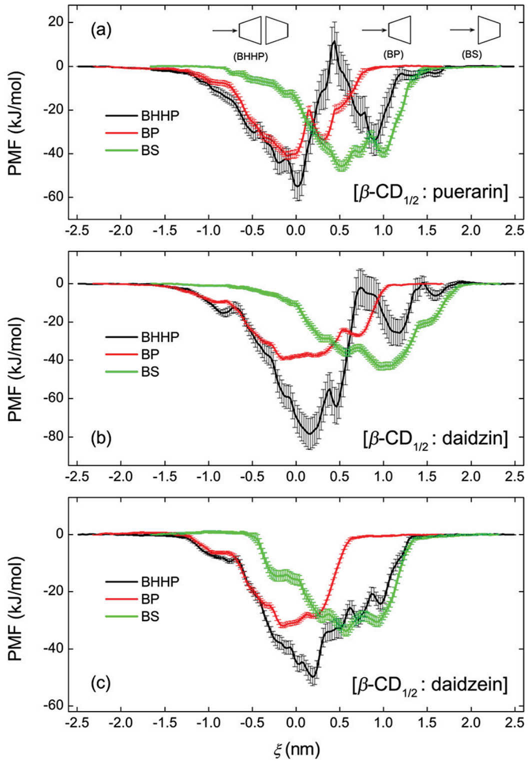


Figure 8. PMF comparison of β-CD monomer in the BP and BS modes with the dimer in the BHHP mode for the complex formation with (a) puerarin, (b) daidzin, and (c) daidzein. PMFs for BP and BS were taken from ref 55.

direct comparison with the dimer. The most stable states for puerarin complexes with the β-CD monomer (BP) and dimer (BHHP) are located approximately in the same position of ξ = ~0.0 nm, and the dimer only gives a small increase in the binding strength (Figure 8a). However, there exists an obvious enhancement in the binding affinity of the head-to-head dimer to daidzin (Figure 8b) and daidzein (Figure 8c).

The cavity of one β-CD monomer does not encapsulate an isoflavone skeleton efficiently, leaving the skeleton in part exposed to the aqueous environment (Figure 3b). The presence of another monomer donating its hydrophobic cavity allows enclosing the exposed moieties of guest. Two β-CD monomers seem enough for encapsulation of such an isoflavone skeleton and cooperative binding of the two monomers forms a more stable inclusion complex (Figures 3 and 8). As seen from panels a and b in Figure 8, approaching the β-CD cavity for the glucose unit of puerarin and daidzin is disfavored thermodynamically (indicated by the central maxima), either in complexation with the monomer or with the dimer, whereas inclusion of the glucose unit inside the cavity seems somewhat favorable (indicated by the local minima). The energy barrier that prevents the glucose unit from further entering the β-CD cavity is higher for puerarin than for daidzin (Figure 8, panels a and b). No significant energy barriers for daidzein binding are detected (Figure 8c).

For evaluation of individual contributions to the binding affinity, the ΔH and ΔS components (see eqs S6 and S7 in the Supporting Information) are weighted by their Boltzmann factors using eq 1 and listed in Table 2. Increment factors (I) relative to the monomer are computed and given in Table 2 as well for comparison. I = 0 means that there is no significant difference between monomer and dimer; I = 1 that the energy contribution is exactly doubled. As shown in Table 2, ΔH_{host} and $\Delta H_{\text{host-host}}$ for the dimer are strengthened significantly with an increment factor (I) larger than 2 in most cases, indicating that atomic positions of host molecules changes obviously (i.e., the host molecule adjusts its configuration for a better encapsulation of its guest), in line with the observed entropy changes of the host ($-T\Delta S_{\text{host}}$, I = 1.6−8.3). Configurations of guest molecules do not change that much upon complexation and smaller values for ΔH_{guest}, $\Delta H_{\text{guest-guest}}$, and $-T\Delta S_{\text{guest}}$ are observed (Table 2).

Water−water enthalpy ($\Delta H_{\text{sol-sol}}$) increases by ~50% for puerarin and daidzin and doubles for daidzein (I = 1). The cooperative effects of two monomers does not make the interaction between host and guest molecules ($\Delta H_{\text{host-guest}}$) exactly twice as large, with an increment factor of I = 0.4−0.7. Upon binding, two β-CD monomers in the dimer are desolvated more intensively ($\Delta H_{\text{host-sol}}$, I = 1.6−8.3), while the guests just show small desolvation increments ($\Delta H_{\text{guest-sol}}$, I = 0.3−0.5). As a result of the desolvation, the water entropy increases correspondingly ($-T\Delta S_{\text{sol}}$, I = 2.2−10.7).

Table 2. Individual Contributions (kJ/mol) of ΔH and ΔS Weighted by Boltzmann Factors for the BHHP Binding Mode (Standard Deviations in Parentheses)

	puerarin		daidzin		daidzein	
$\langle \Delta E\rangle^a$	BHHP	I^b	BHHP	I^b	BHHP	I^b
ΔH_{host}	23(3)	14.0	18(3)	17.0	28(4)	27.0
ΔH_{guest}	−3(1)	0.0	3(1)	0.0	0(1)	0.0
$\Delta H_{\text{host-host}}$	−58(6)	2.9	−11(2)	0.3	−37(4)	2.9
$\Delta H_{\text{guest-guest}}$	4(2)	5.0	−2(1)	0.0	0(1)	0.0
$\Delta H_{\text{sol-sol}}$	−214(9)	0.6	−181(9)	0.5	−203(10)	1.0
$\Delta H_{\text{host-guest}}$	−238(6)	0.4	−254(8)	0.6	−215(8)	0.7
$\Delta H_{\text{host-sol}}$	333(9)	1.1	243(8)	0.7	291(10)	1.4
$\Delta H_{\text{guest-sol}}$	165(8)	0.3	152(6)	0.3	131(7)	0.5
$-T\Delta S_{\text{host}}$	102(5)	2.1	49(3)	1.6	107(6)	8.3
$-T\Delta S_{\text{guest}}$	17(3)	2.1	32(4)	0.8	13(2)	12.0
$-T\Delta S_{\text{sol}}$	−181(6)	2.9	−128(4)	2.2	−152(6)	10.7

[a] Weighted on all complex states along the entire ξ of [−2.5, 2.5] using eq 1. [b] Increment $I = (d - m)/|m|$ where d is the energy item for dimer and m the value averaged on monomers BP and BS. Values for the monomers were taken from ref 55.

DISCUSSION

Cyclodextrin (CD) dimer is a basic building block for the construction of diversified nanoarchitectures such as inclusion complexes, molecular necklaces, nanotubes, nanowires, and vesicles.[2,8,67−69] The cooperative binding of guest molecules to CD cavities is one of the most important driving forces in the assembly and stabilization of these architectures. In order to achieve such a cooperative effect, two CD monomers can be bridged together by a linker,[47,70−73] mostly through covalent reactions of primary hydroxyls with the linker. Bridged bis(CD)s with functional linkers lead to an increase in the binding strength and molecular selectivity compared to native CD monomers.[72] Here we focused on the noncovalent case of 2:1 stoichiometry [CD:guest] complexes, which reveals a thermodynamic background for the stabilization of CD assemblies by cooperative binding of CD cavities to a guest (template) molecule.

Three isoflavone analogues (puerarin, daidzin, and daidzein) were tested as template molecules in this work. The former two are isomers belonging to isoflavone glycosides; puerarin is 8-C-glucoside of daidzein and daidzin 7-*O*-glucoside of daidzein (Figure 1b). The differences in the position of the glucose unit result in different molecule shapes and hence in different binding affinities to the *β*-CD dimer. Puerarin displays as a branch-like structure and daidzin a stick-like one. The glucose unit induces a high-energy barrier and hinders the tested template from further penetrating into the CD cavity, as indicated by the barriers in the PMFs (Figures 3, 4, and 7). As shown in Table 1 and Figure 8, there is no obvious increase in the binding strength of the dimer to puerarin, indicating that a second monomer is not necessary if a stable inclusion of puerarin is of interest. For daidzin and daidzein, further penetration into another monomer's cavity is indeed essential for increased stability. Stick-like template molecules are therefore recommended to induce cooperative effects of CD cavities and to offer a stronger binding force for dimerization of CD monomers. The calculated PMFs detect local minima where the glucose unit locates in the COM region of the dimer. Hydrogen bonds (HB) may induce these favorable minima and hence favor the binding to some extent (Figure 7). Hydroxyl groups of adjacent CDs face each other in this region, yielding a somewhat hydrophilic environment. Thus, a hydrophobic stick-like template with a central hydrophilic moiety is expected to give enhanced binding to CD dimers.

For such isoflavone binding, head-to-head (HH) dimer outperforms the other two orientations of head-to-tail (HT) and tail-to-tail (TT). For HH, the two wide rims of *β*-CDs associate together face to face, thereby maximizing the hydrophobic cavity and allowing efficient encapsulation of template molecules. Thus, BHHP gives the strongest binding. BHTP and BTTS modes disfavor the binding to some extent because the guests encounter two narrow and hydrophilic rims of *β*-CDs (Figure 1d) when forming an inclusion complex and therefore a barrier to entry which means lower k_{on} in case binding rates are of importance. Moreover, the wide rim of CDs makes the template binding easier than the narrow rim and template stabilization of CD dimerization using two wide rims is more thermodynamically favorable. The head-to-head packing therefore appears as a better model for building blocks of CD-based nanostructured materials, in line with Pineiro's report.[27]

The free energy of head-to-head dimerization of *β*-CDs in water was calculated to be −12 kJ/mol using the same force field as in this work;[40] this value is very close to Lopez's work of −14 kJ/mol (although the force fields used were different).[39] Considering the thermodynamic cycle for the binding reaction in Scheme 1, we calculated the binding free energy for each step (Table 3). An obvious cooperative effect for binding of daidzin and daidzein is observed (negative values for $\Delta G_4 - \Delta G_1 = \Delta G_2 - \Delta G_3$ in Table 3). To assess whether the guest (G) is indeed a template for the preferred formation of the CD dimer, one can compare the situation of 2 [CD:G] complexes versus [CD_2:G] complex + G. As shown in the last column of Table 3, we come to a conclusion that only daidzein has a templating effect ($\Delta G_1 + \Delta G_2 - 2\Delta G_3 = -5$), although an enhanced binding strength was observed for all the three guests in some cases.

Scheme 1. Thermodynamic Cycle for CD Dimer Binding Reactions

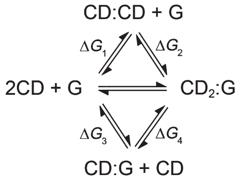


Table 3. Binding Free Energy (kJ/mol) for the Thermodynamic Cycle in Scheme 1

	ΔG_1[a]	ΔG_2[b]	ΔG_3[c]	ΔG_4	$\Delta G_4 - \Delta G_1$[d]	$\Delta G_1 + \Delta G_2 - 2\Delta G_3$[e]
puerarin	−12	−30	−32	−10	2	22
daidzin		−38	−28	−22	−10	6
daidzein		−35	−21	−26	−14	−5

[a] HH dimerization. [b] BHHP binding. [c] 1:1 binding. [d] Cooperative effect. [e] Templating effect.

Template stabilization of CD assemblies can be quantified by the structure-based calculation of binding, which reveals the thermodynamic foundation of the cooperative effects induced by adjacent CD cavities. Upon host−guest complexation CD and template molecules adjust their configurations to minimize the global free energy, leading to fluctuations in atomic positions. For the binding partner, the bonded and torsion energy terms disfavor the complexation, whereas nonbonded interactions tend to favor the binding. The movements of both CD and template molecules are restricted when associated together, leading to entropy loss. For both 1:1 and 2:1 cases, the simulation captured restricted rotations of glucose unit of the guest inside the CD cavity and showed that entropy contributions from the change in flexibility of the molecules in the binding are of crucial importance for proper prediction of free energy differences. Entropic effects are ubiquitous in molecular assembly; however, reliable estimation of entropy for complex systems remains a challenge.[74] In order to establish whether there is any correlation between the free energy and entropy changes, we plot ΔG versus $-T\Delta S_{host}$, $-T\Delta S_{guest}$, and $-T\Delta S_{sol}$ as well as $-T\Delta S_{guest}$ versus $-T\Delta S_{host}$ for [*β*-CD_2:daidzin] in the BHHP mode, as shown in Figures S3a−d, respectively. It is found that the entropy contribution is not correlated to the binding free energy (R^2 = 0.2). The weak correlation of $-T\Delta S_{\text{host}}$ with $-T\Delta S_{\text{guest}}$ (R^2 = 0.4) seems to agree with the finding that entropy loss of guest molecules is accompanied by entropy loss of host molecules, and vice versa (Figure 6).

Desolvation of the binding partner upon binding occurs as well and induces an enthalpy loss (positive $\Delta H_{\text{host-sol}}$ and $\Delta H_{\text{guest-sol}}$). Water molecules that are entrapped inside CD cavity or participate in host and guest solvation are released to the bulk media, yielding a favorable $\Delta H_{\text{sol-sol}}$. The liberation of solvent molecules also allows a greater degree of freedom for water movements and hence an increased ΔS_{sol}. Both these findings are in agreement with the common principle of hydrophobic effect.[75,76] By comparison of the monomer with the dimer, the enthalpic contributions resulting from structural changes and desolvation of host molecules and the entropy contributions from the binding partners and the solvation environment constitute the crucial factors that affect cooperative binding of β-CD dimers to the template molecules, as indicated by the higher increments in Table 2. These factors need to be considered carefully in the design of CD-based supramolecular assemblies based upon cooperative binding.

Thermodynamic analysis on the tested templates shows that the most stable binding states with β-CD dimers do not always correspond to the states where all (un)favorable energy terms achieve their maxima or minima, which confirms our previous conclusion that calculations neglecting flexibility of the binding partners and/or employing implicit solvent will not be able to predict the thermodynamics of complex binding systematically.[40,55] This finding highlights the complexity in molecular assembly and disassembly of CDs in general and is conducive to the regulation of CD-involved aggregates by e.g. template molecules. An in-depth thermodynamic analysis of the binding process described in the present study sets a theoretical foundation for cooperative binding in building blocks of CD-based nanoarchitectures. Such calculations can readily be applied in the design and construction of nanostructures with cooperatively bound units.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Details on the thermodynamic analysis, convergence of entropy calculations, and correlation between ΔG and ΔS. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail twtan@mail.buct.edu.cn; Tel +86 10 64416691 (T.T.).
*E-mail david.vanderspoel@icm.uu.se; Tel +46 18 471 4205 (D.v.d.S.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A grant of computer time by the Swedish research council (SNIC025-12-15) through HPC2N and by Chemcloudcomputing of BUCT is acknowledged. H.Z. thanks the China scholarship council for support. We thank Joppe van der Spoel for technical assistance. This work was supported by the National Basic Research Program of China (973 program: 2013CB733600 and 2011CB710800), the National Nature Science Foundation of China (21106005, 21390202, and 21306006), and National High-Tech R&D Program of China (863 program: 2014AA022101 and 2014AA021904).

■ REFERENCES

(1) Chen, Y.; Zhang, Y.-M.; Liu, Y. Multidimensional Nanoarchitectures Based on Cyclodextrins. *Chem. Commun.* **2010**, *46*, 5622−5633.

(2) Harada, A.; Takashima, Y.; Yamaguchi, H. Cyclodextrin-Based Supramolecular Polymers. *Chem. Soc. Rev.* **2009**, *38*, 875−882.

(3) Lipkowitz, K. B. Applications of Computational Chemistry to the Study of Cyclodextrins. *Chem. Rev.* **1998**, *98*, 1829−1873.

(4) Szejtli, J. Introduction and General Overview of Cyclodextrin Chemistry. *Chem. Rev.* **1998**, *98*, 1743−1753.

(5) van de Manakker, F.; Vermonden, T.; van Nostrum, C. F.; Hennink, W. E. Cyclodextrin-Based Polymeric Materials: Synthesis, Properties, and Pharmaceutical/Biomedical Applications. *Biomacromolecules* **2009**, *10*, 3157−3175.

(6) Jing, J.; Szarpak-Jankowska, A.; Guillot, R.; Pignot-Paintrand, I.; Picart, C.; Auzély-Velty, R. Cyclodextrin/Paclitaxel Complex in Biodegradable Capsules for Breast Cancer Treatment. *Chem. Mater.* **2013**, *25*, 3867−3873.

(7) Gourevich, D.; Dogadkin, O.; Volovick, A.; Wang, L.; Gnaim, J.; Cochran, S.; Melzer, A. Ultrasound-Mediated Targeted Drug Delivery with a Novel Cyclodextrin-Based Drug Carrier by Mechanical and Thermal Mechanisms. *J. Controlled Release* **2013**, *170*, 316−324.

(8) Harada, A. Cyclodextrin-Based Molecular Machines. *Acc. Chem. Res.* **2001**, *34*, 456−464.

(9) Zhao, Y.-L.; Dichtel, W. R.; Trabolsi, A.; Saha, S.; Aprahamian, I.; Stoddart, J. F. A Redox-Switchable α-Cyclodextrin-Based [2]Rotaxane. *J. Am. Chem. Soc.* **2008**, *130*, 11294−11296.

(10) Zhang, Y.-M.; Han, M.; Chen, H.-Z.; Zhang, Y.; Liu, Y. Reversible Molecular Switch of Acridine Red by Triarylpyridine-Modified Cyclodextrin. *Org. Lett.* **2012**, *15*, 124−127.

(11) Chen, Y.; Liu, Y. Cyclodextrin-Based Bioactive Supramolecular Assemblies. *Chem. Soc. Rev.* **2010**, *39*, 495−505.

(12) Kakuta, T.; Takashima, Y.; Harada, A. Highly Elastic Supramolecular Hydrogels Using Host−Guest Inclusion Complexes with Cyclodextrins. *Macromolecules* **2013**, *46*, 4575−4579.

(13) Kettel, M. J.; Hildebrandt, H.; Schaefer, K.; Moeller, M.; Groll, J. Tenside-Free Preparation of Nanogels with High Functional β-Cyclodextrin Content. *ACS Nano* **2012**, *6*, 8087−8093.

(14) Li, J.; Loh, X. J. Cyclodextrin-Based Supramolecular Architectures: Syntheses, Structures, and Applications for Drug and Gene Delivery. *Adv. Drug Delivery Rev.* **2008**, *60*, 1000−1017.

(15) Zhang, J.; Ma, P. X. Cyclodextrin-Based Supramolecular Systems for Drug Delivery: Recent Progress and Future Perspective. *Adv. Drug Delivery Rev.* **2013**, *65*, 1215−1233.

(16) Lai, W.-F. Cyclodextrins in Non-Viral Gene Delivery. *Biomaterials* **2014**, *35*, 401−411.

(17) Dong, Z.; Luo, Q.; Liu, J. Artificial Enzymes Based on Supramolecular Scaffolds. *Chem. Soc. Rev.* **2012**, *41*, 7890−7908.

(18) Nepogodiev, S. A.; Stoddart, J. F. Cyclodextrin-Based Catenanes and Rotaxanes. *Chem. Rev.* **1998**, *98*, 1959−1976.

(19) Liu, Y.; Yang, Y.-W.; Chen, Y.; Zou, H.-X. Polyrotaxane with Cyclodextrins as Stoppers and Its Assembly Behavior. *Macromolecules* **2005**, *38*, 5838−5840.

(20) Harada, A.; Li, J.; Kamachi, M. Synthesis of a Tubular Polymer from Threaded Cyclodextrins. *Nature* **1993**, *364*, 516−518.

(21) Harada, A.; Li, J.; Kamachi, M.; Kitagawa, Y.; Katsube, Y. Structures of Polyrotaxane Models. *Carbohydr. Res.* **1997**, *305*, 127−129.

(22) Udachin, K. A.; Wilson, L. D.; Ripmeester, J. A. Solid Polyrotaxanes of Polyethylene Glycol and Cyclodextrins: The Single Crystal X-Ray Structure of PEG−β-Cyclodextrin. *J. Am. Chem. Soc.* **2000**, *122*, 12375−12376.

(23) Kamitori, S.; Matsuzaka, O.; Kondo, S.; Muraoka, S.; Okuyama, K.; Noguchi, K.; Okada, M.; Harada, A. A Novel Pseudo-Polyrotaxane Structure Composed of Cyclodextrins and a Straight-Chain Polymer: Crystal Structures of Inclusion Complexes of β-Cyclodextrin with

Poly(Trimethylene Oxide) and Poly(Propylene Glycol). *Macromolecules* **2000**, *33*, 1500−1502.

(24) Chatziefthimiou, S. D.; Yannakopoulou, K.; Mavridis, I. M. β-Cyclodextrin Trimers Enclosing an Unusual Organization of Guest: The Inclusion Complex β-Cyclodextrin/4-Pyridinealdazine. *CrystEngComm* **2007**, *9*, 976−979.

(25) Saenger, W.; Jacob, J.; Gessler, K.; Steiner, T.; Hoffmann, D.; Sanbe, H.; Koizumi, K.; Smith, S. M.; Takaha, T. Structures of the Common Cyclodextrins and Their Larger Analogues Beyond the Doughnut. *Chem. Rev.* **1998**, *98*, 1787−1802.

(26) Jana, M.; Bandyopadhyay, S. Hydration Properties of α-, β-, and γ-Cyclodextrins from Molecular Dynamics Simulations. *J. Phys. Chem. B* **2011**, *115*, 6347−6357.

(27) Brocos, P.; Díaz-Vergara, N.; Banquy, X.; Pérez-Casas, S.; Costas, M.; Piñeiro, A. n. Similarities and Differences between Cyclodextrin−Sodium Dodecyl Sulfate Host−Guest Complexes of Different Stoichiometries: Molecular Dynamics Simulations at Several Temperatures. *J. Phys. Chem. B* **2010**, *114*, 12455−12467.

(28) Naidoo, K. J.; Chen, J. Y. J.; Jansson, J. L. M.; Widmalm, G.; Maliniak, A. Molecular Properties Related to the Anomalous Solubility of β-Cyclodextrin. *J. Phys. Chem. B* **2004**, *108*, 4236−4238.

(29) Cai, W.; Sun, T.; Shao, X.; Chipot, C. Can the Anomalous Aqueous Solubility of Beta-Cyclodextrin Be Explained by Its Hydration Free Energy Alone? *Phys. Chem. Chem. Phys.* **2008**, *10*, 3236−43.

(30) Zhang, H.; Feng, W.; Li, C.; Tan, T. Investigation of the Inclusions of Puerarin and Daidzin with β-Cyclodextrin by Molecular Dynamics Simulation. *J. Phys. Chem. B* **2010**, *114*, 4876−4883.

(31) Anconi, C. P. A.; Nascimento, C. S.; De Almeida, W. B.; Dos Santos, H. F. Structure and Stability of $(\alpha\text{-CD})_3$ Aggregate and OEG@$(\alpha\text{-CD})_3$ Pseudorotaxane in Aqueous Solution: A Molecular Dynamics Study. *J. Phys. Chem. B* **2009**, *113*, 9762−9769.

(32) Nascimento, C. S.; Anconi, C. P. A.; Dos Santos, H. F.; De Almeida, W. B. Theoretical Study of the α-Cyclodextrin Dimer. *J. Phys. Chem. A* **2005**, *109*, 3209−3219.

(33) Pozuelo, J.; Mendicuti, F.; Mattice, W. L. Inclusion Complexes of Chain Molecules with Cycloamyloses. 2. Molecular Dynamics Simulations of Polyrotaxanes Formed by Poly(Ethylene Glycol) and α-Cyclodextrins. *Macromolecules* **1997**, *30*, 3685−3690.

(34) Bonnet, P.; Jaime, C.; Morin-Allory, L. α-, β-, and γ-Cyclodextrin Dimers. Molecular Modeling Studies by Molecular Mechanics and Molecular Dynamics Simulations. *J. Org. Chem.* **2001**, *66*, 689−692.

(35) Bonnet, P.; Jaime, C.; Morin-Allory, L. Structure and Thermodynamics of α-, β-, and γ-Cyclodextrin Dimers. Molecular Dynamics Studies of the Solvent Effect and Free Binding Energies. *J. Org. Chem.* **2002**, *67*, 8602−8609.

(36) Tallury, S. S.; Smyth, M. B.; Cakmak, E.; Pasquinelli, M. A. Molecular Dynamics Simulations of Interactions between Polyanilines in Their Inclusion Complexes with β-Cyclodextrins. *J. Phys. Chem. B* **2012**, *116*, 2023−2030.

(37) Liu, P.; Chipot, C.; Shao, X.; Cai, W. How Do α-Cyclodextrins Self-Organize on a Polymer Chain? *J. Phys. Chem. C* **2012**, *116*, 17913−17918.

(38) Lopez, C. A.; de Vries, A. H.; Marrink, S. J. Computational Microscopy of Cyclodextrin Mediated Cholesterol Extraction from Lipid Model Membranes. *Sci. Rep.* **2013**, *3*, 2071.

(39) López, C. A.; de Vries, A. H.; Marrink, S. J. Molecular Mechanism of Cyclodextrin Mediated Cholesterol Extraction. *PLoS Comput. Biol.* **2011**, *7*, e1002020.

(40) Zhang, H.; Tan, T.; Feng, W.; van der Spoel, D. Molecular Recognition in Different Environments: β-Cyclodextrin Dimer Formation in Organic Solvents. *J. Phys. Chem. B* **2012**, *116*, 12684−12693.

(41) Keung, W. M.; Vallee, B. L. Kudzu Root: An Ancient Chinese Source of Modern Antidipsotropic Agents. *Phytochemistry* **1998**, *47*, 499−506.

(42) Lowe, E. D.; Gao, G.-Y.; Johnson, L. N.; Keung, W. M. Structure of Daidzin, a Naturally Occurring Anti-Alcohol-Addiction Agent, in Complex with Human Mitochondrial Aldehyde Dehydrogenase. *J. Med. Chem.* **2008**, *51*, 4482−4487.

(43) Sun, T.; Shao, X.; Cai, W. Self-Assembly Behavior of β-Cyclodextrin and Imipramine. A Free Energy Perturbation Study. *Chem. Phys.* **2010**, *371*, 84−90.

(44) Cai, W.; Sun, T.; Liu, P.; Chipot, C.; Shao, X. Inclusion Mechanism of Steroid Drugs into β-Cyclodextrins. Insights from Free Energy Calculations. *J. Phys. Chem. B* **2009**, *113*, 7836−7843.

(45) Filippini, G.; Goujon, F.; Bonal, C.; Malfreyt, P. Energetic Competition Effects on Thermodynamic Properties of Association between β-CD and Fc Group: A Potential of Mean Force Approach. *J. Phys. Chem. C* **2012**, *116*, 22350−22358.

(46) Zhang, Q.; Tu, Y.; Tian, H.; Zhao, Y.-L.; Stoddart, J. F.; Ågren, H. Working Mechanism for a Redox Switchable Molecular Machine Based on Cyclodextrin: A Free Energy Profile Approach. *J. Phys. Chem. B* **2010**, *114*, 6561−6566.

(47) Wallace, S. J.; Kee, T. W.; Huang, D. M. Molecular Basis of Binding and Stability of Curcumin in Diamide-Linked γ-Cyclodextrin Dimers. *J. Phys. Chem. B* **2013**, *117*, 12375−12382.

(48) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(49) Lemkul, J. A.; Bevan, D. R. Assessing the Stability of Alzheimer's Amyloid Protofibrils Using Molecular Dynamics. *J. Phys. Chem. B* **2010**, *114*, 1652−1660.

(50) Rashid, M. H.; Kuyucak, S. Affinity and Selectivity of Shk Toxin for the Kv1 Potassium Channels from Free Energy Simulations. *J. Phys. Chem. B* **2012**, *116*, 4812−4822.

(51) Hub, J. S.; Winkler, F. K.; Merrick, M.; de Groot, B. L. Potentials of Mean Force and Permeabilities for Carbon Dioxide, Ammonia, and Water Flux across a Rhesus Protein Channel and Lipid Membranes. *J. Am. Chem. Soc.* **2010**, *132*, 13251−13263.

(52) Wennberg, C. L.; van der Spoel, D.; Hub, J. S. Large Influence of Cholesterol on Solute Partitioning into Lipid Membranes. *J. Am. Chem. Soc.* **2012**, *134*, 5351−5361.

(53) Caleman, C.; Hub, J. S.; van Maaren, P. J.; van der Spoel, D. Atomistic Simulation of Ion Solvation in Water Explains Surface Preference of Halides. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6838−6842.

(54) Hub, J. S.; Caleman, C.; van der Spoel, D. Organic Molecules on the Surface of Water Droplets - an Energetic Perspective. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9537−9545.

(55) Zhang, H.; Tan, T.; Hetényi, C.; van der Spoel, D. Quantification of Solvent Contribution to the Stability of Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 4542−4551.

(56) Cezard, C.; Trivelli, X.; Aubry, F.; Djedaini-Pilard, F.; Dupradeau, F. Y. Molecular Dynamics Studies of Native and Substituted Cyclodextrins in Different Media: 1. Charge Derivation and Force Field Performances. *Phys. Chem. Chem. Phys.* **2011**, *13*, 15103−15121.

(57) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(59) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(60) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(61) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. Gromacs 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(62) Zhang, H.; Ge, C.; van der Spoel, D.; Feng, W.; Tan, T. Insight into the Structural Deformations of Beta-Cyclodextrin Caused by

Alcohol Cosolvents and Guest Molecules. *J. Phys. Chem. B* **2012**, *116*, 3880−3889.

(63) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289−6292.

(64) Starr, F. W.; Nielsen, J. K.; Stanley, H. E. Hydrogen-Bond Dynamics for the Extended Simple Point-Charge Model of Water. *Phys. Rev. E* **2000**, *62*, 579−587.

(65) Jain, V.; Maingi, V.; Maiti, P. K.; Bharatam, P. V. Molecular Dynamics Simulations of PPI Dendrimer-Drug Complexes. *Soft Matter* **2013**, *9*, 6482−6496.

(66) van der Spoel, D.; van Maaren, P.; Larsson, P.; Timneanu, N. Thermodynamics of Hydrogen Bonding in Hydrophilic and Hydrophobic Media. *J. Phys. Chem. B* **2006**, *110*, 4393−4398.

(67) Li, G.; McGown, L. B. Molecular Nanotube Aggregates of β- and γ-Cyclodextrins Linked by Diphenylhexatrienes. *Science* **1994**, *264*, 249−251.

(68) Miyake, K.; Yasuda, S.; Harada, A.; Sumaoka, J.; Komiyama, M.; Shigekawa, H. Formation Process of Cyclodextrin Necklace-Analysis of Hydrogen Bonding on a Molecular Level. *J. Am. Chem. Soc.* **2003**, *125*, 5080−5085.

(69) Wu, J.; He, H.; Gao, C. β-Cyclodextrin-Capped Polyrotaxanes: One-Pot Facile Synthesis Via Click Chemistry and Use as Templates for Platinum Nanowires. *Macromolecules* **2010**, *43*, 2252−2260.

(70) Dandawate, P.; Vyas, A.; Ahmad, A.; Banerjee, S.; Deshpande, J.; Swamy, K. V.; Jamadar, A.; Dumhe-Klaire, A.; Padhye, S.; Sarkar, F. Inclusion Complex of Novel Curcumin Analogue CDF and β-Cyclodextrin (1:2) and Its Enhanced in Vivo Anticancer Activity against Pancreatic Cancer. *Pharm. Res.* **2012**, *29*, 1775−1786.

(71) Harada, T.; Pham, D.-T.; Leung, M. H. M.; Ngo, H. T.; Lincoln, S. F.; Easton, C. J.; Kee, T. W. Cooperative Binding and Stabilization of the Medicinal Pigment Curcumin by Diamide Linked γ-Cyclodextrin Dimers: A Spectroscopic Characterization. *J. Phys. Chem. B* **2011**, *115*, 1268−1274.

(72) Liu, Y.; Chen, Y. Cooperative Binding and Multiple Recognition by Bridged Bis(β-Cyclodextrin)s with Functional Linkers. *Acc. Chem. Res.* **2006**, *39*, 681−691.

(73) Liu, Y.; Song, Y.; Chen, Y.; Yang, Z. X.; Ding, F. Spectrophotometric Study on the Controlling Factor of Molecular Selective Binding of Dyes by Bridged Bis(β-Cyclodextrin)s with Diselenobis(Benzoyl) Linkers. *J. Phys. Chem. B* **2005**, *109*, 10717−10726.

(74) Baron, R.; van Gunsteren, W. F.; Hünenberger, P. H. Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation. *Trends Phys. Chem.* **2006**, *11*, 87−122.

(75) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2001**, *106*, 521−533.

(76) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640−647.

D26

International Journal of
Molecular Sciences

MDPI

Article

A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level

István Horváth [1], Norbert Jeszenői [2], Mónika Bálint [3], Gábor Paragi [4,5] and Csaba Hetényi [3,*]

[1] Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720 Szeged, Hungary
[2] Institute of Physiology, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary
[3] Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary
[4] MTA-SZTE Biomimetic Systems Research Group, Dóm tér 8, 6720 Szeged, Hungary
[5] Institute of Physics, University of Pécs, Ifjúság útja 6, 7624 Pécs, Hungary
* Correspondence: hetenyi.csaba@pte.hu

Received: 5 August 2019; Accepted: 4 September 2019; Published: 6 September 2019

Abstract: Optimization of the enthalpy component of binding thermodynamics of drug candidates is a successful pathway of rational molecular design. However, the large size and missing hydration structure of target-ligand complexes often hinder such optimizations with quantum mechanical (QM) methods. At the same time, QM calculations are often necessitated for proper handling of electronic effects. To overcome the above problems, and help the QM design of new drugs, a protocol is introduced for atomic level determination of hydration structure and extraction of structures of target-ligand complex interfaces. The protocol is a combination of a previously published program MobyWat, an engine for assigning explicit water positions, and Fragmenter, a new tool for optimal fragmentation of protein targets. The protocol fostered a series of fast calculations of ligand binding enthalpies at the semi-empirical QM level. Ligands of diverse chemistry ranging from small aromatic compounds up to a large peptide helix of a molecular weight of 3000 targeting a leukemia protein were selected for systematic investigations. Comparison of various combinations of implicit and explicit water models demonstrated that the presence of accurately predicted explicit water molecules in the complex interface considerably improved the agreement with experimental results. A single scaling factor was derived for conversion of QM reaction heats into binding enthalpy values. The factor links molecular structure with binding thermodynamics via QM calculations. The new protocol and scaling factor will help automated optimization of binding enthalpy in future molecular design projects.

Keywords: peptide; interaction; design; affinity; optimization; binding; water; structure; correlation

1. Introduction

Determination of structure and binding thermodynamics of target-ligand complexes is a key step in drug design [1]. Thermodynamic quantities can be measured by experimental methods such as isothermal titration calorimetry (ITC [2–11]). Experimental measurements are often restricted by the lack and high cost of pure and concentrated target (protein) samples. Molecular structures and binding thermodynamics can be also predicted [12–16] by fast and cheap molecular mechanics methods. At the same time, molecular mechanics has serious limitations of calculation of electronic effects in complex structures. Such effects are present in almost all intermolecular interactions including ‘exotic’ cases such as cation-π interactions between aromatic and charged side-chains [4,17] or polarization effects at structural water molecules [18]. Quantum mechanical (QM) approaches can properly handle electronic effects of intermolecular interactions. However, hydration and large size of target-ligand complexes impose further challenges on QM methods as detailed in the following paragraphs.

Hydration largely affects the structure and function of various biomolecules and their complexes [19,20]. Water molecules of the complex interface contribute to the stability and specificity of target-ligand interactions [21–28] by building hydrogen bonding networks [29,30], restraining interatomic distances, and filling cavities [19,31]. Despite their importance, determination of positions of interfacial water molecules is not trivial [32]. Available water positions have been determined mostly [33] by X-ray crystallography. However, even this well-established technique suffers from numerous limitations. Assignation of electron density peaks to possible interface water positions is still not a routine job due to inherent mobility of water and large number of degrees of freedom [34] and the quality of the structure depends on the solute size [35]. Protein hydration in the crystal is not the same as in solution [36] which is further complicated by cryo-artefacts [36]. Overfitting of electron density data and misleading identification of water sites were found to be a bad practice [25]. Other experimental techniques such as nuclear magnetic resonance spectroscopy or cryo-electron microscopy have produced a relatively small number of structures with water positions assigned. To overcome the above limitations of experimental methods, theoretical approaches were developed to help the assignation of water positions. These approaches either assign water positions based solely on solute structures [37] or involve calculation of dynamics [38–45] of water–water interactions. In the present study, a molecular dynamics-based method MobyWat [32,46] will be applied for completion of hydration structures of target-ligand interfaces.

Besides hydration, system size is another challenge of calculation of large complexes at the QM level. Such investigations would require large computer resources if the entire target molecule was calculated. A decomposition of the target-ligand complex into tractable sub-systems can handle this problem. There are at least two approaches to conduct such a decomposition. The first approach applies QM for the binding site and molecular mechanics simulations to the rest of the system [47–52]. Another branch of methods is based on skillful fragmentation of the target and applies QM for the sub-system of target fragments and the ligand. For example, Zhang and Zhang [53] developed a method for molecular fractionation where the protein is decomposed into individual capped fragments. They performed ab initio HF and DFT QM calculations for the target-ligand complexes. Nikitina et al. [54,55] cut the heavy atoms of the target at a distance equal or less than 5 Å from any heavy atom of the ligand. They also used structural water molecules determined by X-ray analysis, inserted new ones according to H-bonding valences of the solute molecules [54] and also proposed an iterative scheme [55] of in silico hydration. They developed correlations for binding enthalpy (ΔH_b) on sets of 8 [54], and 12 [55] complexes, respectively. The complexes included protein targets with small ligands of molecular weight (MW) up to 700 and the calculations were conducted at semi-empirical QM level using the PM3 parametrization. Dobes, Hobza et al. [56] investigated the small-molecule purine inhibitor Roscovitine in complex with cyclin-dependent kinase 2 at B3LYP/6–31G** and MP2 levels of theory. They cut the chains of the kinase target into small fragments of a few amino acids at the C_α-N bond. The peptide bond was maintained and they considered only amino acids and crystal water molecules located within 5 Å from the ligand.

Structure-based calculation of thermodynamic properties such as ΔH_b is a central issue of engineering of efficient drug candidates. Enthalpic optimization of new lead molecules [57–59] is a successful pathway of drug design and requires determination or prediction of ΔH_b of target-ligand complexes. Despite the need for ΔH_b data, there are only a few QM studies on fragment-based calculation of target-ligand binding thermodynamics. Available studies of the previous paragraph mostly work with ligand molecules of moderate size. Complexes of large (peptidic) ligands with numerous hydration sites have not been studied extensively. Moreover, development of automated tools for extraction of structures of complex interfaces and a reliable hydration scheme would be also helpful for such fragment-based QM investigations.

A new protocol was introduced and tested in the present study to help the enthalpic design of drug candidates by answering the above challenges of automation of structure-based calculation of complexes of large ligands. For this purpose, an end-point approach was adopted for the calculation of ΔH_b according to Equations (1) and (2). As the reaction occurs in a biological environment, T and L water molecules hydrate the target and the ligand, respectively. Waters can also remain bound to the partners ($s = 0$), join the complex from the surrounding bulk ($s > 0$) or leave ($s < 0$) during ligand binding. The reaction heat ($\Delta_r H$) of the binding process of Equation (1) can be calculated [14,15,54,55,60–62] according to Hess's law (Equation (2)), where $\Delta_f H$ represents the calculated heat of formation of a reactant or a product as indicated in brackets.

$$\mathrm{Target}[\mathrm{H_2O}]_\mathrm{T} + \mathrm{Ligand}[\mathrm{H_2O}]_\mathrm{L} + \mathrm{s}\ \mathrm{H_2O} = \mathrm{Target{:}Ligand}[\mathrm{H_2O}]_{\mathrm{T+L+s}} \quad (1)$$

$$\Delta_r H = \Delta_f H(\mathrm{Target{:}Ligand}[\mathrm{H_2O}]_\mathrm{s}) - \Delta_f H(\mathrm{Target}) - \Delta_f H(\mathrm{Ligand}) - \mathrm{s}\Delta_f H(\mathrm{H_2O}) \quad (2)$$

This end-point approach is simple and it has been successfully applied in previous publications [14,15,54,55,60–62]. In the present study, it was particularly useful for screening of various solvent models and conducting several trials in reasonable time. In the forthcoming sections, the fine-tuning of the corresponding protocol, and the development of a relationship between calculated reaction heats and experimental binding enthalpy values will be described.

2. Results and Discussion

2.1. Fragmenter

As it was discussed in the Introduction, involving the entire target structure in a QM calculation is not feasible within a reasonable calculation time. Thus, QM calculation of the above $\Delta_f H$ values (Equation (2)) necessitates an extraction of the interface region of the target-ligand complex. However, extraction of the complex interface and automated fragmenting of the target protein has no trivial solution. In the present study, a new protocol was elaborated including a fragmentation method, Fragmenter, to standardize the extraction of target-ligand interfaces (Figure 1). Fragmenter works on a complex structure including a target, a ligand and several water molecules. Amino acids of fragments are selected according to their intermolecular distance cut-off (d_{TL}, Table 1). A brief overview of Fragmenter and the data stream are sketched in Figures S1 and S2 and technical details are provided in Methods.

Fragmenter focuses on the neighboring parts of the target protein which have considerable interactions with the ligand and the interfacial water molecules. The whole ligand molecule and protein residues of interface regions of the complexes are extracted. The residues of the target molecule are preferably extracted as peptide fragments instead of single amino acids. The main goal is to obtain the shortest but continuous peptide chains from the target protein in a standardized way.

Thus, there is still a benefit of a considerably reduced target part, and continuity is also kept wherever it is possible. Parameter n specifies how many adjacent amino acids are added to the fragment chain of amino acids extracted according to d_{TL}. After some experimenting (Table S2), it was found that $n = 0$ produces good correlations (as seen in the following sections), and it was not necessary to investigate $n = 1$ for the systems of the present study. Fragmenter was implemented as a free web service (Figure S4). It provides the extracted complex interface structure (target fragments, ligand and water molecules) as an interactive image, also downloadable as PDB and Mopac input files from the 'results' tab (Figure S5) and also displays a list of estimates of per-residue intermolecular interaction energy (E_{inter}) values to indicate unwanted close contacts.

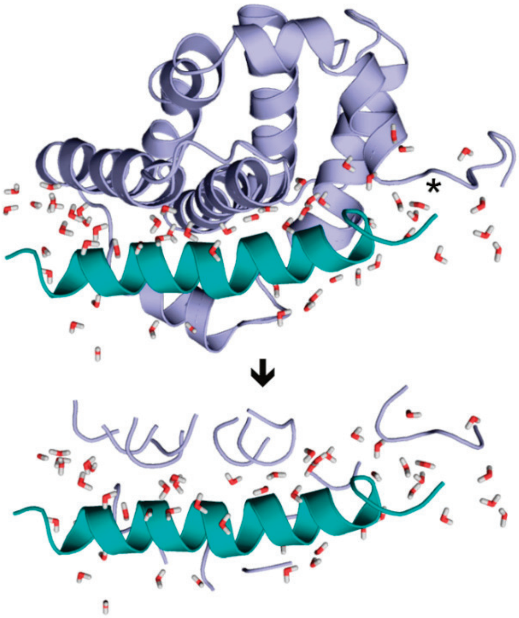


Figure 1. Fragmenter extracts a hydrated interface (bottom) from the target-ligand complex (top). Target (fragments) and ligand are shown in light blue, and green, respectively. System 2roc contains the largest ligand investigated in the present study. In this example, Fragmenter extracted target residues with (d_{TL} = 5.0 Å) considerably reducing the system size used for QM calculation. Interfacial water molecules (d_W = 5.0 Å, sticks) are also retained. Steps of extraction of target fragments are shown in atomic details for the C-terminal region (asterisk) in Figure S3 as an example. Fragmenter is available free of charge as a web service at www.fragmenter.xyz.

2.2. Dry Systems and an Implicit Water Model

Having the Fragmenter protocol developed and implemented, $\Delta_f H$ calculations of the (hydrated) target-ligand complex interfaces were conducted in a simplified and standardized way. Fragmenter was applied on all systems of Table 1 for extraction of the complex interfaces. All systems were prepared for Fragmenter using standard molecular mechanics energy minimization and explicit hydration protocols as described in Methods. The $\Delta_f H$ values were calculated for the individual reactant (ligand and target fragments) and product (complex interface) structures, respectively. The calculations were performed at semi-empirical level using PM7 parameterization, with and without the Mozyme approach (Methods). The resulted, raw energy values are listed in Table S4.

Within the end-point approach (Introduction), calculation of $\Delta_f H$ of the reaction participants (Equation (2)) and a linear scaling (Equation (3)) of $\Delta_r H$ to known experimental ΔH_b(exp) values is necessary for calculation of ΔH_b.

$$\Delta H_b(\text{exp})_i = \alpha \Delta_r H_i + \beta + \varepsilon_i = \Delta H_b(\text{calc})_i + \varepsilon_i, \text{ where } i = 1, 2, \ldots, N \quad (3)$$

In the present study, 15 systems (N = 15) of Table 1 were involved in the derivation of regression coefficients (α, β) yielding ΔH_b(calc) values and residuals (ε). Statistical parameters obtained for the dry complexes and various solvent models are listed in Table 2. Nine of the 15 systems with small ligands up to a MW of 550 were considered in a previous paper [55] as well. In the present study, additional six systems with large peptide ligands were included in the set as they often impose a challenge during lead optimizations due to their size and extensive hydration. Thus, the set of 15 systems involves various ligands with MW up to 3318, two orders of magnitude larger than the previous set. The experimental ΔH_b values cover a wide range between −2.935 and −15.5 kcal/mol (Table 1).

Table 1. Target-ligand systems.

System [a]	Res [b] (Å)	Target	Ligand		Water Count			ΔHb(exp) [d]
			Name	MW [c]	Shell 1	Shell 2	Shell 3	kcal mol^{-1}
3ptb_ben	1.7	beta-trypsin	benzamidine	121.2	1	6	7	−4.507 [2]
3ptb_pme	1.7	beta-trypsin	p-methylbenzamidine	135.2	1	5	6	−4.412 [2]
3ptb_pam	1.7	beta-trypsin	p-aminobenzamidine	136.2	3	4	7	−6.417 [2]
3ptb_pmo	1.7	beta-trypsin	p-methoxybenzamidine	151.2	1	6	7	−3.742 [2]
3ptb_pad	1.7	beta-trypsin	p-amidinobenzamidine	164.2	2	8	10	−2.935 [2]
1k1l	2.5	bovine trypsin	NAPe-piperazine	467.6	5	10	15	−7.863 [4]
1k1m	2.2	bovine trypsin	NAP [e]−4-acetyl-piperazine	508.6	4	12	16	−8.222 [4]
1k1i	2.2	bovine trypsin	NAP [e]-D-pipecolinic acid	508.6	2	13	15	−10.899 [4]
1k1j	2.2	bovine trypsin	NAP [e]-isopipecolinic acid methyl ester	523.6	3	13	16	−9.465 [4]
1jyr	1.55	Grb2 SH2 domain	APS-PTR [e]-VNVQN	1069.0	1	14	15	−7.94 [6]
1rlq	NA	C-src tyrosine kinase SH3 domain	RALPPLPRY	1084.3	2	25	27	−10.2 [7]
2ke1	NA	autoimmune regulator	ARTKQTARKS	1150.3	12	15	27	−9.2 [8]
2bba	1.65	EphB4 receptor	NYLFSPNGPIARAW	1606.8	12	15	27	−15.5 [9]
1jgn	NA	human poly(A)-binding protein	VVKSNLNPNAKEFVPGVKYGNI	2389.8	14	34	48	−14.8 [10]
2roc	NA	induced myeloid leukemia cell differentiation protein homolog	EEEWAREIGAQLRRIADDLNAQYERRM	3317.6	14	38	52	−14.3 [11]

[a] System codes are derived from the PDB identifiers, and abbreviated ligands names (where applicable). [b] Resolution (available for crystallographic structures). [c] Molecular weight. [d] Experimental binding enthalpy values are given at their original level of precision except those with three decimal digits converted from $kJmol^{-1}$, where 1 J = 4.184 cal. Sources of values are indicated as references in superscript. [e] NAP: N-alpha-(2-naphthylsulfonyl)-N-(3-amidino-L-phenylalaninyl); PTR: o-phosphotyrosine.

In the first step of the present investigations, no solvent models were applied (s = 0 in Equations (1) and (2)). That is, dry input structures without explicit water molecules were calculated in vacuo. The complete lack of water models resulted no correlation between the calculated and experimental ΔH_b values (column Vacuum/Dry in Table 2, Figure 2). The application of an implicit water model (COnductor-like Screening MOdel, COSMO [63]) increased the correlation (column COSMO/Dry in Table 2). However, this correlation can still be improved as reflected by the cross-validation. In general, the use of COSMO proved advantageous if compared with the vacuum/dry results (Table 2). There was a single case of System 2ke1 where ΔH_b(exp) could not be converted to 298.15 K and the original value at 296.15 K (Table S3) was used for the regressions of Table 2. To check the influence of this data point on the results, linear regressions were performed without System 2ke1, as well. The statistical parameters showed (Table S5) that leaving out System 2ke1 did not improve the results in vacuo and COSMO yields considerable correlation.

Table 2. Per-system residuals (ε) and statistical parameters of linear regressions obtained with different water models.

System	Vacuum				COSMO				
	Dry	**Shell 1**	**Shell 2**	**Shell 3**	**Dry**	**Shell 1**	**Shell 2**	**Shell 3**	**Shell 3** [b]
					|ε| [a]				
3ptb_ben	3.70	3.18	2.90	1.93	3.73	2.33	2.45	1.18	0.85
3ptb_pme	3.60	3.09	3.05	2.03	0.53	1.12	2.48	0.95	1.22
3ptb_pam	1.74	1.29	0.88	0.03	0.01	0.02	0.44	0.92	3.03
3ptb_pmo	4.45	3.91	3.71	2.64	3.59	3.25	3.32	2.24	0.33
3ptb_pad	5.65	5.11	5.04	4.25	3.03	3.12	4.32	3.22	1.37
1k1l	0.56	0.39	0.10	0.10	2.59	1.05	0.56	0.43	1.74
1k1m	0.16	0.56	0.36	0.79	2.71	1.17	0.75	1.50	3.12
1k1i	2.67	3.19	2.95	3.51	2.80	3.62	2.88	3.34	4.60
1k1j	1.43	1.81	1.60	2.10	0.57	2.60	1.47	2.32	3.75
1jyr	0.78	0.28	0.53	0.09	1.73	0.36	0.40	0.53	0.36
1rlq	0.67	0.64	0.27	0.33	0.37	2.13	0.61	0.24	0.16
2ke1	2.54	4.67	4.66	5.28	4.38	4.21	5.73	2.46	2.89
2bba	7.25	6.38	7.21	6.23	4.34	2.13	6.58	3.77	3.31
1jgn	5.88	5.24	4.75	2.56	1.61	0.27	3.25	0.25	1.36
2roc	4.96	4.11	3.74	1.26	2.76	1.24	3.05	1.71	3.92
R^2	0.06	0.18	0.19	0.44	0.51	0.65	0.33	0.73	0.93
R^2(cv) [c]	0.00	0.01	0.02	0.22	0.34	0.54	0.07	0.65	0.91
F	0.81	2.77	3.14	10.20	13.46	24.28	6.36	34.55	179.66
RMSE [a]	4.02	3.76	3.72	3.10	2.90	2.45	3.40	2.17	2.65
t_α	0.90	1.66	1.77	3.19	3.67	4.93	2.52	5.88	13.40
t_β	−5.56	−5.04	−4.68	−3.90	−2.18	−3.99	−4.24	−2.81	-

[a] Unit: kcalmol^{-1}. [b] Linear regression with β = 0 (last column), and β ≠ 0 (other columns). [c] Leave-one-out cross-validated coefficient of determination.

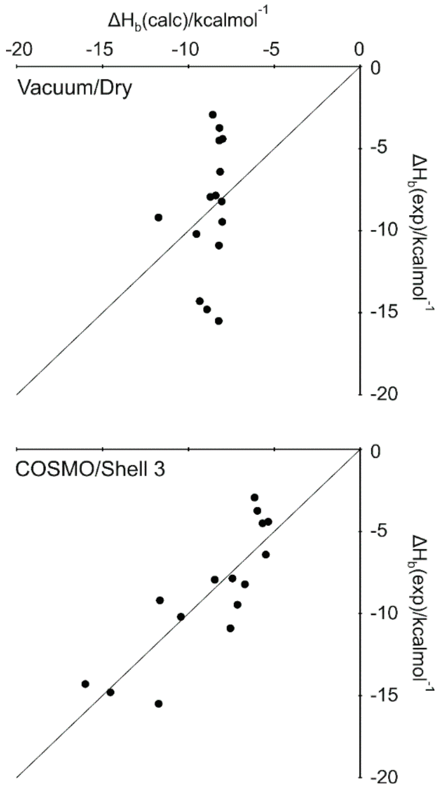


Figure 2. Correlation plots obtained without (Vacuum/Dry) and with (COSMO/Shell3) the hybrid water model.

2.3. Explicit Hydration and a Hybrid Model

A systematic investigation on explicit hydration was conducted to further improve the correlations of the previous section. It is challenging to give a straightforward definition for the origin of water molecules in the complexes, and prediction of ligand-bound water molecules is rather uncertain due to the relatively small binding surface of ligands. Thus, T = L = 0 was set and all interface water molecules were considered as if they had originated from the surrounding bulk solvent ($s > 0$, in Equations (1) and (2)). Hydration structure of the target-ligand complex was built up by the MobyWat method [32] and extracted by Fragmenter as part of the interfaces. MobyWat can produce complete, void-free hydration structures of complex interfaces. This is guaranteed by a soaking step during the systematic evaluation of a series of snapshots of molecular dynamics simulations accounting for water–water interactions besides solute-water ones. Thus, MobyWat can find all experimental reference water positions in many cases [32] and assign water positions not detectable by experimental [25,33,34,64,65] measurements.

In the present study, three shells were defined according to d_w (Table 1 and Table S1) using interfacial water molecules (Figure 3A). Shell 1 contains water molecules closest to the solutes (d_w = 3.5 Å). Shell 2 holds waters with intermediate positions (3.5 Å < d_w < 5.0 Å). Shell 3 consists of all interfacial water molecules of Shells 1 and 2 with a d_w = 5.0 Å.

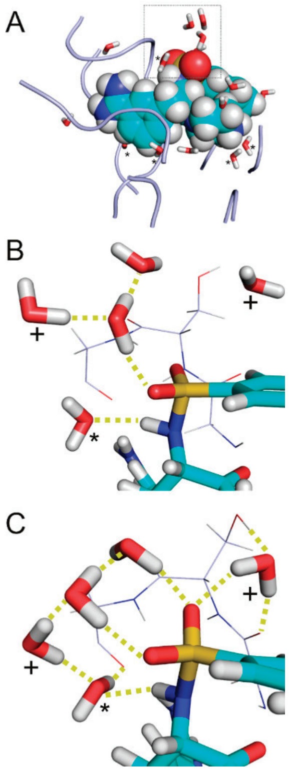


Figure 3. Extracted complex interface of System 1k1l. (**A**) Initial structure equipped with water molecules and energy-minimized at the molecular mechanics level. Target fragments and ligand are shown in ribbon and space filling representations, respectively. Water molecules in Shell 1 (d_W = 3.5 Å, sticks marked with asterisk) are positioned close to the solute partners and play a bridging role. The rest of Shell 3 (d_W = 5.0 Å) waters belong to Shell 2 (sticks without asterisks) and located at the edges of the interface, close to the bulk. Shell 3 = Shell 1 + Shell 2. (**B**) A rotated close-up of the box in Panel A showing the surrounding of the sulphonyl group of the ligand (sticks) and the neighboring residues $G_{216}SG_{218}$ of the target (lines) where the numbering follows that of the crystallographic structure (PDB ID 1k1l). Hydrogen bonds are marked with yellow dotted lines. (**C**) Structure in Panel B after relaxation at semi-empirical level using PM7 parameterization and Mozyme. Water molecules with a displacement above 1.5 Å after relaxation are marked with crosses.

The use of explicit water molecules in vacuum improved the 'dry' correlations to an R^2 of 0.44 (all systems, column Vacuum/Shell 3 in Table 2) and 0.65 (without System 2ke1, Table S5). Cross-validation indicates that this improvement of correlation is robust only without System 2ke1. At this point, it seemed reasonable to check whether a hybrid model using both implicit and explicit hydration further improves the correlation. Indeed, the hybrid model (column COSMO/Shell 3 in Table 2) provided the best agreement between calculated and experimental ΔH_b with an R^2 of 0.73 (Figure 2) using all interfacial water molecules. Notably, Shell 1 waters also yielded considerable correlation (R^2 of 0.65). In both cases, the correlations survived the challenge of cross-validation. The intermediate water positions alone (Shell 2) yielded a stable correlation only without System 2ke1 (Table S5).

To investigate the effect of ligand size and target diversity on the stability of the above correlation (COSMO/Shell 3), the set of systems in Table 1 was split into two sub-sets according to ligand MW. The first sub-set contains nine systems with small ligands of MW < 600. All these ligands have a common target, beta trypsin. The second sub-set contained six systems with large ligands of MW > 1000 and various targets. Linear regressions were performed separately for the two sub-sets and ΔH_b(calc) values were calculated by the two regression equations, respectively. Overall statistical parameters obtained (Table S6) were comparable to those of the regression for all systems (column COSMO/Shell 3 in Table 2) detailed above. Thus, stability of the correlations is not influenced by ligand size and target diversity of the systems in the case of the hybrid model.

2.4. Scaling Factor

The above COSMO/Shell 3 model with $\beta \neq 0$ in Equation (3) is significant and robust regarding its overall regression parameters. However, the t_β value (Table 2) indicates that the level of significance of regression coefficient β is moderate ($p = 0.015$). Thus, a linear regression with $\beta = 0$ was also developed and the corresponding statistical parameters are listed in the last column of Table 2 (Table S7). In this way, a model of high significance ($p < 0.01$) of all parameters was obtained and Equation (3) was simplified. The resulting Equation (4) includes only the value of regression coefficient α, which serves as a single, unit-independent scaling factor for conversion of calculated $\Delta_r H$ into ΔH_b.

$$\Delta H_b = 0.031\ (\pm 0.002)\ \Delta_r H \tag{4}$$

A similar value of 0.032 (±0.002) was obtained for the scaling factor if System 2ke1 was not involved in the regression. Via QM calculations, this factor serves as a direct link between molecular structure and binding thermodynamics of molecular complexes.

2.5. Case Studies on Hydration Structures

In two-thirds of the 15 systems, application of Shell 1 or 3 explicit water molecules resulted in the decrease of residuals (COSMO models in Table 2). Shell 2 waters have similar effect in one-third of the cases. For example, in the case of System 1k1l, the residuals decreased from 2.59 (dry) to 0.43 (Shell 3, $\beta \neq 0$) and 1.74 kcal/mol (Shell 3, $\beta = 0$, Table 2), and a similar trend can be observed for the vacuum values.

In the interface of System 1k1l extracted after molecular mechanics energy-minimization (Figure 3A), Shells 1 and 2 contain 5 and 10 water molecules, respectively (Table 1). The water molecules of Shell 1 (Figure 3A) are located at the bottom of the interface bridging between the target and ligand (solute) partners. Shell 2 waters mostly occur at the opening of the interface towards the bulk (right side of Figure 3A) waters/region. As it was expected, large clusters of waters gathered around charged or polar groups. For example, the sulfonyl group (Figure 3B) of the ligand is surrounded by a group of water molecules, and only one of them belongs to Shell 1. No interactions were observed between the waters and the closest target fragment ($G_{216}SG_{218}$).

During semi-empirical QM relaxation (Figure 3C), positions and orientations of some water molecules were changed. For example, two water molecules (marked with crosses in Figure 3C) were shifted by 3.2 and 1.8 Å. The orientation of Shell 1 water molecule (marked with asterisk in Figure 3C) was changed to interact with the target fragment. Such changes resulted in an extensive H-bonding network of water molecules stabilizing the target-ligand interaction around the sulfonyl group. Formation of new hydrogen bonds imply that some of the Shell 2 water molecules became Shell 1 (not marked in Figure 3C).

While the hydration structure underwent a remarkable transformation during semi-empirical QM relaxation, the conformation of the target fragment was preserved. The above example of System 1k1l (Figure 3) showed how water molecules in the different shells contribute to the completion of the target-ligand interface structure and a consequent decrease in residuals of calculated ΔH_b.

Besides small, rigid ligands like the phenylalanine derivative of System 1k1l, large peptide ligands were also involved in the present study. For example, System 2bba (Figure 4) has a penta-decapeptide ligand (Table 1) and a relatively extensive hydration structure of 27 water molecules in the extracted interface. In the case of 2bba, the largest decrease from 4.34 to 2.13 kcal/mol of the residual (COSMO models in Table 2) was obtained with Shell 1 water molecules. A detailed overview of the hydration structure shows that water molecules of Shell 1 (asterisks in Figure 4) mostly positioned at the bottom of the binding pocket and play a bridging role between the target and ligand partners. In this case, application of Shell 2 waters in addition to Shell 1 ones was not beneficial as they increased the residual. However, the final residual with Shell 3 is still below the dry model.

Beyond bridging and space filling roles presented in Figures 3 and 4, interfacial hydration also exerts a shielding effect [66] on target-ligand intermolecular interactions, as well. Despite the importance of the hydration structure, crystallography often does not supply crucial water positions or erroneously assigns waters in close contact (see also Introduction). This leads to limitations of the use of experimental complex structures in drug design.

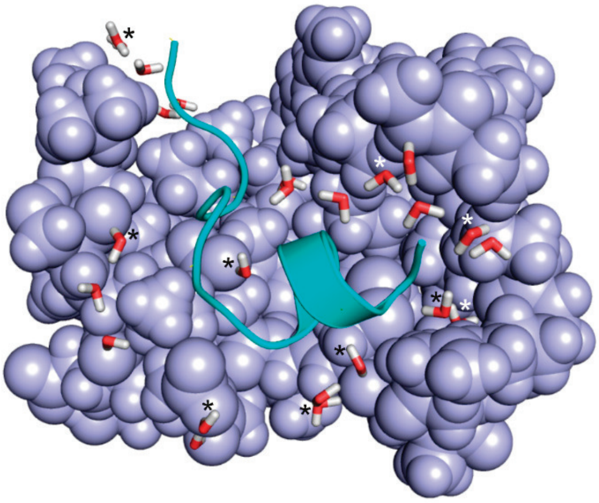


Figure 4. Extracted complex interface of System 2bba after relaxation at semi-empirical level using PM7 parameterization and Mozyme. Target fragments and ligand are shown in light blue space filling and green cartoon representations, respectively. Water molecules (sticks) in Shell 1 are marked with asterisks. Non-marked waters belong to Shell 2.

The present study has overcome such limitations of experimental determination of hydration structures, and calculation of ΔH_b was possible using complete interfacial hydration structures resulted exclusively by MobyWat calculations (see Methods). Besides hydration structures, missing ligand positions of four Systems (3ptb_pad, 3ptb_pam, 3ptb_pme, 3ptb_pmo) were also produced by computational modeling. Thus, modeling provided atomic resolution data reliably completing experimental structures and yielding robust correlations of the present study. Notably, modeling steps (Figure 5) of building the hydration structure and the full complex require only moderate computational resources, and can be accomplished on a single workstation. With the application of a parallelized MD engine and a supercomputing facility, the calculation time can be reduced to a couple of hours. The Fragmenter step takes some seconds.

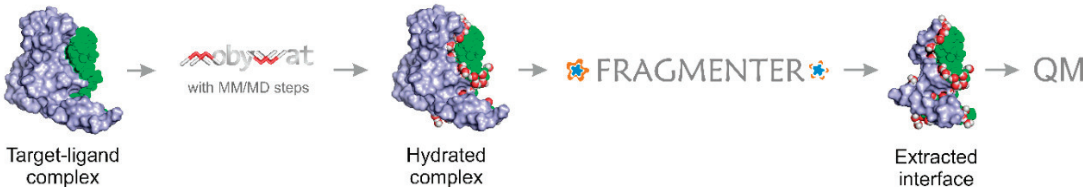


Figure 5. Modeling steps used for preparation of hydrated target-ligand interface for QM calculations shown on the example of System 1jgn. The procedure starts from the complex of the target (grey) and ligand (green) molecules. Program MobyWat [32,46] provides accurate hydration structure with MD-based calculations of positions of water molecules (red and white) in the interface. MobyWat is downloadable free of charge at www.mobywat.com. Finally, Fragmenter, a web service extracts the complex interface with considerably reduced target part for subsequent use in QM calculations. Fragmenter was introduced in the present study and available at www.fragmenter.xyz.

3. Methods

3.1. Preparation of Complexes

The primary input structures of all systems (Table 1) were obtained from the Protein Databank (PDB [67]). All crystallographic water molecules were removed. Missing atoms of solute side chains (both protein and ligand) were reconstructed with Swiss PDB Viewer [68]. In the case of missing terminal and non-terminal amino acids, acetyl and amide capping groups were added with the Schrödinger Maestro program package v. 9.6 [69] to the N-and C-terminus, respectively. In cases of homodimer structures, chain A was selected for calculations.

3.2. Parameters of Non-Amino Acid Ligands

For non-standard (non-amino-acid) ligands or residues molecular mechanics force field parameters were obtained from the GAFF force field [70]. Considering a non-standard residue, it was first capped on both terminals, with Ace- and -NHMe groups and pre-minimized with PC Model 9 [71] using MMFF94 force field [72]. Subsequently, semi-empirical quantum mechanics optimization was performed with MOPAC-2009 [73] using the PM6 parameterization with a 0.001 gradient norm [74]. In all cases, the force constant matrices were positive definite. Then, the completely minimized molecules were uploaded to RED server [75] to perform ab initio geometry optimization to obtain partial charges by RESP-A1B charge fitting (compatible with the AMBER99SB-ILDN force field). The calculations were performed with the Gaussian09 software [76], using HF/6-31G* split valence basis set [77]. The caps on the termini were excluded from charge derivation, charge restraints were applied on these atoms. Normal mode analysis was performed using GAMESS [78] to ensure that the final geometry is in energy minimum. Bond stretching, angle bending, and torsional parameters were assigned with the parmchk utility of AmberTools 1.5 [79] and used together with the partial charges to build GROMACS [80,81] residue topology entries for the non-standard residues.

3.3. Calculation of Interfacial Hydration Structure

MobyWat [46] predictions along with GROMACS MD simulations were used for calculation of water positions in the target-ligand complex. A uniform procedure was followed based on Method 3 of a previous study [32] briefly described in the following points. An overview of the modeling steps described in the forthcoming sections is provided in a flow chart of Figure 5.

3.4. Molecular Mechanics Energy-Minimization during MobyWat Predictions

For pre-MD minimization, the target or complex structure was placed in a cubic box using a distance criterion of 1 nm between the solute and the box. Void spaces of the box were filled up by explicit TIP3P water molecules [82] with the standard gmx solvate routine of GROMACS. Counter-ions (sodium or chloride) were added to neutralize the system. A uniform, procedure was applied in all cases prior to the MD steps, including a steepest descent (sd) followed by a conjugate gradient (cg) step. Exit tolerance levels were set to 10^3 and 10 kJ·mol^{-1}·nm^{-1} while maximum step sizes were set to 0.5 and 0.05 nm, respectively. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJmol^{-1}nm^{-2}. All calculations were performed with programs of the GROMACS software package [81], using the AMBER99SB-ILDN force field [83]. The above energy-minimization was performed twice, once for the target and once for the re-assembled target-ligand complex (see below).

3.5. Molecular Dynamics of the Protein Target

After energy-minimization, 5-ns-long NPT MD simulations were carried out with a time step of 2 fs. For temperature-coupling the velocity rescale [84] and the Parrinello–Rahman algorithm were used. Solute and solvent were coupled separately with a reference temperature of 300 K and a coupling time constant of 0.1 ps. Pressure was coupled by the Parrinello–Rahman algorithm [85–87] and a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1} and reference pressure of 1 bar. Particle Mesh-Ewald summation was used for long range electrostatics. Van der Waals and Coulomb interactions had a cut-off at 11 Å. Coordinates were saved at regular time-intervals of 1 ps yielding 1.001×10^3 frames. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ·mol^{-1}·nm^{-2}. Periodic boundary conditions were treated before analysis to make the solute whole and recover hydrated solute structures centered in the box. Each frame was fit to the original protein crystal structure using Cα atoms. The final trajectory including all atomic coordinates of all frames was converted to portable binary files. The target structure, and the surrounding (surface) water molecules were extracted as the last frame of the 5-ns-long MD simulation. At this point, there is a difference between the present study and Method 3 applied previously [32]. In Method 3, surface water molecules had been provided by MobyWat using 1-ns-long MD simulation. In the present study, the final frame of a 5-ns-long MD simulation was applied.

3.6. Re-Assembly of the Target-Ligand Complex

The target-ligand complex was re-assembled. For this, the target part of the holo and the hydrated apo systems were fitted on the top of each-other and the ligand was used together with the hydrated target (soaking), and interfacial water molecules were extracted. A water molecule was considered interfacial if intermolecular distance was smaller than/equal to a pre-defined maximal threshold (d_{max}) of 5 Å for both the ligand and target partners. Water molecules conflicting with the ligand structure were excluded using the editing mode of MobyWat at a minimum distance limit (d_{min}) of 1.75 Å prior the second MD simulation.

3.7. Molecular Dynamics of the Target-Ligand Complex

The MD simulation protocol described above for protein targets was performed for the re-assembled target-ligand complex structure, as well. In this case, all frames of the final trajectory of the target-ligand complex (in a water box) were used in the next step for production of interfacial water positions.

3.8. Production of Interfacial Water Positions

After the MD simulation of the target-ligand complex, MobyWat prediction of interfacial water positions was performed with dmax, clustering and prediction tolerances of 5.0, 3.0, and 3.0 Å, respectively. The MER clustering algorithm of MobyWat was applied. At this point, the present procedure differs from Method 3 [32]. As a result, a list of predicted water oxygen positions was produced by MobyWat in PDB format.

3.9. Molecular Mechanics Energy-Minimization after MobyWat

The MobyWat-supplied oxygen atoms of predicted water positions were equipped with hydrogen atoms and energy minimization was performed for the hydrated complexes. A four-step protocol was applied for energy minimization of complexes with predicted water positions following an sd-cg-sd-cg pattern with parameters of sd and cg methods described above. During the first two steps, all solute heavy atoms and the oxygen of the predicted interfacial water molecules were position restrained and bulk waters and ions were released. In the last two steps, position restraints were not applied on predicted waters, only solute heavy atoms were position restrained. Other details were the same as described in Section 3.4 above.

3.10. Extraction of Target-Ligand Interfaces by Fragmenter

Fragmenter automatically extracts target-ligand interfaces of large complexes and is freely available as a web service at www.fragmenter.xyz. Algorithm details and connections between input, algorithm, implementation, and output scripts are presented in Figures S1 and S2. In brief, the extraction is based on the selection determined by the target-ligand (d_{TL}) and the water-solute (d_w) distances as well as the inter-residual distance (n). In the main loop (Figure S1), a target amino acid residue is extracted if it has at least one heavy atom with $d_{cls} \leq d_{TL}$, where d_{cls} is the spatial distance measured between the closest heavy atoms of the actual target and ligand molecules. The maximal distance allowed between the closest heavy atoms of the target and the ligand (d_{TL}) can provided by the user and a default value is set to 3.5 Å. The same distance between solute partners and water molecules (d_W) is also defined and applied for extraction of interfacial waters. Connecting amino acids and terminating groups are also inserted. The length of fragment peptides is influenced by the maximal inter-residual topological distance (n) of the target. Parameter n specifies how many adjacent amino acids are added to the fragment chain of amino acids extracted above by the d_{TL} criterion. If $n > 0$, then the fragment was grown by adding n connecting amino acid residues. If $n = 0$ only amino acids with $d_{cls} \leq d_{TL}$ are added to the fragment chain. If $n = 1$, the sequential first neighbors are also attached to the terminus (termini) of the fragment chain, even if the attached amino acids have a $d_{cls} > d_{TL}$, etc. (Table S2).

3.10.1. Input

The actual content of the query form of the 'submit' tab of the web interface (Figure S4) is saved as a single input file (project_ID.inp) generated according to a template inputfile.inp (Figure S6). This template contains the system variables, php path, the path of the createqinput.sh, and the template for the input parameters from the website. The 'submit' tab allows setting distance (d_{TL}, d_W, and n) and other parameters of Table S1. Fragmenter offers an option to freeze (restrain) atomic positions by labeling certain groups of heavy atoms such as backbone C_α-atoms, heavy atoms, all heavy atoms in the Mopac input file. The definition of these restraints, additional Mopac parameters, and other administrative details are also collected in the project_ID.inp file. The latter parameters include the path and the file name of the complex structure, the process name (for the SLURM workload manager), the path of the php executable and Fragmenter scripts the mopac license file (for SLURM) and the path of the mopac and php executable. Using the above path and file information, setting of system variables is performed by script genqinput.sh. The user does not need to care about server configuration (e.g., server specific php executable path), it is stored on the server. Clicking on the 'submit' button the script calculate.php checks the integrity of the complex structure (Figures S1 and S2) by a PDB to PDB file conversion using OpenBabel [88]. In the case of conversion errors Fragmenter terminates and the errors are displayed on a separate page. Then it collects and transforms the input parameters from inputfile.inp and from the site from the user for the script createqinput.sh and calls createqinput.sh. Script genqinput.sh requires only one input file (project_ID.inp), which contains all necessary parameters for the run (Figure S6).

3.10.2. Main Algorithm

Having all input data in project_ID.inp, script createqinput.sh calls script fragment.php, the main engine of Fragmenter and creates the output files using other php classes (point.php, atom.php, charge.php, ligand.php). Among the classes (i) atom.php represents the atom objects with coordinates, type; (ii) module charge.php calculates the charge the ligand and generated fragment chains; (iii) module ligand.php handles a ligand object, contains the atoms, bonds, it reads and writes the pdb files; (iv) Point.php is a small class reserved for the coordinates of the atoms. Utils.php collects technical parameters, for example operating system dependent information, config file handling, etc.

Script fragment.php (Figure S2) includes steps for input processing, fragmenting, and working with output files. Target and ligand objects are obtained from the input steps, target, ligand residues, and water molecules are detected based on their chain IDs and residue types (WAT, SOL, H2O), respectively. Accordingly, the input structure is split into ligand, target and water molecules, the residues are sorted by their residue IDs and only heavy atoms of the target are examined. In the main loop of fragment.php, the target amino acid residues are selected according to dTL and n by ligand.php and point.php. Single residue-gaps are excluded by connecting two neighboring fragments by selecting the connecting residue, as well.

Having all target fragments produced in the previous steps, each of them are terminated by a uniform procedure as represented by the cycle of fragment.php (Figures S1 and S2). In the case of a free N-terminus a protonated amino group is built automatically by adding hydrogen atoms in a correct geometry. Similarly, in the case of a free C-terminus, a carboxylate anion is left unchanged. After merging ligand and water molecules with the target fragments into a new PDB file, the total charge of the complex is calculated and stored in the remark section of the file. For all cut target chains, Ac- (at N-terminus) and -NHMe (at C-terminus) blocking groups are built on both or non-free ends using atoms of previous and/or next amino acids of the chain and adding three hydrogen atoms to the methyl group (Figure S3). Following the generation of all fragments, the interface water molecules are extracted according to the intermolecular distance cut-off (d_w, Table S1). After extraction of the water molecules, their total net charges are calculated by charge.php using individual charges of amino acids (Table S3) at pH 7. Special care was taken for disulfide bridges between side-chains of cysteine amino acids. Following the main loop (Figure S1), Cys residues connected via disulfide bridges are also selected and added to the fragments. Total net charge (Table S3) of the target fragments is calculated. In the case of disulfide bridges or protonated sulfhydryl group the charge of Cys is automatically set to zero, otherwise −1. The charge of His is calculated according to the protonation state of the imidazole ring (−1, 0, +1).

3.10.3. Target-Ligand Intermolecular Interaction Energy

Fragmenter calculates target-ligand intermolecular interaction energy (E_{inter}) for the extracted interface, which is expressed as the sum of Lennard-Jones (LJ) and Coulomb potentials (Equation (5)). For both the LJ and Coulomb potentials, Amber force field parameters are used [83,89]. A per-residue list of the E_{inter} is printed in the 'results' table. The list can be used for identification of target residues colliding with the ligand as large E_{inter} values. In such cases, further MM energy-minimization may be required to achieve a complex structure appropriate for QM investigations.

$$E_{inter} = E_{LJ} + E_{Coulomb} = \sum_{i,j}^{N_T N_L}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0\ \varepsilon_r r_{ij}}\right] \tag{5}$$

$$A_{ij} = \ \varepsilon_{ij}R_{ij}^{12}\ ;\ B_{ij} = \ 2\varepsilon_{ij}R_{ij}^{6}\ ;\ R_{ij} = \ R_i + R_j\ ;\ \varepsilon_{ij} = \ \sqrt{\varepsilon_i\ \varepsilon_j}$$

where, ε_{ij} is the potential well depth at equilibrium between the ith (ligand) and jth (target) atoms; ε_0 is the permittivity of vacuum; $\varepsilon_r = 1$, relative permittivity; R_{ij} is the inter-nuclear distance at equilibrium between ith (ligand) and jth (target) atoms; q is the partial charge of an atom; r_{ij} is the actual distance between the ith (ligand) and jth (target) atoms; N_T is the number of target atoms; N_L is the number of ligand atoms.

3.10.4. Output

Fragmenter stores the output files in an output directory and provides a download link to all of them in the 'results' site (Figure S5). The ligand, the selected water molecules and the target fragments are downloadable as a complex in PDB format. The final charge of the complex is stored in the remark section of the PDB file. Fragmenter also provides additional separate PDB files and also converts them into downloadable Mopac input files. These files include the structures of the complex with/without water molecules, separate ligand, or target fragments. The 'output' tab also features the fragmented complex in a small window and it can be manipulated by the user. The visualization and rotation is performed by JSmol [90] implemented in the web page.

3.11. Calculation of Heats of Formation

Mopac 2012 [91] was used for structural relaxation and calculation of heats of formation of the extracted complex structures, separate ligand, water molecules, and target fragments. Hamiltonian of the Parametric Method number 7 (PM7 [92]) was applied. The exit criterion of the energy-minimization was defined as a gradient norm of 1.0. The value was set according to the instructions of the Mopac Support Team, and it is a magnitude smaller than the value of 10 suggested by the Manual [92]. There were only four vacuum calculations where the final gradient norm was slightly higher than 1, and the largest one of the four was 2.5. To reduce computational cost, the localized molecular orbital approach of Mozyme [93] was applied. Total net charges of the molecules were calculated from individual net charges of the amino acids (Table S3). The charges were indicated in the command line, checked manually and automatically with keyword GEO-OK. To prevent unwanted termination of calculations, keyword PREC was applied. Eigenvector following [94] was used as a default geometry optimization. Molecular mechanics correction to peptide bonds was applied by keyword MMOK. Except the cases of in vacuo calculations, the COSMO (COnductor-like ScreeningMOdel) model [63] was used. For this, a value of 78.3 was set at the EPS key word which is the dielectric constant of water at 293.15 K and 101325 Pa. $\Delta_f H$ of water was calculated with the above keywords in vacuum and using the COSMO model, respectively. In the cases of four systems, integrity of disulphide bridges was conserved by restraining the coordinates of S atoms during COSMO calculations.

3.12. Statistics

Simple linear regressions were performed between calculated $\Delta_r H$ and experimental ΔH_b(exp) values in all cases of Table 2 and Tables S4–S7. ΔH_b(exp) values were obtained from various publications as listed in Table 1. Statistical parameters of the regressions including regression coefficients (α and β in Equation (3)), coefficients of determination (R^2), *t*-values, *F*-values, residuals and root mean square error (RMSE) values are listed in Table 2. Leave-one-out cross-validated R^2 values were also calculated to check the stability of the correlations. Significance values of regression coefficients mentioned in the main text were calculated by two-sided *t*-test. For correlation plots, ΔH_b(calc) values were calculated using $\Delta_r H$ values and the regression coefficients (Equation (3)).

4. Conclusions

Structure-based calculation of binding thermodynamics is challenging at the QM level. To overcome the limitations of system size and hydration, a new protocol was introduced combining a MobyWat-based prediction of hydration structure with Fragmenter, a tool designed for extraction of the target-ligand interface with peptide fragments representing the target molecule. The protocol allowed fast QM calculations on a series of target-ligand interfaces with systematically adjusted hydration models. High correlations were achieved with a hybrid model involving a shell of explicit water molecules of calculated positions and the implicit solvation method COSMO. At semi-empirical QM level, and PM7 parameterization, a single, statistically significant scale factor was obtained for conversion of calculated reaction heats into experimental binding enthalpy values. The results of the present study will be particularly helpful in enthalpic optimization of drugs and in the molecular design of stable complexes and new ligands, in general. Further development and tests of the protocol have been also initiated for applications at the highest level of QM theory.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/18/4384/s1.

Author Contributions: Conceptualization, C.H.; Methodology, N.J., M.B., I.H., and C.H.; Software, I.H. and C.H.; Validation, I.H., N.J., and M.B.; Formal analysis, I.H., and C.H.; Investigation, I.H., N.J., and M.B.; Data curation, C.H., G.P., M.B., and I.H.; Writing, C.H., I.H., M.B, N.J., and G.P.; Visualization, C.H. and I.H.; Supervision, C.H.; Project administration, C.H.; Funding acquisition, C.H. and I.H.

Funding: This work was supported by the Hungarian National Research, Development and Innovation Office (K123836). This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We acknowledge the grant of computer time from CSCS Swiss National Supercomputing Centre, and the Governmental Information Technology Development Agency, Hungary. The University of Pécs is acknowledged for a support by the 17886-4/23018/FEKUTSTRAT excellence grant. C.H.'s work was supported by a grant co-financed by Hungary and the European Union (EFOP-3.6.2-16-2017-00008).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

COSMO	Conductor-like screening model
E_{inter}	Intermolecular interaction energy
ITC	Isothermal titration calorimetry
LJ	Lennard-Jones
MD	Molecular dynamics
MM	Molecular mechanics
MW	Molecular weight
PDB	Protein Databank
PM7	Parametric method 7
QM	Quantum mechanical
RMSE	Root mean square error

References

1. Bajusz, D.; Ferenczy, G.; Keseru, G. Structure-Based Virtual Screening Approaches in Kinase-Directed Drug Discovery. *Curr. Top. Med. Chem.* **2017**, *17*, 2235–2259. [CrossRef]
2. Talhout, R.; Engberts, J.B. Thermodynamic analysis of binding of p-substituted benzamidines to trypsin. *Eur. J. Biochem.* **2001**, *268*, 1554–1560. [CrossRef]
3. Sleigh, S.H.; Seavers, P.R.; Wilkinson, A.J.; Ladbury, J.E.; Tame, J.R.H. Crystallographic and Calorimetric Analysis of Peptide Binding to OppA Protein. *J. Mol. Biol.* **1999**, *291*, 393–415. [CrossRef]
4. Dullweber, F.; Stubbs, M.; Musil, D.; Stürzebecher, J.; Klebe, G. Factorising ligand affinity: A Combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.* **2001**, *313*, 593–614. [CrossRef]
5. Palencia, A.; Cobos, E.S.; Mateo, P.L.; Martinez, J.C.; Luque, I. Thermodynamic Dissection of the Binding Energetics of Proline-rich Peptides to the Abl-SH3 Domain: Implications for Rational Ligand Design. *J. Mol. Biol.* **2004**, *336*, 527–537. [CrossRef]
6. McNemar, C.; Snow, M.E.; Windsor, W.T.; Prongay, A.; Mui, P.; Zhang, R.; Durkin, J.; Le, H.V.; Weber, P.C. Thermodynamic and structural analysis of phosphotyrosine polypeptide binding to Grb2-SH2. *Biochemistry* **1997**, *36*, 10006–10014. [CrossRef]

7. Wang, C.; Pawley, N.H.; Nicholson, L.K. The role of backbone motions in ligand binding to the c-Src SH3 domain. *J. Mol. Biol.* **2001**, *313*, 873–887. [CrossRef]
8. Org, T.; Chignola, F.; Chignola, F.; Hetényi, C.; Gaetani, M.; Rebane, A.; Liiv, I.; Maran, U.; Mollica, L.; Bottomley, M.J.; et al. The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. *EMBO Rep.* **2008**, *9*, 370–376. [CrossRef]
9. Chrencik, J.E.; Brooun, A.; Recht, M.; Kraus, M.L.; Koolpe, M.; Kolatkar, A.; Bruce, R.H.; Martiny-Baron, G.; Widmer, H.; Pasquale, E.B.; et al. Thermodynamic and structural analysis of phosphotyrosine polypeptide binding to Grb2-SH2. *Biochemistry* **2006**, *14*, 321–330.
10. Kozlov, G.; De Crescenzo, G.; Lim, N.S.; Siddiqui, N.; Fantus, D.; Kahvejian, A.; Trempe, J.-F.; Elias, D.; Ekiel, I.; Sonenberg, N.; et al. Structural basis of ligand recognition by PABC, a highly specific peptide-binding domain found in poly(A)-binding protein and a HECT ubiquitin ligase. *EMBO J.* **2004**, *23*, 272–281. [CrossRef]
11. Day, C.L.; Smits, C.; Fan, F.C.; Lee, E.F.; Fairlie, W.D.; Hinds, M.G. Structure of the BH3 domains from the p53-inducible BH3-only proteins Noxa and Puma in complex with Mcl-1. *J. Mol. Biol.* **2008**, *380*, 958–971. [CrossRef]
12. Lu, Z.; Zhang, Y. Interfacing ab initio Quantum Mechanical Method with Classical Drude Osillator Polarizable Model for Molecular Dynamics Simulation of Chemical Reactions. *J. Chem. Theory Comput.* **2008**, *4*, 1237–1248. [CrossRef]
13. Vanommeslaeghe, K.; Guvench, O.; MacKerell, A.D. Molecular Mechanics. *Curr. Pharm. Des.* **2014**, *20*, 3281–3292. [CrossRef]
14. Ganoth, A.; Friedman, R.; Nachliel, E.; Gutman, M. A Molecular Dynamics Study and Free Energy Analysis of Complexes between the Mlc1p Protein and Two IQ Motif Peptides. *Biophys. J.* **2006**, *91*, 2436–2450. [CrossRef]
15. Fenley, A.T.; Muddana, H.S.; Gilson, M.K. Entropy—Enthalpy transduction caused by conformational shifts can obscure the forces driving protein—Ligand binding. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 20006–20011. [CrossRef]
16. Huggins, D.J. Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations. *Biophys. J.* **2015**, *108*, 928–936. [CrossRef]
17. Dougherty, D.A. Cation-π Interactions Involving Aromatic Amino Acids. *J. Nutr.* **2007**, *137* (Suppl. S1), 1504–1508. [CrossRef]
18. Majumdar, S.; Maiti, S.; Ghosh Dastidar, S. Dynamic and Static Water Molecules Complement the TN16 Conformational Heterogeneity Inside the Tubulin Cavity. *Biochemistry* **2016**, *55*, 335–347. [CrossRef]
19. Poole, P.L.; Finney, J.L. Hydration-induced conformational and flexibility changes in lysozyme at low water content. *Int. J. Biol. Macromol.* **1983**, *5*, 308–310. [CrossRef]
20. Jukič, M.; Konc, J.; Gobec, S.; Janežič, D. Identification of Conserved Water Sites in Protein Structures for Drug Design. *J. Chem. Inf. Model.* **2017**, *57*, 3094–3103. [CrossRef]
21. Choi, H.; Kang, H.; Park, H. New solvation free energy function comprising intermolecular solvation and intramolecular self-solvation terms. *J. Chem.* **2013**, *5*, 5–8. [CrossRef]
22. Cui, G.; Swails, J.M.; Manas, E.S. SPAM: A Simple Approach for Profiling Bound Water Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 5539–5549. [CrossRef]
23. García-Sosa, A.T.; Stuart, F.C.; Mancera, R.L. Including Tightly-Bound Water Molecules in de Novo Drug Design. Exemplification through the in Silico Generation of Poly(ADP-ribose)polymerase Ligands. *J. Chem. Inf. Model.* **2005**, *45*, 624–633. [CrossRef]
24. Huggins, D.J.; Tidor, B. Systematic placement of structural water molecules for improved scoring of protein—Ligand interactions. *Protein Eng. Des. Sel.* **2011**, *24*, 777–789. [CrossRef]
25. Ladbury, J.E. Just add water! The effect of water on the specificity of proteinligand binding sites and its potential application to drug design. *Cell Chem. Biol.* **1996**, *3*, 973–980.
26. Lloyd, D.G.; García-Sosa, A.T.; Alberts, I.L.; Todorov, N.P.; Mancera, R.L. The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comp. Aided Mol. Des.* **2004**, *18*, 89–100. [CrossRef]
27. Michel, J.; Tirado-Rives, J.; Jorgensen, W.L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411. [CrossRef]
28. Zheng, Z. Computational Modeling of Solvent Effects on Protein-Ligand Interactions Using Fully Polarizable Continuum Model and Rational Drug Design. *Commun. Comput. Phys.* **2013**, *13*, 31–60. [CrossRef]

29. Kunstmann, S.; Gohlke, U.; Broeker, N.K.; Roske, Y.; Heinemann, U.; Santer, M.; Barbirz, S. Solvent Networks Tune Thermodynamics of Oligosaccharide Complex Formation in an Extended Protein Binding Site. *J. Am. Chem. Soc.* **2018**, *140*, 10447–10455. [CrossRef]
30. Brysbaert, G.; Blossey, R.; Lensink, M.F. The Inclusion of Water Molecules in Residue Interaction Networks Identifies Additional Central Residues. *Front. Mol. Biosci.* **2018**, *5*, 88. [CrossRef]
31. Quiocho, F.A.; Wilson, D.K.; Vyas, N.K. Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* **1989**, *340*, 404–407. [CrossRef]
32. Jeszenői, N.; Bálint, M.; Horváth, I.; van der Spoel, D.; Hetényi, C. Exploration of interfacial hydration networks of target—Ligand complexes. *J. Chem. Inf. Model.* **2016**, *56*, 148–158. [CrossRef]
33. Afonine, P.V.; Grosse-Kunstleve, R.W.; Adams, P.D.; Urzhumstev, A. Bulk-solvent and overall scaling revisited: Faster calculations, improved results. *Acta Cryst. D Biol. Cryst.* **2013**, *69 Pt 4*, 625–634. [CrossRef]
34. Badger, J. Modeling and refinement of water molecules and disordered solvent. *Meth. Enzymol.* **1997**, *277*, 344–352.
35. Finney, J.L. The organization and function of water in protein crystals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **1977**, *278*, 3–32. [CrossRef]
36. Halle, B. Protein hydration dynamics in solution: A critical survey. *Philos. Trans. R. Soc. Lond. B* **2004**, *359*, 1207–1224. [CrossRef]
37. Elsässer, S.J.; Huang, H.; Lewis, P.W.; Chin, J.W.; Allis, C.D.; Patel, D.J. DAXX envelops a histone H3.3–H4 dimer for H3.3-specific recognition. *Nature* **2012**, *491*, 560–565. [CrossRef]
38. Abel, R.; Young, T.; Farid, R.; Berne, B.; Friesner, R. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831. [CrossRef]
39. Pearlstein, R.; Hu, Q.; Zhou, J.; Yowe, D.; Levell, J.; Dale, B.; Kaushik, V.; Daniels, D.; Hanrahan, S.; Sherman, W.; et al. New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: Analysis of the epidermal growth factor-like repeat A docking site using WaterMap. *Proteins* **2010**, *78*, 2571–2586. [CrossRef]
40. Vukovic, S.; Brennan, P.E.; Huggins, D. Exploring the role of water in molecular recognition: Predicting protein ligandability using a combinatorial search of surface hydration sites. *J. Phys. Condens. Matter.* **2016**, *28*, 344007. [CrossRef]
41. Hylsová, M.; Carbain, B.; Fanfrlik, J.; Lepsik, M. Explicit treatment of active-site waters enhances quantum mechanical/implicit solvent scoring: Inhibition of CDK2 by new pyrazolo [1,5-a] pyrimidines. *Eur. J. Med. Chem.* **2017**, *126*, 1118–1128. [CrossRef]
42. García-Sosa, A.T.; Mancera, R.L. Free Energy Calculations of Mutations Involving a Tightly Bound Water Molecule and Ligand Substitutions in a Ligand-Protein Complex. *Mol. Inf.* **2010**, *29*, 589–600. [CrossRef]
43. Lee, S.H.; Rossky, P.J. A comparison of the structure and dynamics of liquid water at hydrophobic and hydrophilic surfaces—A molecular dynamics simulation study. *J. Chem. Phys.* **1994**, *100*, 3334–3345. [CrossRef]
44. Watanabe, G.; Nakajima, D.; Hiroshima, A.; Suzuki, H.; Yoneda, S. Analysis of water channels by molecular dynamics simulation of heterotetrameric sarcosine oxidase. *Biophys. Physicobiol.* **2015**, *12*, 131–137. [CrossRef]
45. Patodia, S.; Bagaria, A.; Chopra, D. Molecular Dynamics Simulation of Proteins: A Brief Overview. *J. Phys. Chem. Bioph.* **2014**, *4*, 1. [CrossRef]
46. Jeszenői, N.; Horváth, I.; Bálint, M.; van der Spoel, D.; Hetényi, C. Mobility-based prediction of hydration structures of protein surfaces. *Bioinformatics* **2015**, *31*, 1959–1965. [CrossRef]
47. Field, M.J.; Bash, P.A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comp. Chem.* **1990**, *11*, 700–733. [CrossRef]
48. Hayik, S.A.; Dunbrack, R.; Merz, K.M. A Mixed QM/MM Scoring Function to Predict Protein-Ligand Binding Affinity. *J. Chem. Theory Comp.* **2010**, *6*, 3079–3091. [CrossRef]
49. Menikarachchi, L.C.; Gascón, J.A. QM/MM Approaches in Medicinal Chemistry Research. *Curr. Top. Med. Chem.* **2010**, *10*, 46–54. [CrossRef]
50. Van der Kamp, M.W.; Mulholland, A.J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52*, 2708–2728. [CrossRef]
51. Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249. [CrossRef]

52. Vidossich, P.; Magistrato, A. QM/MM Molecular Dynamics Studies of Metal Binding Proteins. *Biomolecules* **2014**, *4*, 616–645. [CrossRef]
53. Zhang, D.W.; Zhang, J.Z.H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein—Molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599–3605. [CrossRef]
54. Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. Semiempirical calculations of binding enthalpy for protein–ligand complexes. *Int. J. Quantum Chem.* **2004**, *97*, 747–763. [CrossRef]
55. Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. Mixed Implicit/Explicit Solvation Models in Quantum Mechanical Calculations of Binding Enthalpy for Protein–Ligand Complexes. *Int. J. Quantum Chem.* **2006**, *106*, 1943–1963. [CrossRef]
56. Dobeš, P.; Otyepka, M.; Strnad, M.; Hobza, P. Interaction Energies for the Purine Inhibitor Roscovitine with Cyclin-Dependent Kinase 2: Correlated Ab Initio Quantum-Chemical, DFT and Empirical Calculations. *Chemistry* **2006**, *12*, 4297–4304. [CrossRef]
57. Freire, E. Do Enthalpy and Entropy Distinguish First in Class from Best in Class? *Drug Discov. Today* **2008**, *13*, 869–874. [CrossRef]
58. Ferenczy, G.G.; Keserű, G.M. Thermodynamics guided lead discovery and optimization. *Drug Discov. Today* **2010**, *15*, 919–932. [CrossRef]
59. Hann, M.M.; Keserü, G.M. Finding the sweet spot: The role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug. Discov.* **2012**, *11*, 355–365. [CrossRef]
60. Zhang, D.W.; Xiang, Y.; Zhang, J.Z.H. New Advance in Computational Chemistry: Full Quantum Mechanical ab Initio Computation of Streptavidin—Biotin Interaction Energy. *J. Phys. Chem. B* **2003**, *107*, 12039–12041. [CrossRef]
61. Zhang, D.W.; Xiang, Y.; Gao, A.M.; Zhang, J.Z.H. Quantum mechanical map for protein-ligand binding with application to β-trypsin/benzamidine complex. *J. Chem. Phys.* **2004**, *120*, 1145–1148. [CrossRef]
62. Brown, S.; Shirts, M.; Mobley, D. Free-energy calculations in structure-based drug design. *Drug Des. Struct. Ligand Based Approaches* **2010**, *2010*, 61–86.
63. Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *J. Chem. Soc. Perk. Trans.* **1993**, *2*, 799–805. [CrossRef]
64. Weichenberger, C.X.; Afonine, P.V.; Kantardjieff, K.; Rupp, B. The solvent component of macromolecular crystals. *Acta Cryst. D Biol. Cryst.* **2015**, *71*, 1023–1038. [CrossRef]
65. Halle, B. Biomolecular cryocrystallography: Structural changes during flash-cooling. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4793–4798. [CrossRef]
66. Schmidtke, P.; Barril, X.; Luque, F.J.; Murray, J.B. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J. Am. Chem. Soc.* **2011**, *133*, 18903–18910. [CrossRef]
67. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
68. Guex, N.; Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714–2723. [CrossRef]
69. *Schrödinger Release 2019-3: Maestro*; Schrödinger, LLC: New York, NY, USA, 2019.
70. Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174. [CrossRef]
71. Gille, A.L.; Dutmer, B.C.; Gilbert, T.M. PCMODEL 9.2. *J. Am. Chem. Soc.* **2009**, *131*, 5714. [CrossRef]
72. Halgren, T. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519. [CrossRef]
73. Stewart, J.J.P. *MOPAC2009, 2009*; Steward Computational Chemistry: Colorado Springs, CO, USA, 2008.
74. Stewart, J.J.P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213. [CrossRef]
75. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J.C.; Cieplak, P.; Dupradeau, F.Y. R.E.D. Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **2011**, *39*, W511–W517. [CrossRef]
76. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; et al. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, USA, 2009.
77. Krishnan, R.; Binkley, J.S.; Seeger, R.; Pople, J.A. Self-consistent molecular-orbital methods. XX. Basis set for correlated wave-functions. *J. Chem. Phys.* **1980**, *72*, 650–654. [CrossRef]

78. Schmidt, M.W.; Baldridge, K.K.; Boatz, J.A.; Elbert, S.T.; Gordon, M.S.; Jensen, J.H.; Koseki, S.; Matsunaga, N.; Nguyen, K.A.; Su, S.J.; et al. General atomic and molecular electronic-structure system. *J. Comp. Chem.* **1993**, *14*, 1347–1363. [CrossRef]
79. Case, D.; Darden, T.; Cheatham Iii, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Walker, R.; Zhang, W.; Merz, K. *AmberTools, 15*; Amber; University of California: San Francisco, CA, USA, 2015.
80. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25. [CrossRef]
81. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M.R.; Smith, J.C.; Kasson, P.M.; van der Spoel, D.; et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854. [CrossRef]
82. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
83. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958. [CrossRef]
84. Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101. [CrossRef]
85. Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald—An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. [CrossRef]
86. Nose, S.; Klein, M.L. Constant pressure molecular-dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055–1076. [CrossRef]
87. Parrinello, M.; Rahman, A. Polymorphic transitions in single-crystals—A new molecular-dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190. [CrossRef]
88. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef]
89. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.J.; Luo, R.; Duan, Y. Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *J. Phys. Chem. B* **2012**, *116*, 7088–7101. [CrossRef]
90. Hanson, R.M.; Prilusky, J.; Renjian, Z.; Nakane, T.; Sussman, J.L. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Isr. J. Chem.* **2013**, *53*, 207–216. [CrossRef]
91. Stewart, J.J.P. Openmopac Online Manual. 2016. Available online: http://www.openmopac.net (accessed on 1 March 2013).
92. Stewart, J.J.P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1–32. [CrossRef]
93. Stewart, J.J.P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quant. Chem.* **1996**, *58*, 133–146. [CrossRef]
94. Baker, J. An Algorithm for the Location of Transition States. *J. Comp. Chem.* **1986**, *7*, 385. [CrossRef]

D27

PCCP

PAPER

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 31714

Target–ligand binding affinity from single point enthalpy calculation and elemental composition†

Viktor Szél, Balázs Zoltán Zsidó, Norbert Jeszenői and Csaba Hetényi *

Received 15th September 2023,
Accepted 1st November 2023

DOI: 10.1039/d3cp04483a

rsc.li/pccp

Reliable target–ligand binding thermodynamics data are essential for successful drug design and molecular engineering projects. Besides experimental methods, a number of theoretical approaches have been introduced for the generation of binding thermodynamics data. However, available approaches often neglect electronic effects or explicit water molecules influencing target–ligand interactions. To handle electronic effects within a reasonable time frame, we introduce a fast calculator QMH-L using a single target–ligand complex structure pre-optimized at the molecular mechanics level. QMH-L is composed of the semi-empirical quantum mechanics calculation of binding enthalpy with predicted explicit water molecules at the complex interface, and a simple descriptor based on the elemental composition of the ligand. QMH-L estimates the target–ligand binding free energy with a root mean square error (RMSE) of 0.94 kcal mol^{-1}. The calculations also provide binding enthalpy values and they were compared with experimental binding thermodynamics data collected from the most reliable isothermal titration calorimetry studies of systems including various protein targets and challenging, large peptide ligands with a molecular weight of up to 2–3 thousand. The single point enthalpy calculations of QMH-L require modest computational resources and are based on short runs with open source and/or free software like Gromacs, Mopac, MobyWat, and Fragmenter. QMH-L can be applied for fast, automated scoring of drug candidates during a virtual screen, enthalpic engineering of new ligands or thermodynamic explanation of complex interactions.

Introduction

Prediction of the binding affinity of target–ligand complexes is a cornerstone of drug design and molecular engineering. Correct estimation of binding affinity is a key to identification of potent ligands in the early (screening) stages and also saves time by reducing the amount of costly synthesis and testing steps.[1–3] The binding affinity of a target–ligand complex is expressed in terms of binding free energy (ΔG_b).[4–6] The calculated ΔG_b also guides the ranking and selection of the best binding modes during computational docking of the ligand to the target. ΔG_b is composed of binding enthalpy (ΔH_b) and entropy (ΔS_b) according to $\Delta G_b = \Delta H_b - T\Delta S_b$ (where T is the thermodynamic temperature). While ΔS_b measures the change of energy partitioning among available degrees of freedom[7] (change of the degree of disorder), ΔH_b accounts for the change of interactions[8,9] between the ligand, the surrounding target, and the water molecules.

Numerous scoring methods have been introduced[10–15] and modified[16] for the fast calculation of ΔG_b using the formulae of molecular mechanics (MM) force fields accounting for the enthalpic contributions like van der Waals, electrostatic and hydrogen-bonding interactions. The MM-based scoring functions have improved a lot since the early 1990s,[6,17–21] and are fairly successful in ranking and selection of correct binding modes during computational docking.[10,11,22] At the same time, they show a relatively large error[22] if correlated with experimental ΔG_b data. These limitations of MM-based scoring functions may originate from the lack of calculation of electronic effects, such as polarization, hydrogen bonding, aromatic interactions,[11,23] and poor representation of entropic contributions.[6] In addition, despite the importance[24] of explicit water molecules interacting with the solute (target and ligand) partners, they are generally absent during docking and scoring, and the implemented implicit solvent models often cannot provide precise results.[25]

A recent increase in computational speed allowed the development of quantum mechanics (QM) scoring approaches for the handling of the above-mentioned electronic effects (Table 1). However, relatively few QM-based studies performed a systematic correlation of calculated and experimental ΔG_b (Table 1) values. It was shown that cutting out the interfacial regions of the complex including the ligand and the

Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary. E-mail: hetenyi.csaba@pte.hu

† Electronic supplementary information (ESI) available. See DOI: **https://doi.org/10.1039/d3cp04483a**

Table 1 Previous studies on quantum chemical calculation of binding thermodynamics quantities of target–ligand complexes

Study	Number of systems (sub-set)	M_W of largest ligand	Calculated quantity	Predicted quantity[a]	R^2 (best)	Regression error[b] (kcal mol^{-1})	Calculation method
Nikitina *et al.* (2004)28	8 (full set)	700	ΔH_b	ΔH_b	—	1.85[b]	PM3-M,[d] calculated water positions
Raha *et al.* (2005)[29]	165 (full set)	1535	ΔG_b	ΔG_b	0.55	1.98[b]	PM3 interaction energy, Lennard-Jones term, Poisson–Boltzmann solvation, empirical dispersion and conformational entropy[f]
	57 (Wang subset)	na	ΔH_b		0.46	—	PM3
	16 (SPI subset)	613 (na)			0.88	—	PM3
Nikitina *et al.* (2006)[30]	12 (full set)	507	ΔH_b	ΔH_b	0.96	0.93[b]	PM3-C,[e] calculated water positions
Fanfrlík *et al.* (2010)[31]	11 (HIV-1 protease subset)	705	ΔG_b	ΔG_b	0.71	—	PM6-DH2-M[d]-C[e] enthalpy, rigid-rotor harmonic oscillator entropy, ligand desolvation (SMD)[f]
			ΔH_b		0.10	—	PM6-DH2-M[d] enthalpy
Dobes *et al.* (2011)[32]	15 (full set)	482	ΔG_b	ΔG_b	0.52	4.31[b]	PM6-DH2-M[d]-C[e] enthalpy, rigid rotor harmonic oscillator entropy, ligand desolvation (SMD)[g]
	15 (full set)		ΔH_b		0.87	1.17[b]	PM6-DH2-M[d]-C[e] enthalpy
Gonzalez *et al.* (2017)[33]	8 (HLA-DR1 subset)	1506	ΔH_b	ln(IC$_{50}$)	0.81	—	PM7-C,[e] crystallographic waters
			Electronic interaction energy		0.81	—	FMO-DFTB, crystallographic waters
	14 (HLA-DR2 subset)		ΔH_b		0.61	—	PM7-C,[e] crystallographic waters
			Electronic interaction energy		0.75	—	FMO-DFTB, crystallographic waters
Ehrlich *et al.* (2017)[23]	25 (Fxa subset)	548	ΔG_b	ΔG_b	0.47	2.8[c]	HF-3c electronic energy, DFTB3-3D thermostatic correction, C[e]-RS solvation free energy[h]
	16 (TYK2 subset)	392			0.55	2.7[c]	
Hylsová *et al.* (2017)[34]	21 (full set)	562	ΔG_b	ln(IC$_{50}$)	0.68	—	DFT-D3 and PM6-D3X4-C[e] combined, crystallographic and predicted waters[i]
Pecina *et al.* (2018)[35]	10 (full set)	391	ΔG_b	ΔG_b	0.69	—	DFTB3-D3H4X interaction energy, PM6-M[d]-C,[e] crystallographic waters, based on 10 docked poses
					0.58		DFTB3-D3H4X interaction energy, PM6-M[d]-C[e]
					0.56		DFTB3-D3H4X interaction energy, PM6-M[d]-C,[e] crystallographic waters, based on one conformation
Horváth *et al.* (2019)[26]	15 (full set)	3318	ΔH_b	ΔH_b	0.93	2.65[b]	PM7-M[d]-C[e], calculated waters

[a] A measured quantity correlated with the calculated quantity. [b] Root mean square error (RMSE). [c] Mean absolute deviation. [d] MOZYME linear scaling method. [e] COSMO implicit solvation model. [f] $\Delta G_b = \Delta H_b + \Delta LJ_6 + \Delta\Delta G_{solv} + \Delta S_{solv} + \Delta S_{conf}$, where ΔH_b is the enthalpy of binding, ΔLJ_6 is the Lennard-Jones term, $\Delta\Delta G_{solv}$ is the solvation free energy change during complexation, ΔS_{solv} is the solvent entropy change, and ΔS_{conf} is the conformational entropy change. [g] $\Delta G_b = \Delta H_b + T\Delta S + \Delta E_{def}$(ligand) + $\Delta\Delta G_{solv}$(ligand), where ΔH_b is the binding enthalpy, $T\Delta S$ is the rigid rotor harmonic oscillator entropic term, ΔE_{def}(ligand) is the deformation energy for the ligand, and $\Delta\Delta G_{solv}$(ligand) is the solvation free energy change for the ligand. [h] $\Delta G_b = \Delta E_{el} + \Delta G_{RRHO} + \Delta G_{solv}$, where ΔE_{el} is the electronic interaction energy, ΔG_{RRHO} is the rigid rotor harmonic oscillator entropic term, and ΔG_{solv} is the solvation free energy change. [i] $\Delta G_b = \Delta E_{int} + \Delta\Delta G_{solv} + \Delta G'^{w}_{conf}(L) - T\Delta S_{solv}$, where ΔE_{int} is the gas-phase interaction energy, $\Delta\Delta G_{solv}$ is the interaction solvation/desolvation free energy, $\Delta G'^{w}_{conf}(L)$ is the change of the conformational free energy of the ligand, and ΔS_{solv} is the entropy of the explicit water molecules.

surrounding interacting target residues (fragments) can help to reduce system size and computational costs.[23,26,27] This fragmenting (Fig. 1) approximation was applied on twenty-five complex structures for the estimation of ΔG_b[23] with an interaction energy calculation at the Hartree–Fock (HF) and composite hybrid PBEh-3c density-functional theory (DFT) levels, in combination with a DFTB3-D3 thermostatic correction supported by a conductor-like screening model for real solvents (COSMO-RS)[36] for solvation effects. Other studies also applied target fragmenting successfully,[28,30,37] and therefore, it was proved to be a valid approach of speeding up QM calculations of ligand–target complexes. While target fragmenting considerably speeds up calculations, the full optimization process at the QM level still requires considerable time for large ligands even at the semi-empirical level. While there are promising studies,[38] it has not been thoroughly investigated whether single point self-consistent field (1SCF) calculations after fast MM-preoptimization of the complex can substitute QM optimization in ΔH_b calculations.

Prediction of ΔH_b alone is also important for at least two reasons. Firstly, ΔH_b is a component of ΔG_b (dominant in a large number of target–ligand complexes), and used as a separate term in the scoring functions (see the footnotes of Table 1). Such QM-based ΔG_b calculators (Table 1) tend to estimate ΔH_b at the HF level with semi-empirical parametrizations Austin Model 1 (AM1), Parametric Method 3 (PM3), PM6 or PM7[29,35,39–41] and classical approximations are introduced for the estimation of entropic changes and implicit solvation.[27,39] Secondly, ΔH_b is key to the efficiency and selectivity of the ligands. A favourable (large negative) ΔH_b likely results in strong and specific interactions[42] and drugs of increased efficiency.[43–46] Calculated ΔH_b is also promising in

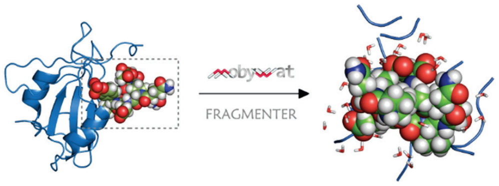


Fig. 1 Fragmenting approaches generate target fragments (peptides or amino acids) surrounding the ligand molecule. For example, in the present study, the structure of human growth factor receptor bound protein 7 src-homology 2 domain (Grb7 SH2) (target, blue cartoon) was one of the targets investigated in complex with PQPE-pY-VNQPD peptide (ligand, space filling or sticks). The ligand-binding pocket of the raw PDB structure (grey dashed frame on the left, PDB ID 1mw4) was converted into the hydrated, energy-minimized, and extracted interface structure (on the right) containing fragments of the target. The conversion procedure was performed by key programs MobyWat, Gromacs, and Fragmenter, used for the prediction of water positions, energy-minimization, and fragmentation, respectively, as described in the Methods section.

prediction of binding selectivity. It has been shown[47] that relative differences in ΔH_b values of a set of ligands calculated by a PM7/MOZYME/COSMO combination correlated well (R = 0.7) with relative IC_{50} values measured on Src homology region 2 domain-containing phosphatase targets. This combination of the semi-empirical QM approaches with COSMO implicit solvent model for the calculation of ΔH_b has been a choice in various studies with PM3,[28,30] PM6[48] and PM7[33,49] parametrizations among which PM7 may be considered as the most robust one.[50] In a comparative study, PM7 outperformed PM6, PM6-DH+, and PM6-D3 methods in differentiating between decoy *versus* native docked poses based on single-point complex formation enthalpies,[51] and similar results were obtained with PM7/COSMO.[52]

While the above studies (Table 1) hint that application of the COSMO solvent model is useful in QM-based ΔH_b calculations, there are many situations in drug design[25] where the use of explicit water molecules would be advantageous.[53] For example, in the case of large, peptide or charged ligands with numerous and/or highly occupied hydration sites, appropriate modeling of the water structure is necessary. However, there are relatively few studies[26,28,30,34,53] that published correlations on target–ligand complex structures equipped with explicit water molecules.

In the present study, we investigate if the estimation of ΔH_b is possible using an appropriately hydrated, MM-optimized complex structure and 1SCF calculation at the semi-empirical QM level. We also test the combination of the calculated ΔH_b with ligand-based descriptors to develop a ΔG_b calculator for fast scoring of target–ligand complex structures obtained from experiments or docking calculations.

Methods

Systems and data

The selection of target–ligand complex systems was directed by the availability of accurate isothermal titration calorimetry (ITC) data (both ΔH_b and ΔG_b) and the related experimental complex structure (preferably X-ray or NMR). The present study aims at calculation of both ΔH_b and ΔG_b for all systems, and ITC is the only method which measures both quantities. Collections of only K_d, K_i, or IC_{50} values were not considered for the present study. Thus, ITC databases or studies investigating protonation–deprotonation processes in different buffers were preferred, since it is a crucial aspect of accurate thermodynamic measurement.[54] In a few cases, raw data were corrected for the proton transfer (Table S1, ESI†). Most of ITC measurements were conducted at 298.15 K, however in four cases (1py1, 2v8c, 2v8f, 1axc) the temperature was 303.15 K and in one case (2ke1) 296.15 K (Table S1, ESI†). Without measured heat capacities these binding enthalpies could not be converted to their respective values at 298.15 K. However for 3ptb systems (3ptb, 3ptb_pme, 3ptb_pam, 3ptb_pmo, 3ptb_pad) recalculated data were used (from 298.25 to 298.15 K) from the previous study[26] (Table S1, ESI†).

All target–ligand complex structures were obtained from the Protein Databank (Table S1, ESI†). Fitted crystallographic water and other solvent molecules, ions were removed. Missing atoms of target residues were reconstructed with Swiss PDB Viewer.[55] In cases of homo-oligomer structures, chain A was selected for calculations. In the case of 1axc, 1jyr, 2roc, 2v8c, 2v8f and 3ask missing ligand residues were reconstructed with PyMol. For 1abo,1bbz, 1hcs, 1lcj, 4j9f, 4j9g and 4j9i acetyl and amido caps were also added to ligands in agreement with ITC measurements.

Non-standard amino acid residues among the systems were phosphotyrosine, phosphoserine and trimethyllysine for which molecular mechanics force field parameters were determined as described in a previous study.[26]

MM energy-minimization and molecular dynamics

The following procedure (Fig. 2a) was applied on all systems in Table S1 (ESI†) except the ones that were prepared in the same way in our previous study.[26] According to the MobyWat surface and interface hydration procedures,[56] the procedure was applied first on the (dry) target and subsequently on the target–ligand complex, respectively.

The MobyWat M3 procedure produces a void-free hydration structure.[56] Prior to molecular dynamics runs, the target (complex) structures were energy-minimized with molecular mechanics in a two-step fashion including steepest descent (sd) and conjugate gradient (cg) algorithms. The hydrogenated structures were placed in a cubic box using a distance criterion of 1 nm between the solute and the edge of the container. The box was filled up with explicit TIP3P water molecules and (if it was necessary) counter-ions (sodium or chloride) were added to neutralize the system. Exit tolerance levels were set to 10^3 and 10 kJ mol^{-1} nm^{-1} while maximum step sizes were set to 0.5 and 0.05 nm for the sd and cg steps, respectively. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ mol^{-1} nm^{-2}. Calculations were performed with programs of the GROMACS software package,[57] using the AMBER99SB-ILDN force field.[58]

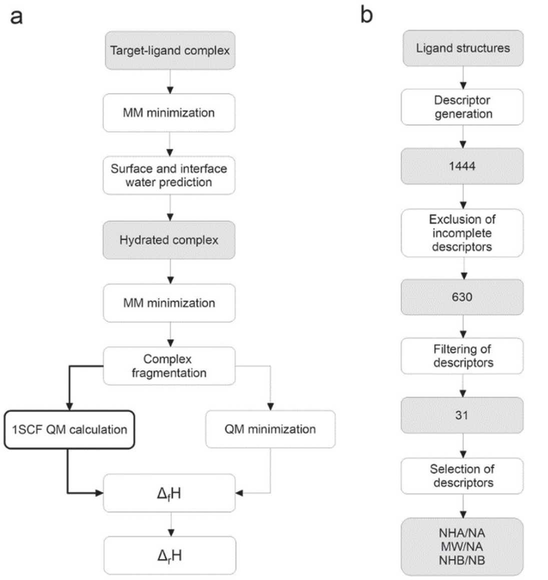


Fig. 2 (a) Key steps of the calculation of a reaction enthalpy value of ligand binding from the target–ligand complex structure. The present study focuses on the single-point (1SCF, thick arrows) alternative. (b) The procedure used for generating, filtering and selecting ligand-based descriptors for calculation of ΔG_b.

After energy-minimization, 1 ns-long *NPT* MD simulation was carried out with a time step of 2 fs on both target (complex) structures. Position restraints were applied on solute heavy atoms at a force constant of 10^3 kJ mol^{-1} nm^{-2} to limit their movement and make the hydration spots directly comparable. For temperature-coupling the velocity rescale and the Parrinello–Rahman algorithms[59–61] were used. Solute and solvent were coupled separately with a reference temperature of 300 K and a coupling time constant of 0.1 ps. Pressure was coupled using the Parrinello–Rahman algorithm and with a coupling time constant of 0.5 ps, compressibility of 4.5×10^{-5} bar^{-1} and reference pressure of 1 bar. Particle Mesh–Ewald summation was used for long range electrostatics. van der Waals and Coulomb interactions had a cut-off at 11 Å. Coordinates were saved at regular time-intervals of 1 ps yielding 1×10^3 frames. Periodic boundary conditions were treated before analysis to make the solute whole and recover hydrated solute structures centered in the box. Each frame was fit to the original protein crystal structure using C$_\alpha$ atoms to make them suitable for MobyWat. The final trajectory including all atomic coordinates of all frames was converted to portable binary files.

Calculation and optimization of interfacial water positions

The interface water positions were calculated[56] from the above MD trajectories with the all-inclusive identity based (IDa) algorithm of the program MobyWat.[62] The maximum distances from the target (dmax), prediction (ptol) and clustering tolerances (ctol) were set to 5, 2.5 and 1 Å, respectively. MobyWat calculation of interface water molecules was performed in two main steps. In the first main step, surface water positions were calculated from MD trajectories of the hydrated target molecule. For this, MobyWat first selected candidate water molecules for all frames based on a desired distance limit (dmax) from the target, then an occupancy list was constructed containing every different water ID on different lines in decreasing order with respect to the number of occurrences among all frames. Clustering was applied to all rows (all different water IDs) of the occupancy list using the ctol parameter to define the distance between elements of the same cluster. The largest cluster was selected from all clusters to give the first predicted water oxygen atom by averaging the spatial coordinates of included molecules. In the next steps, clusters were selected in a descending order of their sizes, and checked if their distance was larger than the prediction tolerance (ptol) from previously predicted water oxygen positions to give further predicted oxygen positions until all clusters are done. Finally target, ligand and predicted water oxygen atoms were merged together excluding clashing water oxygens with the Editing mode of MobyWat to produce the starting for the second main step.

In the second main step, the whole previous procedure was repeated for the assembled target–ligand complex equipped with predicted water oxygens (surface) from the first step including hydrogenation, minimization, molecular dynamics and MobyWat prediction with dmax, ptol and ctol set to 5, 3, and 3 Å, respectively, to obtain an appropriately hydrated interface.

As a final MM optimization of the hydrated complex, hydrogen atoms were added to the system (including interface water oxygens from MobyWat) and energy minimization was performed according to a four-step protocol[56] of sd-cg-sd-cg pattern with parameters of sd and cg methods described previously. During the first two steps, all solute heavy atoms and the oxygen of the predicted interface water molecules were position restrained and bulk waters and ions were released. In the last two steps, position restraints were not applied on predicted waters, only on solute heavy atoms.

Single point QM calculation

The interface regions of the above MM-optimized, hydrated complex systems were extracted using the Fragmenter server.[52] In the interface region, the Fragmenter retained peptide fragments of the target together with bound ligand and interfacial water molecules. In this way, the calculation time can be reduced, using the extracted interface instead of the whole complex. In the Fragmenter step, the target–ligand and solute-water cut-off distances were both set to 5 Å and edges were equipped with acetyl (N-terminal) or methylamino (C-terminal) groups.

$\Delta_f H$ values were calculated with Mopac 2016[63] using PM7[64] PM6-DH2X[65,66] and PM6-D3H4X[65,66] parametrizations. $\Delta_f H$ values were separately calculated for the extracted, hydrated complex interface, the target fragments, and the ligand in their bound conformations. Only one self-consistent field iteration was applied with the 1SCF keyword. Charges of input structures were assigned manually as the sum of consisting amino acid and other charged groups (Table S2, ESI†). The non-vacuum calculations were performed utilizing the COSMO[67] solvation model setting relative permittivity to 78.3 as a specific value for aqueous medium. Calculation of larger systems (number of atoms >1000) with COSMO also requires the faster localized molecular orbital approach of MOZYME[68] in Mopac. Notably, in the case of single-point calculation, the results are essentially identical with and without MOZYME. $\Delta_f H$ of a single water molecule was also calculated for the use in eqn (1). $\Delta_f H(H_2O)$ of −57.65, −65.03, and −65.20 kcal mol^{-1} values were obtained in the case of vacuum, COSMO solvent model after MM minimization, and after QM minimization, respectively. All files of single point QM calculations and raw $\Delta_f H$ values are available online (see the section on data availability).

QM minimization

For comparison with the results of single point calculations, full QM energy minimization was also performed on the molecules described in the previous section using PM7 parametrization with MOZYME, COSMO solvation and L-BFGS optimization algorithm.[63] The gradient norm for all calculations was set to 3 kcal mol^{-1} Å^{-1} (keyword GNORM) which is a reasonable limit taking into consideration that it is usually reachable even for bigger systems, furthermore $\Delta_f H$s do not change significantly (<1%) after the gradient drops below 3. In some cases, the calculations were stopped even before reaching the gradient limit, and $\Delta_f H$ became stationary. All files of QM minimizations and raw $\Delta_f H$ values are available online (see the section on data availability).

Generation of the descriptor pool

Descriptors were generated (Fig. 2b) for all ligands with the PaDEL-Descriptor program package.[69] Besides the original ''parents'', additional ''daughter'' descriptors were also derived by normalization with the total number of atoms (NA) in the molecule. 1444 unique descriptors were generated and further investigated. The generation of 814 descriptors was incomplete as the program was unable to produce values for all ligands in their cases. These incomplete (na/infinity) or all-zero descriptors were excluded from statistical analyses. The remaining descriptors were further filtered into a sub-set if they had a correlation of an R^2 of at least 0.09 ($|r| \geq 0.30$) with ΔG_b, and a correlation of a maximal R^2 of 0.04 ($|r| < 0.20$) with $\Delta_r H$. The $R^2 = 0.09$ threshold was chosen according to the 95% confidence interval for r is $[-0.30,0.30]$ assuming H_0: $r = 0$ according to Fisher's method[70] for $N = 43$ assuming normal distribution. From among the remaining 31 descriptors (Fig. 2b) further ones lacking generalizability with very specific atom and ring counts were omitted. Finally the best 3 descriptors were selected based on minimal correlation with $\Delta_r H$ minimizing collinearity for regression analysis (Table S3, ESI†). For the historical record, we would mention that program Codessa (ver. 2.20)[71,72] was also applied on the full data set for initial guessing of the ligand-based descriptor as a second variable in the ΔG_b calculator equation. The MW/NA descriptor was identified (where MW is the molecular weight) indeed with Codessa and the (successful) idea of the above normalization of parent descriptors by NA was inspired by this finding. Notably, we could not use the executable of this ''old'' version of Codessa on MS Windows, but it still worked under the wine-4.5 windows emulator of Fedora Core Linux 29 with somewhat limited functionality.

Statistics

Simple (eqn (2)) or multiple (eqn (3)) linear regression analyses were performed with ($\beta \neq 0$) and without ($\beta = 0$) intercept between the calculated (predictor, x) and experimental (outcome, y) variables including all investigated systems. The program R[73] was applied for all analyses. The goodness of fit was measured by the coefficient of determination (R^2) and the stability of the models was featured by leave-one-out cross-validated R^2 values. The global error of the models was expressed as the root mean squared error (RMSE). Significance (p-values) of the regression coefficients (α and β) was calculated by a two-sided t-test at 0.05 significance level. Further detailed statistical parameters including residuals (ε) and errors of the regression coefficients can be found in the respective tables in the ESI.†

A stepwise, systematic procedure was followed for the exclusion of the outliers. The procedure started with fitting the linear model ($\beta \neq 0$) to the full dataset and the data point with the largest absolute residual (ε) was excluded. The fitting and exclusion steps were repeated for the reduced data sets until the RMSE dropped below a pre-set limit of 2.65 kcal mol^{-1} in the case of model Hybrid 1. For comparability, the same number of data points (N_o = 12 in Table S5, ESI†) were left out in all tailored sets of all ΔH_b models. In the case of ΔG_b models, an RMSE limit of 1.00 kcal mol^{-1} was applied for the termination of the procedure and finally the same 5 data points were excluded in the cases of both descriptors in Table S14 (ESI†).

Results and discussion

Systems and models

The main criteria of the collection of a set of target–ligand complexes (Table S1, ESI†) were the availability of both experimental ΔH_b and ΔG_b values from high quality ITC measurements, and atomic resolution structure of the same complex system in the Protein Databank (PDB).[74] Thus, our results are based on experimental data both on the input (structure) and output (ΔH_b and ΔG_b) sides. Additional structural calculations using MobyWat[56,62] contributed optimized water positions.

Notably, the number of systems fulfilling the above selection criteria is rather limited especially at large, negative ΔH_b values that are essential for a solid correlation with a large data range. Finally, a set of $N = 43$ systems (Table S1, ESI†) were collected also including peptide ligands and 15 systems from our previous study.[26]

For each complex, an end-point approach was applied for the estimation of ΔH_b from the calculated enthalpy of reaction ($\Delta_r H$, eqn (1)) of the binding process. A simple linear regression (eqn (2), where α and β are regression coefficients, ε is the residual, and N is the count of complex systems) was used as a statistical model.

$$\text{Target}[\text{H}_2\text{O}]_x + \text{ligand}[\text{H}_2\text{O}]_y + z\ \text{H}_2\text{O} = \text{target:ligand}[\text{H}_2\text{O}]_{x+y+z}$$

$$\Delta_r H = \Delta_f H(\text{target:ligand}[\text{H}_2\text{O}]_{x+y+z}) - \Delta_f H(\text{target}[\text{H}_2\text{O}]_x) - \Delta_f H(\text{ligand}[\text{H}_2\text{O}]_y) - z\ \Delta_f H(\text{H}_2\text{O}) \qquad (1)$$

$$\Delta H_{b,i}\ (\text{exp}) = \Delta H_{b,i}\ (\text{calc}) + \varepsilon_i = \alpha \Delta_r H_i + \beta + \varepsilon_i;\ i = 1, 2, \ldots, N \qquad (2)$$

$\Delta_r H$ was obtained from the calculated enthalpy of formation ($\Delta_f H$) values of the reaction partners according to Hess's law (eqn (1)). The energy-minimized, hydrated,[56,62] and extracted interface structures (Fig. 1) were used for calculation of $\Delta_f H$ values at the semi-empirical QM level with PM7 parametrization.[26] Interface structures from both MM and QM energy-minimizations were used for calculation of $\Delta_f H$. In the case of MM-minimized structures, a single-point QM calculation of $\Delta_f H$ was performed (referred to as 1SCF calculations in the present study). Hydration effects were included by the COSMO implicit[67] and hybrid (implicit + explicit[26]) water models. Three hybrid models were investigated depending on the assignation of explicit water molecules (Fig. S1, ESI†). Vacuum calculations were also performed for comparison. The technical details of the calculations are described in the Methods section, the technical details of the protocol, the resulted raw data files including input and output structures, and $\Delta_f H$ and $\Delta_r H$ values are available online (see the section on data availability).

Comparison of binding enthalpy models

In the first part of the study, the MM-minimized structures and 1SCF calculations were used to produce $\Delta_r H$ for the linear regressions of eqn (2) without intercept ($\beta = 0$). The final results for a tailored set ($N = 31$) in terms of the squared correlation coefficient (R^2) and the root mean square error (RMSE) values are shown in Fig. 3a. It was found that the hybrid water model outperformed the implicit and explicit water models alone, and vacuum calculations resulted in very weak correlations if any. The best statistics (an RMSE of 2.84 kcal mol^{-1} and an R^2 of 0.93, Table S4, ESI†) were obtained for the Hybrid 1 model. This is comparable to other results in the literature based on a similar number of data points (Table 1, ESI†). The tailored set was constructed by a systematic exclusion of data points using a stepwise re-fitting procedure (Methods; Table S6, ESI†) from the full set ($N = 43$). This exclusion of data points of large residuals (ε_i in eqn (1)) is quite common and necessary to avoid any unwanted influence of these points biasing the least squares fit regression results. Large residuals may originate

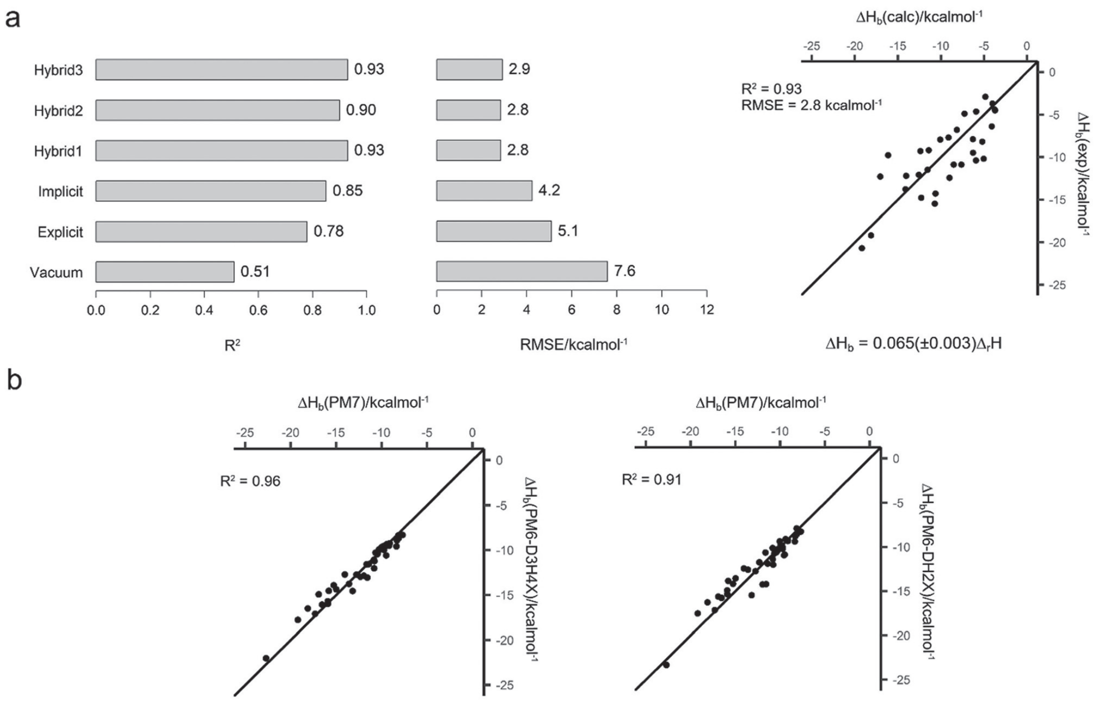


Fig. 3 (a) Comparison of the performance of hydration models used for single point calculation of binding enthalpy. R^2 (left) and RMSE (middle) values of linear regressions ($\beta = 0$) between calculated and experimental binding enthalpies are shown for the tailored ($N = 31$) data set. On the right the correlation plots between calculated and experimental binding enthalpies are also shown for the best model (Hybrid 1). Regression equations with the value $\alpha(\pm\Delta\alpha)$ are displayed below. (b) Comparison of the results obtained using different semi-empirical parametrizations. Correlation plots between binding enthalpies calculated (single point QM, Hybrid 3 model, $\beta = 0$, $N = 43$) with PM7 and PM6-D3H4X (left), or PM6-DH2X (right) methods.

from the inaccuracy of the models and/or the error of isothermal titration calorimetric (ITC) measurements of the experimental ΔH_b values used for the linear regression (eqn (2)). Although ITC is the most reliable source of available binding thermodynamics data, the error of ITC measurements can be as large as 20% of the measured value.[54,75] This relatively large experimental error is often rooted in the inaccuracy of determination of protein (target) concentration, which alone can be as large as 10% as shown by an inter-laboratory study.[76] Analytical studies show that the most reliable and advised protein quantification methods are based on amino acid analysis.[77,78] Unfortunately, in ITC studies, amino acid analysis has been very rarely (almost never) considered and the much less accurate[78] direct single UV spectrophotometric measurement has been applied instead, which was also the case (Table S6, ESI†) for the ITC data used in the present study. Thus, a 20% error of ΔH_b can be expected which can be as large as 4–5 kcal mol^{-1} if considering the system with the largest ΔH_b in our study (Table S1, ESI†). The results for the full set (N = 43, Table S7 and Fig. S2, ESI†) also emphasize the best performance of the hybrid models. It is important that the regression coefficients (α in eqn (2)) obtained for the full (Fig. S2, ESI†) and tailored (Fig. 3a) sets, respectively, have less than 15% difference in their values also indicating the stability of the linear model obtained. Simple linear regressions were also performed with an intercept ($\beta \neq 0$) and resulted in similar trends as the above results with $\beta = 0$. However, the relatively high error of β (Tables S8 and S9, ESI†) shows that the use of the intercept does not increase the significance and quality of the models that was also concluded in our previous study.[26]

We also investigated if the time-consuming QM minimization step formerly used in our[26] and other[31] studies further improves the results or not. Thus, QM minimizations of the extracted interfaces were performed for all systems (see the Methods section). A comparison of the statistical results of the 1SCF and QM-minimized approaches does not show (Table S10, ESI†) any improvement after QM minimization in the ΔH_b calculation.

Finally, a comparison of the performance of different semi-empirical QM parametrizations was performed, and the 1SCF calculations were repeated using PM6-DH2X[65,79,80] and PM6-D3H4X[65,66,79] parametrizations which were reported[50] to produce slightly better intermolecular interaction energy values than PM7. The correlation plots (Fig. 3b) and statistical parameters (Table S11, ESI†) show that there is no significant difference in the 1SCF calculation results for the full and tailored sets between the performance of PM7 and the other two semi-empirical QM parametrizations. These results further emphasize the reproducibility and robustness of the above ΔH_b calculation protocol based on a single-point MM minimization followed by 1SCF QM scoring at the semi-empirical level.

Calculation of binding free energy

Accurate ΔG_b values are inevitable for the selection of the best drug candidates and also direct the computational docking procedures (see the Introduction section). However, ΔG_b calculators are often challenged by the missing or incorrectly assigned water structure around the ligand molecule. As the Hybrid 1 model provided good correlations with experimental ΔH_b in the previous section, we speculated if the calculated, QM-based $\Delta_r H$'s can be combined with ligand-based descriptors for the construction of new ΔG_b calculators similarly to previous, MM-based scoring function developments.[16,17,19] For this, multiple linear regressions (MLR's) were carried out systematically using $\Delta_r H$ (from the Hybrid 1 model) as a first variable (eqn (3) where α_1, α_2, and β are regression coefficients, ε is the residual, and N is the count of complex systems). The second variable (D_L in eqn (3)) was selected from among 1444 ligand-based descriptors through systematic filtering of the descriptor pool (Methods; Table S12, ESI†) resulting in 31 descriptors (Methods; Table S13, ESI†).

$$\Delta G_{b,i}\,(\text{exp}) = \Delta G_{b,i}\,(\text{calc}) + \varepsilon_i = \alpha_1 \Delta_r H_i + \alpha_2 D_{Li} + \beta + \varepsilon_i;\ i = 1,2,\ldots,N \quad (3)$$

The final hits were shortlisted (Table S13, ESI†) and the top descriptor NHA/NA was finally selected as D_L according to its significance, generalizability, and predictive power. NHA/NA is a ''daughter'' obtained by normalization of a ''parent'' descriptor (the number of heavy atoms, NHA) by the molecular size (represented by the total number of atoms, NA). This normalization is particularly important as the size-dependent (parent) descriptors tend to correlate with ΔG_b[81,82] and also with $\Delta_r H$ (Table S3, ESI†), as an increased ligand size often results in more interactions with the target. However, a good MLR model should avoid inter-correlation of the variables of eqn (3). This requirement can be fulfilled by the daughter descriptor NHA/NA that is size-independent and does not correlate with $\Delta_r H$ but is still descriptive for ΔG_b (Fig. 4; Table S3, ESI†).

High correlations between NHA/NA and other descriptors (Table S3, ESI†) suggest that other D_L-s also predict ΔG_b. However, NHA/NA is the simplest choice as it can be readily calculated from the chemical composition of the ligand. The regressions of eqn (3) were performed with both $\beta = 0$ and $\beta \neq 0$ (Table S14, ESI†). After leaving out 5 outliers, the correlation plots show an even distribution of data points around the main diagonal indicating that our ΔG_b calculators based on NHA/NA (Fig. 4a and Table S14, ESI†) predict the experimental values well. The $\beta \neq 0$ model showed overall better statistics and was selected as the final equation for the new calculator named QMH-L. (Abbreviations QMH and L refer to the QM-based calculation of ΔH_b, and the ligand-based descriptor, respectively.) The regression metrics of our calculators (Fig. 4 and Table S14, ESI†) are comparable to other studies using combined QM-MM procedures augmented by PB(GB)SA solvation.[83] The RMSE values below 1 kcal mol^{-1} (Fig. 4) can be considered as excellent.

Besides the above good statistical parameters, the physical meaning of the terms of QMH-L can also be fully explained. As it was shown in the previous section, $\Delta_r H$ of eqn (3) accounts for target–ligand interaction (and the ΔH_b component of ΔG_b). The regression coefficients in the final equation (Fig. 4a) reflect the expected physical meaning of the variables in the MLR.

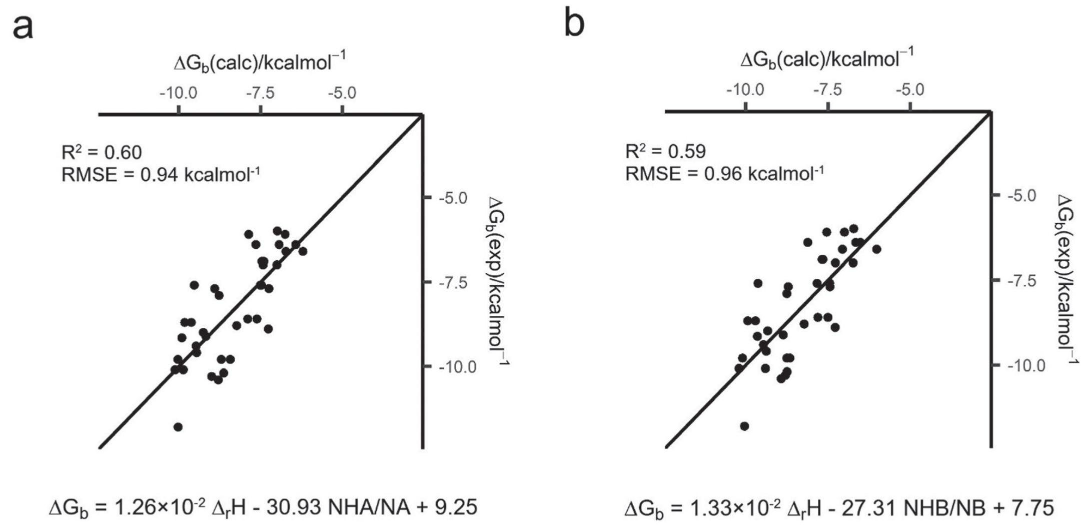
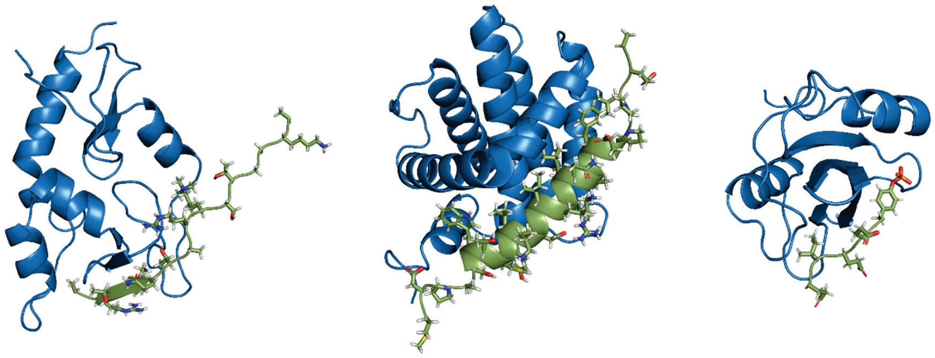


Fig. 4 Correlation plots between calculated and experimental binding free energies. Only Δ_rH (single point QM, PM7, Hybrid 1 model) and an additional ligand-based descriptor (D_L) were used as variables in the regression models ($\beta \neq 0$, eqn (3)). (a) D_L = NHA/NA and (b) D_L = NHB/NB. The regression equations are displayed below the plots, in both cases 5–5 outliers were left out (N = 38; Table S15, ESI†).

The positive sign of α_1 means that a more negative Δ_rH leads to more favourable ΔG_b as expected. As NHA/NA can be considered the density of heavy atoms in a ligand, the negative sign of α_2 implies that a higher density of heavy atoms contributes to a more negative ΔG_b (as NHA/NA $>$ 0 per def.). To understand the physical meaning of NHA/NA and the negative sign of α_2 at a structural level, it is better to start with a related descriptor NHB/NB (the ratio of the number of bonds between heavy atoms and the number of all bonds) that can be considered as the density of bonds between heavy atoms. NHB/NB



ID	3ql9	2rod	1hcs
Ligand	ARTKQTAR-Kme3-STGGKA	AELPPEFAAQLRKIGDKVYCTWSAPD	Ac-pY-EEIE-NH2
MW	1607.9	2905.3	797.2
Δ_rH (kcal/mol)	-215.6	-294.7	-124.9
NHA/NA	0.4667	0.5062	0.5670
NHB/NB	0.4644	0.5134	0.5670

Fig. 5 The structure of three selected peptide ligands (green sticks and cartoon) in complex with their targets (blue cartoon). System IDs and values of ligand-based descriptors are also tabulated. Kme3 and pY refer to a tri-methylated lysine, and phosphorylated tyrosine residue, respectively. While Δ_rH depends on molecular size (the larger the MW, the more negative the Δ_rH), daughter descriptors NHA/NA and NHB/NB do not show correlation with molecular size and Δ_rH due to the normalization by NA. Long alkyl chains of K and R residues (3ql9) result in smaller NHA/NA if compared with the relatively compact, thin ligand of system 1hcs containing side-chains with dehydrogenated, short E's and a ring (in Y). In terms of NHA/NA, the side-chain of K has a butylene ($-C_4H_8-$) group, which means a 1/3 contribution with a weight of 4 to the overall NHA/NA, while in E, the contribution of the ethylene group is only 2 × 1/3 and E also has a carboxylate of a large 3/3 contribution. In Y, the phenylene ring also has a large contribution of 3/5.

correlates with NHA/NA very well ($R^2 = 0.9$) and also provides a good correlation with ΔG_b (Fig. 4b). Sample systems with various ligands and descriptor values are shown in Fig. 5 and a full list of descriptor values of all systems is provided in Table S15 (ESI†). A comparison of the systems (Fig. 5) shows that peptide ligands (Fig. 5) with several long side-chains (residues *K* or *R* in system 3ql9) and long alkyl chains with numerous H atoms have a small NHB/NB (or NHA/NA) while others with compact and/or restrained side-chains, rings, unsaturated, aryl, or carboxylate groups with no or fewer H atoms adopt intermediate (0.5 in system 3ptb) or large (system 1hcs) values. Thus, a higher NHB/NB (or NHA/NA) means a lower occurrence of freely rotatable alkyl- or other (massively hydrogenated) groups and describes a smaller loss of internal rotational degrees of freedom of the ligand during binding to the target. A value of 1/3 may be considered as a realistic lower limit of NHA/NA which corresponds to an alkylene group (C_nH_{2n}). All-in-all, a higher NHB/NB (or NHA/NA) corresponds to a smaller decrease in molecular entropy (S), that is, a smaller negative contribution to ΔS_b, resulting in a larger negative contribution to ΔG_b (due to the negative sign of α_2, Fig. 4). Besides internal rotations, the loss of rotational S of the entire molecule is also smaller for compact and thin molecules of higher NHA/NA. Consequently, NHA/NA can be considered as an entropic term describing the contribution of frozen rotations to ΔG_b, which is often approximated by the number of free torsions (N_{tor}) in MM-based scoring functions. However, N_{tor} depends on the ligand size (MW or NA), and therefore, it easily yields meaningless positive ΔG_b values for large, flexible ligands like peptides, and erroneously correlates with the enthalpic term (see the considerations above on the unwanted inter-correlation of terms in the ΔG_b equation). As NHA/NA is size-independent, such limitations will not restrict its use, and therefore, it can be considered to complement the enthalpic terms (Coulomb, Lennard-Jones) in MM-based scoring functions similar to the present case of $\Delta_r H$ (Fig. 3a).

Conclusions

There are various MM-based methods available for calculation of binding thermodynamics of target–ligand complexes. However, MM calculators can rarely handle electronic effects of highly charged or polarizable systems. Interfacial water is often involved in such effects due to its high dipole moment, further complicating the estimation of target–ligand interactions at the MM level. While QM methods are expected to give a solution for this problem, QM optimization often requires enormous computer time for large target–ligand complexes. In the present study, single point QM calculations and a hybrid water model were tested on various systems including protein targets and challenging, large peptide ligands with a MW up to 2–3 thousand (Fig. 5). The fragmented complex structures from fast MM-based optimization were piped into single point QM calculations of ΔH_b. The precision of the present ΔH_b calculator is comparable to that of full length QM optimizations. At the same time, the present approach reduced the computational cost at least by three orders of magnitude if compared with the full length QM minimizations. Binding data of the highest quality ITC measurements were used for validation, and therefore, the test of the robustness of the calculators was possible separately for ΔH_b and ΔG_b. The QM-calculated $\Delta_r H$s were first tested for prediction of ΔH_b, and then adopted for further development of the ΔG_b calculators. This successive building of the calculators allowed the discovery and the elucidation of the physical meaning of the descriptor NHA/NA which can be simply obtained from the chemical composition of the ligand, and completed the final calculator called QMH-L. Thus, the final ΔG_b equation includes only two variables, the $\Delta_r H$ and a ligand-based descriptor NHA/NA. Notably, NHA/NA may be adopted by other (MM-based) scoring functions as an entropic term complementing the enthalpic terms. The single-point protocol and the calculators developed in the present study offer a good compromise in terms of accuracy, applicability, and computational cost. The calculated ΔH_b and ΔG_b values involve electronic effects at the semi-empirical QM level, and the protocol does not need further parametrization of the ligands which is often a bottle-neck of drug screening projects. The single point protocol requires modest computational resources (under 1 min for the largest systems on a i5-8250U 1.6 GHz quad-core processor) and is based on short runs with open source and/or free software like Gromacs, Mopac, Moby-Wat, and Fragmenter. The calculators can be applied for fast, automated scoring of drug candidates during a virtual screen, engineering of new complexes or thermodynamic explanation of target–ligand interactions.

Author contributions

V. S. performed research, wrote and revised the manuscript. B. Z. Z. and N. J. participated in research. C. H. designed and supervised research, wrote and revised the manuscript.

Data availability

A compressed file including the resulted raw data files, input and output structures, $\Delta_f H$ and $\Delta_r H$ values supporting the findings of this study. A pdf file describing the technical details of the protocol. A compressed file including sample files to the description of the protocol. All files were deposited as compressed archives at **https://zenodo.org/records/10055343**.

List of abbreviations

1SCF	Single point self-consistent field
AM1	Austin Model 1
cg	Conjugate gradient
COSMO	Conductor-like screening model
DFT	Density-functional theory
ctol	Clustering tolerance
dmax	Maximum distance

HF	Hartree–Fock
ITC	Isothermal titration calorimetry
MD	Molecular dynamics
MLR	Multiple linear regression
MM	Molecular mechanics
mtol	Match tolerance
MW	Molecular weight
NA	Number of atoms
NB	Number of bonds
NHA	Number of heavy atoms
NHB	Number of bonds between heavy atoms
N_{tor}	Number of torsions
PDB	Protein Databank
PM3	Parametric Method 3
ptol	Prediction tolerance
QM	Quantum mechanics
RMSE	Root mean square error
sd	Steepest descent

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The development teams of Gromacs and Mopac are acknowledged for the release and maintenance of their open source program packages. C. H. is thankful to former colleague and supervisor Prof. Mati Karelson for his kind gift of a life-long license (no. 06031) of Codessa (ver. 2.20). The authors acknowledge the support from the Governmental Information Technology Development Agency, Hungary. The project has been supported by the European Union, co-financed by the European Social Fund. Project name and code: Comprehensive Development for Implementing Smart Specialization Strategies at the University of Pécs, [EFOP-3.6.1-16-2016-00004]. The authors acknowledge that the results of this research have been achieved using the Distributed Computing Initiative resource Archer2 based in the United Kingdom at the National Supercomputing Service with support from the Partnership for Advanced Computing in Europe aisbl.

References

1 P. H. M. Torres, A. C. R. Sodero, P. Jofily and F. P. Silva-Jr, Key topics in molecular docking for drug design, *Int. J. Mol. Sci.*, 2019, **20**(18), 1–29.

2 C. N. Cavasotto, M. G. Aucar and N. S. Adler, Computational chemistry in drug lead discovery and design, *Int. J. Quantum Chem.*, 2019, **119**(2), 1–19.

3 X. Lin, X. Li and X. Lin, A review on applications of computational methods in drug screening and design, *Molecules*, 2020, **25**(6), 1–17.

4 X. Du, Y. Li, Y. L. Xia, S. M. Ai, J. Liang and P. Sang, *et al.*, Insights into protein–ligand interactions: Mechanisms, models, and methods, *Int. J. Mol. Sci.*, 2016, **17**(2), 1–34.

5 V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis and E. Kazlauskas, Binding affinity in drug design: experimental and computational techniques, *Expert Opin. Drug Discovery*, 2019, **14**(8), 755–768.

6 N. Brooijmans and I. D. Kuntz, Molecular recognition and docking algorithms, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 335–373.

7 R. Claveria-Gimeno, S. Vega, O. Abian and A. Velazquez-Campoy, A look at ligand binding thermodynamics in drug discovery, *Expert Opin. Drug Discovery*, 2017, **12**(4), 363–377.

8 A. Schön and E. Freire, Enthalpy screen of drug candidates, *Anal. Biochem.*, 2016, **513**, 1–6.

9 C. Baggio, P. Udompholkul, E. Barile and M. Pellecchia, Enthalpy-Based Screening of Focused Combinatorial Libraries for the Identification of Potent and Selective Ligands, *ACS Chem. Biol.*, 2017, **12**(12), 2981–2989.

10 N. S. Pagadala, K. Syed and J. Tuszynski, Software for molecular docking: a review, *Biophys. Rev.*, 2017, **9**(2), 91–102.

11 V. B. Sulimov, D. C. Kutov and A. V. Sulimov, Advances in Docking, *Curr. Med. Chem.*, 2018, **26**(42), 7555–7580.

12 A. R. Ortiz, M. T. Pisabarro, F. Gago and R. C. Wade, Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis, *J. Med. Chem.*, 1995, **38**(14), 2681–2691.

13 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.*, 1997, **267**(3), 727–748.

14 G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart and R. K. Belew, *et al.*, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, *J. Comput. Chem.*, 1998, **19**(14), 1639–1662.

15 O. Troot and A. J. Olson, Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading, *J. Comput. Chem.*, 2010, **31**, 16.

16 C. Hetényi, G. Paragi, U. Maran, Z. Timár, M. Karelson and B. Penke, Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins, *J. Am. Chem. Soc.*, 2006, **128**(4), 1233–1239.

17 R. Rajamani and A. C. Good, Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development, *Curr. Opin. Drug Discovery Dev.*, 2007, **10**(3), 308–315.

18 A. N. Jain, Scoring functions for protein-ligand docking, *Curr. Protein Pept. Sci.*, 2006, **7**(5), 407–420.

19 M. H. J. Seifert, Targeted scoring functions for virtual screening, *Drug Discovery Today*, 2009, **14**(11–12), 562–569.

20 E. Yuriev, M. Agostino and P. A. Ramsland, Challenges and advances in computational docking: 2009 in review, *J. Mol. Recognit.*, 2011, **24**(2), 149–164.

21 D. Santos-Martins, L. Solis-Vasquez, A. F. Tillack, M. F. Sanner, A. Koch and S. Forli, Accelerating Auto

Dock 4 with GPUs and Gradient-Based Local Search, *J. Chem. Theory Comput.*, 2021, **17**(2), 1060–1073.

22 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu and Y. Li, *et al.*, Comparative Assessment of Scoring Functions: The CASF-2016 Update, *J. Chem. Inf. Model.*, 2019, **59**(2), 895–913.

23 S. Ehrlich, A. H. Göller and S. Grimme, Towards full Quantum-Mechanics-based Protein–Ligand Binding Affinities, *Chem. Phys. Chem.*, 2017, **18**(8), 898–905.

24 D. Scaramozzino and P. M. Khade Applied sciences Protein Fluctuations in Response to Random External Forces. 2022.

25 B. Z. Zsidó and C. Hetényi, The role of water in ligand binding, *Curr. Opin. Struct. Biol.*, 2021, **67**, 1–8.

26 I. Horváth, N. Jeszenői, M. Bálint, G. Paragi and C. Hetényi, A fragmenting protocol with explicit hydration for calculation of binding enthalpies of target-ligand complexes at a quantum mechanical level, *Int. J. Mol. Sci.*, 2019, **20**(18), 4384–4403.

27 C. N. Cavasotto and M. G. Aucar, High-Throughput Docking Using Quantum Mechanical Scoring, *Front. Chem.*, 2020, 8.

28 E. Nikitina, V. Sulimov, V. Zayets and N. Zaitseva, Semiempirical Calculations of Binding Enthalpy for Protein-Ligand Complexes, *Int. J. Quantum Chem.*, 2004, **97**(2), 747–763.

29 K. Raha and K. M. Merz, Large-Scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes, *J. Med. Chem.*, 2005, **48**(14), 4558–4575.

30 E. Nikitina, V. Sulimov, F. Grigoriev, O. Kondakova and S. Luschekina, Mixed implicit/explicit solvation models in quantum mechanical calculations of binding enthalpy for protein–ligand complexes, *Int. J. Quantum Chem.*, 2006, **106**(8), 1943–1963.

31 J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka and P. Hobza, A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands, *J. Phys. Chem. B*, 2010, **114**(39), 12666–12678.

32 P. Dobeš, J. Fanfrlík, J. Řezáč, M. Otyepka and P. Hobza, Transferable scoring function based on semiempirical quantum mechanical PM6-DH2 method: CDK2 with 15 structurally diverse inhibitors, *J. Comput.-Aided Mol. Des.*, 2011, **25**(3), 223–235.

33 R. González, C. F. Suárez, H. J. Bohórquez, M. A. Patarroyo and M. E. Patarroyo, Semi-empirical quantum evaluation of peptide – MHC class II binding, *Chem. Phys. Lett.*, 2017, **668**, 29–34.

34 M. Hylsová, B. Carbain, J. Fanfrlík, L. Musilová, S. Haldar and C. Köprülüoğlu, *et al.*, Explicit treatment of active-site waters enhances quantum mechanical/implicit solvent scoring: Inhibition of CDK2 by new pyrazolo[1,5-a]pyrimidines, *Eur. J. Med. Chem.*, 2017, **126**(2017), 1118–1128.

35 A. Pecina, J. Brynda, L. Vrzal, R. Gnanasekaran, M. Hořejší and S. M. Eyrilmez, *et al.*, Ranking Power of the SQM/COSMO Scoring Function on Carbonic Anhydrase II–Inhibitor Complexes, *Chem. Phys. Chem.*, 2018, **19**(7), 873–879.

36 A. Klamt, Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena, *J. Phys. Chem.*, 1995, **99**(7), 2224–2235.

37 C. N. Cavasotto, N. S. Adler and M. G. Aucar, Quantum chemical approaches in structure-based virtual screening and lead optimization, *Front. Chem.*, 2018, **6**(MAY), 1–7.

38 J. Liu, J. Wan, Y. Ren, X. Shao, X. Xu and L. Rao, DOX_BDW: Incorporating Solvation and Desolvation Effects of Cavity Water into Nonfitting Protein–Ligand Binding Affinity Prediction, *J. Chem. Inf. Model.*, 2023, **63**(15), 4850–4863.

39 A. Pecina, R. Meier, J. Fanfrlík, M. Lepšík, J. Řezáč and P. Hobza, *et al.*, The SQM/COSMO filter: Reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation, *Chem. Commun.*, 2016, **52**(16), 3312–3315.

40 H. Ajani, A. Pecina, S. M. Eyrilmez, J. Fanfrlík, S. Haldar and J. Řezáč, *et al.*, Superior Performance of the SQM/COSMO Scoring Functions in Native Pose Recognition of Diverse Protein-Ligand Complexes in Cognate Docking, *ACS Omega*, 2017, **2**(7), 4022–4029.

41 A. Pecina, S. M. Eyrilmez, C. Köprülüoğlu, V. M. Miriyala, M. Lepšík and J. Fanfrlík, *et al.*, SQM/COSMO Scoring Function: Reliable Quantum-Mechanical Tool for Sampling and Ranking in Structure-Based Drug Design, *ChemPlusChem*, 2020, **85**(11), 2362–2371.

42 G. Klebe, Applying thermodynamic profiling in lead finding and optimization, *Nat. Rev. Drug Discovery*, 2015, **14**(2), 95–110.

43 H. Ohtaka and E. Freire, Adaptive inhibitors of the HIV-1 protease, *Prog. Biophys. Mol. Biol.*, 2005, **88**(2), 193–208.

44 I. Bertini, V. Calderone, M. Fragai, A. Giachetti, M. Loconte and C. Luchinat, *et al.*, Exploring the subtleties of drug-receptor interactions: The case of matrix metalloproteinases, *J. Am. Chem. Soc.*, 2007, **129**(9), 2466–2475.

45 E. Freire, Do enthalpy and entropy distinguish first in class from best in class?, *Drug Discovery Today*, 2008, **13**(19–20), 869–874.

46 S. Merighi, C. Simioni, S. Gessi, K. Varani and P. A. Borea, Binding thermodynamics at the human cannabinoid CB1 and CB2 receptors, *Biochem. Pharmacol.*, 2010, **79**(3), 471–477.

47 S. F. L. S. Rocha, Sant'Anna CMR. A procedure combining molecular docking and semiempirical method PM7 for identification of selective Shp2 inhibitors, *Biopolymers*, 2019, **110**(11), e23320.

48 C. A. Ortiz-Mahecha, W. A. Agudelo, M. A. Patarroyo, M. E. Patarroyo and C. F. Suárez, MHCBI: A pipeline for calculating peptide-MHC binding energy using semi-empirical quantum mechanical methods with explicit/ implicit solvent models, *Briefings Bioinf.*, 2021, **22**(6), 1–8.

49 L. Wei, Y. Chen, J. Liu, L. Rao, Y. Ren and X. Xu, *et al.*, Cov_DOX: A Method for Structure Prediction of Covalent Protein–Ligand Bindings, *J. Med. Chem.*, 2022, **65**(7), 5528–5538.

50 J. Hostaš, J. Řezáč and P. Hobza, On the performance of the semiempirical quantum mechanical PM6 and PM7

methods for noncovalent interactions, *Chem. Phys. Lett.*, 2013, **568–569**, 161–166.

51 G. A. Urquiza-Carvalho, W. D. Fragoso and G. B. Rocha, Assessment of semiempirical enthalpy of formation in solution as an effective energy function to discriminate native-like structures in protein decoy sets, *J. Comput. Chem.*, 2016, **37**(21), 1962–1972.

52 A. V. Sulimov, D. C. Kutov, E. V. Katkova and V. B. Sulimov, Combined Docking with Classical Force Field and Quantum Chemical Semiempirical Method PM7, *Adv. Bioinf.*, 2017, **2017**, 7167691.

53 W. Chen, H. He, J. Wang, J. Wang and C. E. A. Chang, Uncovering water effects in protein-ligand recognition: importance in the second hydration shell and binding kinetics, *Phys. Chem. Chem. Phys.*, 2022, **25**(3), 2098–2109.

54 S. Geschwindner, J. Ulander and P. Johansson, Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip?, *J. Med. Chem.*, 2015, **58**(16), 6321–6335.

55 N. Guex and M. C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling, *Electrophoresis*, 1997, **18**(15), 2714–2723.

56 N. Jeszenoi, M. Bálint, I. Horváth, D. Van Der Spoel and C. Hetényi, Exploration of Interfacial Hydration Networks of Target-Ligand Complexes, *J. Chem. Inf. Model.*, 2016, **56**(1), 148–158.

57 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar and R. Apostolov, *et al.*, GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit, *Bioinformatics*, 2013, **29**(7), 845–854.

58 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis and R. O. Dror, *et al.*, Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 8.

59 M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, *J. Appl. Phys.*, 1981, **52**(12), 7182–7190.

60 S. Nosé and M. L. Klein, Constant pressure molecular dynamics for molecular systems, *Mol. Phys.*, 1983, **50**(5), 1055–1076.

61 T. Darden, D. York and L. Pedersen, Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems, *J. Chem. Phys.*, 1993, **98**(12), 10089–10092.

62 N. Jeszenoi, I. Horváth, M. Bálint, D. Van Der Spoel and C. Hetényi, Mobility-based prediction of hydration structures of protein surfaces, *Bioinformatics*, 2015, **31**(12), 1959–1965.

63 J. J. P. Stewart, MOPAC 2016. Stewart Computational Chemistry, 2016.

64 J. J. P. Stewart, Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters, *J. Mol. Model.*, 2013, **19**(1), 1–32.

65 J. Řezáč and P. Hobza, Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods, *J. Chem. Theory Comput.*, 2012, **8**(1), 141–151.

66 J. Řezáč and P. Hobza, A halogen-bonding correction for the semiempirical PM6 method, *Chem. Phys. Lett.*, 2011, **506**(4–6), 286–289.

67 A. Klamt and G. Schüürmann, COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.

68 J. J. P. Stewart, Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations, *Int. J. Quantum Chem.*, 1996, **58**(2), 133–146.

69 C. W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.*, 2011, **32**(7), 1466–1474.

70 B. Rosner, *Fundamentals of biostatistics*, Cengage learning, 2015.

71 A. R. Katritzky, V. S. Lobanov and M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, *Chem. Soc. Rev.*, 1995, **24**(4), 279–287.

72 M. Karelson, V. S. Lobanov and A. R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev.*, 1996, **96**(3), 1027–1044.

73 R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: **https://www.r-project.org/**.

74 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat and H. Weissig, *et al.*, The Protein Data Bank, *Nucleic Acids Res.*, 2000, **28**(1), 235–242.

75 L. Baranauskiene, V. Petrikaite, J. Matuliene and D. Matulis, Titration calorimetry standards and the precision of isothermal titration calorimetry data, *Int. J. Mol. Sci.*, 2009, **10**(6), 2752–2762.

76 J. Tellinghuisen and J. D. Chodera, Systematic errors in isothermal titration calorimetry: Concentrations and baselines, *Anal. Biochem.*, 2011, **414**(2), 297–299.

77 M. Fountoulakis and H. W. Lahm, Hydrolysis and amino acid composition analysis of proteins, *J. Chromatogr. A*, 1998, **826**(2), 109–134.

78 K. Reinmuth-Selzle, T. Tchipilov, A. T. Backes, G. Tscheuschner, K. Tang and K. Ziegler, *et al.*, Determination of the protein content of complex samples by aromatic amino acid analysis, liquid chromatography-UV absorbance, and colorimetry, *Anal. Bioanal. Chem.*, 2022, **414**(15), 4457–4470.

79 J. J. P. Stewart, Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements, *J. Mol. Model.*, 2007, **13**(12), 1173–1213.

80 M. Korth, M. Pitoňák, J. Řezáč and P. Hobza, A Transferable H-Bonding Correction for Semiempirical.pdf, *J. Chem. Theory Comput.*, 2010, **6**(1), 344–352.

81 I. D. Kuntz, K. Chen, K. A. Sharp and P. A. Kollman, The maximal affinity of ligands, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**(18), 9997–10002.

82 A. T. Garcia-Sosa, U. Maran and C. Hetenyi, Molecular Property Filters Describing Pharmacokinetics and Drug Binding, *Curr. Med. Chem.*, 2012, **19**(11), 1646–1662.

83 E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu and J. Z. H. Zhang, *et al.*, End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design, *Chem. Rev.*, 2019, **119**(16), 9478–9508.

hetenyi.csaba_83_23

D28

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 20 2007, pages 2678–2685
doi:10.1093/bioinformatics/btm431

Structural bioinformatics

Structure-based calculation of drug efficiency indices

Csaba Hetényi[1,2,*], Uko Maran[1], Alfonso T. García-Sosa[1] and Mati Karelson[1]

[1]Institute of Chemical Physics, University of Tartu, 2 Jakobi Street, 51014 Tartu, Estonia and
[2]Department of Organic Chemistry, Faculty of Pharmacy, Semmelweis University, Budapest, Hungary

Received on March 17, 2007; revised on August 6, 2007; accepted on August 18, 2007

Advance Access publication September 5, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The efficiency indices (EI's) have been derived from the experimental binding affinities of drug candidates to macromolecules. These 'two-in-one' measures include information on both pharmacodynamics and pharmacokinetics of the candidate molecules. The time-consuming experimental measurement of binding affinities of extensive molecule libraries may become a bottle-neck of large scale generation and application of EI's.

Results: To overcome this limitation, structure-based calculation of new EI's is introduced using the modified free energy function of the popular program package AutoDock. The results are validated on experimental binding data of biochemical systems such as potent inhibitors bound to β-secretase, a key enzyme of Alzheimer's disease and various drug–protein complexes. Application of new EI's is tested. Thermodynamics of EI's and their role in virtual high - throughput screening of drugs and in the development of docking programs are discussed.

Contact: csabahete@yahoo.com

Supplementary information: Accompanies this manuscript on the publisher's web site.

1 INTRODUCTION

The mechanism of drug action generally involves a long chain of interactions with the molecules of the human body. There are numerous experimental and *in silico* drug design tools describing the terminal link of these chains, i.e. the estimation of equilibrium binding affinities (BA) of drug candidates (ligands) to the targeted macromolecules. Although BA is undoubtedly a key property, other pharmacokinetic and non-equilibrium links in the chain such as absorption, distribution and excretion of the candidate molecules also affect drug-likeness (Swinney, 2004, 2006).

Accordingly, most of the current *in silico* molecular design strategies (Lipinski and Hopkins, 2004) include modeling steps for the equilibrium binding and also for the pharmacokinetics of drugs. Atomic level techniques have been introduced for structural calculation of binding in ligand–target complexes. Computational molecular docking (Fig. 1) is the most advanced among these techniques (Brooijmans and Kuntz, 2003). The BA values of the ligands can be calculated directly from docked ligand–protein complex structures with free energy (scoring) functions. Another important step is the optimization of pharmacokinetics and drug-likeness of ligand databases using empirical rules of selection (Lipinski *et al.*, 1997). These rules define limit values of simple, size-dependent molecular descriptors, e.g. the molecular weight (M_{W}) which can be used for filtering of compound databases.

Recently, new measures, the efficiency indices (EI) were introduced (Abad-Zapatero and Metz, 2005; Hopkins *et al.*, 2004) linking the above mentioned different steps of drug design. EI's have promptly gained applications connecting structural diversity and biological activity of drugs (Schuffenhauer *et al.*, 2006) and in optimization of synthetic receptors (Chen *et al.*, 2006). The introduction of EI's was inspired by earlier studies (Kuntz *et al.*, 1999) showing the usefulness of normalization of BA with the number of heavy atoms (N_{HAT}) for drug design purposes.

In EI's, the normalized quantities [Equation (1)] are represented by commonly used measures of BA such as the experimental free energy of binding (ΔG_{E}), the negative logarithms of experimental dissociation constant (pK_{d}), inhibition constant (pK_{i}) or inhibitor concentration at 50% inhibition ($p\mathrm{IC}_{50}$). The above mentioned simple descriptors, i.e. M_W (Abad-Zapatero and Metz, 2005) or $\mathrm{N}_{\mathrm{HAT}}$ (Hopkins *et al.*, 2004) are typical examples of the size-dependent normalizing factors (SNF).

$$\mathrm{EI} = \frac{\mathrm{BA}}{\mathrm{SNF}} \qquad (1)$$

The EI's were originally defined with experimental BA values (Abad-Zapatero and Metz, 2005; Hopkins *et al.*, 2004). However, the use of structure-based, calculated binding free energy (ΔG_{C}) values from scoring (Hetényi *et al.*, 2006) of docked ligand–protein structures instead of ΔG_{E} may become successful alternative for obtaining EI's. Remarkably, computational docking has an advantage of producing atomic level protein–ligand complex structures within reasonable time. The calculation of ΔG_{C} (scoring) is either performed along with the docking calculations or independently in post-docking mode (Fig. 1). In both cases it requires negligible time and, therefore, allows reduction of time-consuming and expensive biochemical measurements of BA's. Picking up the speed of *in silico* docking and scoring, the calculation of EI's can become an essential part of high-throughput, structure-based virtual compound screening and drug design. The aim of the present study is to introduce and investigate rapid calculation of various EI's on the basis of a set of biologically relevant structural and thermodynamic experimental data.

*To whom correspondence should be addressed.

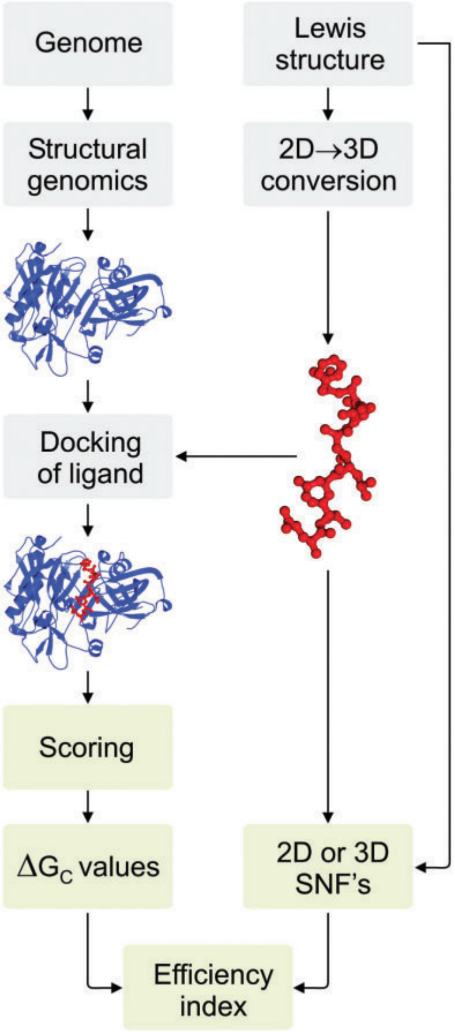


Fig. 1. The pathway of *in silico* drug design connecting genome and drug efficiency. Structural genomics projects generate new protein structures at an unprecedented rate (Yang and Tung, 2006). To efficiently use this increasing amount of 3D information for drug design, high-throughput methods are necessary, which can reduce the complexity of drug (ligand)–protein interactions to comparable measures (indices). The sequence of gray boxes show, that starting from the 2D Lewis structures of a ligand, 3D ligand–protein complexes can be obtained via conversion, modeling and docking. In the present study (beige boxes), a set of biologically relevant ligand–protein complexes were used for calculation of binding free energy (ΔG_C). A representative complex of β-secretase (blue), a key enzyme of Alzheimer's disease and its potent peptidic inhibitor ligand, GluValAsnLeu(Ψ)AlaAlaGluPhe (red) is included in this figure. Further references on the role of β-secretase can be found in works of Hetényi *et al.* (2006) and Hong *et al.* (2000). Both 2D and 3D representation of the ligand molecules can be used for calculation of size-dependent normalization factors (SNF). The ratio of ΔG_C and SNF is the efficiency index, which is a practical 'two-in-one' measure of drug design. This figure was prepared using PyMol (DeLano, 2006).

Table 1. The 20 drug–protein complexes of the external validation set

PDB code	Protein	Drug
1aj6	Gyrase	Novobiocin
1cea	Plasminogen	Aminocaproic acid
1dhi	Dihydrofolate reductase	Methotrexate
1dwc	Alpha-thrombin (small subunit)	Argatroban
1f5l	Urokinase-type plasminogen activator	Amiloride
1fkf	Fk 506 binding protein	Tacrolimus
1h61	Pentaerythitol tetranitrate reductase	Hydrocortisone
1hvy	Thymidylate synthase	Raltitrexed
1hxw	HIV protease	Ritonavir
1j3j	Dihydrofolate reductase	Pyrimethamine
1jt1	FEZ-1, class B3 metallo-beta-lactamase	Captopril
1m2z	Glucocorticoid receptor	Dexamethasone
1odi	Purine nucleoside phosphorylase	Adenosine
1ohr	Aspartylprotease	Nelfinavir
1p62	Deoxycytidine kinase	Gemcitabine
1sqn	Progesterone receptor	Norethindrone
1t7j	Drug resistant HIV protease	Amprenavir
1uw6	Acetylcholine-binding protein	Nicotine
2aou	Histamine N-methyltransferase	Amodiaquine
2gss	Glutathione S-transferase P1-1	Ethacrynic acid

2 METHODS

2.1 Binding data and structure-based free energy calculation of protein–ligand systems

ΔG_E and ΔG_C values of 53 protein–ligand complexes were adopted from a previous study (Hetényi *et al.*, 2006) and listed in Supplementary Material. Proteins having large, peptidic ligands ($M_W > 350$) and physiological importance such as the β-secretase enzyme of Alzheimer's disease (Fig. 1), HIV-1 protease, streptavidin and immunoglobulins were prioritized for the study. The atomic coordinates of 41 of the complexes, were obtained from the Protein Databank (PDB, Berman *et al.*, 2000). The 12 β-secretase- inhibitor systems (om12, om13, om14, om15, om16, om17, om18, om19, om22, om23, om24 and om99-1) with no PDB structures available were modeled by modification of the 1fkn structure. Details on the systems, modeling and minimization of the complexes can be found in the previous paper (Hetényi *et al.*, 2006). Although the peptidic ligands of these systems may become excellent lead compounds, they cannot be considered as drugs (Rishton, 2003). Thus, a set of an additional 20 drug–protein complexes (Table 1) having both PDB structures and ΔG_E values was collected and used in the external validation and application tests of the new EI's introduced in this study. The sources and the procedure of collection of these data are described in details in the Supplementary Material. Altogether the 53 + 20 ligands represent a wide range of compounds including larger, lead-like non-drugs and actual drugs.

The ΔG_C's were calculated using the minimized protein–ligand complexes, according to the modified AutoDock 3.0 (AD3, Morris *et al.*, 1998) and AutoDock4 (AD4, Huey *et al.*, 2007) scoring functions [Equation (2)].

$$\Delta G_C(\mathrm{AD3}) = \underbrace{f_{\mathrm{elec}} \sum_{i,j} \frac{q_j q_i}{\varepsilon(r_{ij}) r_{ij}} + f_{\mathrm{vdw}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)}_{\Delta H_C}$$

$$+ \underbrace{f_{\mathrm{hbond}} \sum_{i,j} \xi(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right)}_{\Delta H_C} + \underbrace{f_{\mathrm{sol}} \sum_{i,j} S_i V_j e^{\left(\frac{r_{ij}^2}{2\sigma^2} \right)}}_{\Delta G_{\mathrm{s,C}}} \quad (2)$$

The f coefficients were determined empirically from a multi-linear regression (MLR) to a set of 30 protein–ligand complexes (AutoDock calibration set) with known binding constants (Morris *et al.*, 1998). The indices i and j correspond to ligand and protein atoms, respectively. The Coulombic term includes the partial charges (q) and a distance-dependent dielectric function (ε) (Morris *et al.*, 1996). A, B, C and D are the Lennard–Jones parameters in the dispersion/repulsion 12-6 and H-bonding 12-10 formulas and r denotes the distance between the atomic pairs. $\xi(t)$ is a directional weight depending on angle t at the H-bonds. S and V denote the solvation parameter and fragmental volume, respectively, in the solvation function of Stouten *et al.* (1993). In the scoring function of AutoDock 3.0, only the C atoms of the ligand molecules are involved in the solvation model. The exponential term is an envelope function with a constant-value of $\sigma = 3.5$ Å. For simplicity, the sum of Coulombic and Lennard–Jones (enthalpic) terms is marked as ΔH_C and the last, desolvation term is marked as $\Delta G_{s,C}$. Remarkably, the AutoDock4 scoring function has different parametrization of the ΔG_C(AD4) part, especially for the desolvation term. Details on the new AD4 scoring function can be found in the original paper of Huey *et al.* (2007). In the present study, all systems were re-scored using the epdb command of AutoDock4. Besides the ΔG_C(AD4), i.e. the intermolecular enthalpic + desolvation terms, the full AD4 binding free energy [ΔG_{full}(AD4)] was also calculated and checked for applicability in EI calculations.

2.2 Regression analyses

The LR's were statistically analyzed and the SNF values were obtained using the program package CODESSA (ver. 2.0) (Karelson *et al.*, 1996; Katritzky *et al.*, 1995). Results of the regression analyses, i.e. mean square errors and t-values of the regression coefficients, the F-values, and the squares of the correlation coefficients (r^2) of the regressions are tabulated in section Results. The principal moments of inertia were calculated for the binding conformations of ligand molecules using the Analyze program of the TINKER software package (Ren and Ponder, 2003). Numerical values used in the calculations and the correlations of SNF's are tabulated in Supplementary Material.

3 RESULTS

3.1 Definition of new EI's

The list and definitions of the SNF's [Equation (1)] corresponding to new, and formerly published (Abad-Zapatero and Metz, 2005; Hopkins *et al.*, 2004) EI's can be found in Table 2. Some of these SNF's are commonly applied as two-dimensional (2D) descriptors in quantitative structure-activity relationship (QSAR) equations (Devillers and Balaban, 1999) and show relatively large degree of correlation with each other (Supplementary Material). The more complicated SNF's contain information also on molecular complexity involving internal (topological) distances and branching of the ligands resulting in their more unique profile and, in some cases, moderate correlations with each other. Whereas 2D descriptors are derived solely from the Lewis formula, i.e. the empirical connectivity list or molecular graph of the ligands, calculation of ΔG_C requires the knowledge of spatial atomic positions in the protein–ligand complex. In a recent study (Hetényi *et al.*, 2006), it was found that ΔG_E's of even large, flexible peptides (Fig. 1) can be predicted with a modified scoring function (ΔG_C) of the docking program package AutoDock 3.0 (Morris *et al.*, 1998). As ΔG_C shows a significant correlation with the ΔG_E values (Hetényi *et al.*, 2006) it was selected to represent BA in the structure-based, calculated EI values throughout the present investigations [BA = ΔG_C in Equation (1)].

3.2 Correlation of experimental and calculated EI's

To test the reliability and predictive value of calculated EI's, simple linear regression (LR) analyses were performed with EI's obtained from the measured ΔG_E values [Equation (3)].

$$\frac{\Delta G_{E,k}}{\text{SNF}_k} = \alpha \frac{\Delta G_{C,k}}{\text{SNF}_k} + \beta + \varepsilon_k \quad (k = 1,2,\ldots,N) \qquad (3)$$

Where α, and β represent the regression coefficient and the intercept, respectively. The ε_k's are the residuals at each data point. The total number of data points (N), i.e. the number of protein–ligand systems adopted from the previous study (Hetényi *et al.*, 2006) was 50. A systematic series of LR's were developed for EI's based on the SNF's of Table 2 and ΔG_C's calculated with the scoring schemes of AutoDock3.0 and AutoDock4, respectively. The results and statistical parameters of the LR's are summarized in Table 3 and in the Supplementary Material.

All LR's are statistically significant, and show higher r^2 values than the correlation ($r^2 = 0.706$) obtained between ΔG_E and ΔG_C (Hetényi *et al.*, 2006). Importantly, the high r^2 values in Table 3 are not trivial consequences of this correlation in the previous work, as the SNF values are different for the 50 different ligand molecules [Equation (3)].

An advantage of 2D descriptors such as the Wiener index (W) involved in the best correlation (Fig. 2; Table 3) is that they can be unambiguously and rapidly calculated from the internal connectivity information coded in the molecular graph (Table 2). For example, W involves a simple summation of shortest topological distances in a molecule. Comparably good correlations could be achieved at all other SNF's including Balaban index (J) which is also defined by internal topological distances (Table 2) and was found to be useful as a QSAR descriptor in prediction of the entropic parts of ΔG_E (Hetényi *et al.*, 2006). In addition, even the three outlier protein–ligand systems (1hhj, om22, om24) of the previous study (Hetényi *et al.*, 2006) could be involved in the models [$N = 53$ in Equation (3)]. In case of W the level of correlation ($r^2 = 0.962$) did not decrease when the three former outliers were included.

3.3 Cross-validation of the correlations

There were different methods applied for cross-validation of the correlations presented in Table 3. The cross-validated correlation coefficients (r^2_{cv}) of the leave-one-out (LOO) and leave-50%-out (L50%O) methods (Table 3) shows that exclusion of one or more data points from the models does not decrease the level of correlation dramatically. A set of 20 drug–protein complexes (Table 1) was used as an external validation set (EXT). Most of the corresponding r^2 values are above 0.5 showing that the models can predict the EI values for smaller, drug ligands not included in the training set (50 systems). Notably, ΔG_C(AD4) produced higher $r^2_{\text{cv,EXT}}$ values for the external validation than ΔG_C(AD3), probably due to the more advanced solvation terms and the larger compound database included in its parametrization. The ΔG_{full}(AD4) function did not result better EI-correlations (data not shown) than

Table 2. Codes and definitions of size-dependent normalizing factors (SNF) of ligands used in the denominator of efficiency indices [Equation (1)]

Code	Name of SNF	Definition	References
1D SNF's			
N_{AT}	Number of atoms		Karelson (2000)
N_{HAT}	Number of heavy atoms		Hopkins *et al.* (2004)
N_{B}	Number of σ-bonds		Karelson (2000)
M_W	Molecular weight		Abad-Zapatero and Metz (2005)
2D SNF's			
W	Wiener index	$W = \frac{1}{2}\sum_{i,j}^{N_{\mathrm{SA}}} d_{ij}$	Wiener (1947)
$^0\chi$	Randic index ($n = 0$)	$^n\chi = \sum_{n\mathrm{-lengthpaths}}^{N_{\mathrm{SB}}} (\delta_{i1}, \ldots, \delta_{in+1})^{-1/2}$	Randić (1975)
$^1\chi$	Randic index ($n = 1$)		
$^2\chi$	Randic index ($n = 2$)		
$^3\chi$	Randic index ($n = 3$)		
$^0\chi^v$	Kier&Hall index ($n = 0$)	$^n\chi^v = \sum_{n\mathrm{-lengthpaths}}^{N_{\mathrm{SB}}} (\xi_{i1}, \ldots, \xi_{in+1})^{-1/2}; \quad \xi_i = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1}$	Kier and Hall (1976)
$^1\chi^v$	Kier&Hall index ($n = 1$)		
$^2\chi^v$	Kier&Hall index ($n = 2$)		
$^3\chi^v$	Kier&Hall index ($n = 3$)		
$^1\kappa$	Kier shape index ($n = 1$)	$^n\kappa = (N_{\mathrm{SA}} + \alpha)(N_{\mathrm{SA}} + \alpha - 1)^2(^nP + \alpha)^2$	Kier (1990)
$^2\kappa$	Kier shape index ($n = 2$)	$^n\kappa = (N_{\mathrm{SA}} + \alpha - 1)(N_{\mathrm{SA}} + \alpha - 2)^2(^nP + \alpha)^2$	
$^3\kappa$	Kier shape index ($n = 3$)	$\begin{cases} ^n\kappa = (N_{\mathrm{SA}} + \alpha - 1)(N_{\mathrm{SA}} + \alpha - 3)^2(^nP + \alpha)^2, \text{ if } N_{SA} \text{ is odd} \\ ^n\kappa = (N_{\mathrm{SA}} + \alpha - 3)(N_{\mathrm{SA}} + \alpha - 2)^2(^nP + \alpha)^2, \text{ if } N_{SA} \text{ is even} \end{cases}$	
Φ	Kier flexibility index	$\phi = (^1\kappa^2\kappa)/N_{\mathrm{SA}}$	Kier (1990)
J	Balaban index	$J = \left(\frac{q}{\mu+1}\right)\sum_{i,j}^{q}\left(s_i s_j\right)^{-1/2}; \quad \mu = q - n + 1$	Balaban (1982)
3D SNF's			
$I_A I_B I_C$	Product of principal moments of inertia	$I_A I_B I_C = \prod_x I_x; \quad I_x = \sum_i A_{M,i} r_{x,i}^2; \quad (x = A,\ B \text{ or } C)$	Karelson (2000)
$\mathrm{GI_B}$	Gravitation index (all bonds)	$\mathrm{GI_B} = \sum_{i<j}^{N_{\mathrm{B}}} \frac{A_{M,i}A_{M,j}}{r_{ij}^2}$	Karelson (2000)
$\mathrm{GI_P}$	Gravitation index (all pairs)	$\mathrm{GI_P} = \sum_{i<j}^{N_{\mathrm{A}}} \frac{A_{M,i}A_{M,j}}{r_{ij}^2}$	Karelson (2000)

N_{SA}: number of atoms in the molecular graph (hydrogens excluded); N_{SB}: number of bonds in the graph; *n*: length of bonding path (topological distance, order of descriptor); d_{ij}: entry of the distance matrix corresponding to the number of bonds in the shortest path connecting the pair of atoms *i* and *j*; δ: coordination number of atoms; *v*: valence of atom in a molecule; ξ: value of atomic connectivity; Z_{i}: total number of electrons in atom *i*; Z_i^v: number of valence electrons in atom *i*; H_i: number of hydrogens directly attached to atom *i*; ^{n}P: number of paths of length n in the molecular graph; α: sum of all ratios of the *i*th atomic radius and radius of sp^3 carbon atom for all atoms in the graph minus 1; *q*: number of edges in the molecular graph; *m*: number of vertices in the graph; μ: cyclometric number; s_i and s_j: distance sums obtained by summation of row *i* and column *i* (or row *j* and column *j*) of the distance matrix; A_M: atomic mass; $r_{x,i}$: distance of the *i*th atom from principal axis *x*; *A*, *B*, *C*: principal axes; N_{A}: number of atoms; N_{B}: number of bonds; r_{ij}: interatomic distance.

ΔG_{C}(AD4), and, therefore ΔG_{C}(AD4) was selected for the final evaluations (Table 3).

The results of the cross-validated correlations in Table 3 allow us to conclude that structure-based calculation of EI's works for both the 'traditional' (Abad-Zapatero and Metz, 2005; Hopkins *et al.*, 2004) and the newly introduced 2D SNF's (W, χ's, J, etc.). The formulas in Equation (3) and Table 2 and the validated models can be coded and applied as EI-calculators during the *in silico* drug design process (Fig. 1). Direct implementation of EI-calculator algorithms in docking/scoring program packages such as AutoDock is also possible.

3.4 Applications

(1) To check the applicability of two new EI's with the best correlations (Table 3), the distributions of ΔG_{E} and EI values were compared for the sets including the 50 peptidic compounds (non-drugs) and the 20 drugs, respectively. It was found (Fig. 3 and Supplementary Material) that overlapping distributions of ΔG_{E}'s (Fig. 3A) of drugs and non-drugs are separated for the EI's (Fig. 3B). There are one or two orders of magnitude difference (Table 4) in the median/average values of EI's for both W and $I_A I_B I_C$ and there are considerably large gaps between the minimum values of drugs and non-drugs, as well. These results emphasize the applicability of the new EI's in separation of drugs from non-drugs.

(2) The introduction of EI's in a virtual screening process improves the selectivity of screening. As a test case, the binding pocket of progesterone receptor was used as a target in the docking of 1760 compounds including an abridged version of the NCI Diversity Set (NCI/NIH; Lindstrom *et al.*, 2003) and the native drug ligand norethindrone (1sqn, Table 1). ΔG_{C}'s were collected and W- and $I_A I_B I_C$-based EI's were calculated. Details on

Table 3. Statistical parameters of linear regressions [Equation (3)] obtained for efficiency indices based on SNF's of different dimensionality

SNF	r^2	$r^2_{\mathrm{cv,LOO}}$	$r^2_{\mathrm{cv,L50\%O}}$	$r^2_{\mathrm{cv,EXT}}$	F-value	r^2	$r^2_{\mathrm{cv,LOO}}$	$r^2_{\mathrm{cv,L50\%O}}$	$r^2_{\mathrm{cv,EXT}}$	F-value
	AutoDock3.0					AutoDock4				
1D SNF's										
N_{AT}	0.857	0.845	0.852	0.493	286.90	0.839	0.826	0.835	0.718	250.28
N_{HAT}	0.896	0.886	0.891	0.593	413.66	0.887	0.877	0.884	0.758	378.11
N_{B}	0.865	0.854	0.863	0.522	308.23	0.848	0.835	0.844	0.731	266.98
M_W	0.889	0.879	0.889	0.607	386.55	0.881	0.870	0.879	0.767	354.22
2D SNF's										
W	**0.962**	**0.954**	**0.954**	**0.910**	**1216.46**	**0.960**	**0.953**	**0.952**	**0.931**	**1139.22**
$^0\chi$	0.891	0.882	0.890	0.589	394.40	0.884	0.873	0.879	0.757	364.67
$^1\chi$	0.893	0.884	0.889	0.582	402.59	0.884	0.873	0.883	0.752	364.56
$^2\chi$	0.918	0.910	0.911	0.686	540.18	0.913	0.905	0.908	0.803	503.64
$^3\chi$	0.916	0.908	0.912	0.856	524.80	0.906	0.897	0.904	0.909	461.11
$^0\chi^{\mathrm{v}}$	0.892	0.882	0.830	0.548	397.04	0.881	0.870	0.878	0.749	355.47
$^1\chi^{\mathrm{v}}$	0.886	0.876	0.874	0.571	372.51	0.871	0.860	0.856	0.764	325.15
$^2\chi^{\mathrm{v}}$	0.914	0.906	0.910	0.692	509.66	0.904	0.895	0.903	0.832	449.73
$^3\chi^{\mathrm{v}}$	0.905	0.897	0.896	0.803	458.55	0.889	0.880	0.884	0.896	384.51
$^1\kappa$	0.870	0.859	0.866	0.602	322.41	0.862	0.850	0.860	0.782	300.81
$^2\kappa$	0.791	0.774	0.788	0.603	181.45	0.777	0.759	0.776	0.801	167.30
$^3\kappa$	0.781	0.764	0.726	0.719	170.81	0.784	0.768	0.729	0.859	174.61
Φ	0.742	0.723	0.739	0.667	137.79	0.729	0.709	0.727	0.845	129.26
J	0.871	0.855	0.860	0.847	325.16	0.854	0.834	0.845	0.889	280.01
3D SNF's										
$\boldsymbol{I_AI_BI_C}$	**0.966**	**0.929**	**0.938**	**0.961**	**1345.18**	**0.963**	**0.933**	**0.938**	**0.963**	**1246.91**
$\mathrm{GI_B}$	0.900	0.891	0.890	0.660	432.20	0.892	0.882	0.887	0.787	396.71
$\mathrm{GI_P}$	0.927	0.919	0.926	0.796	606.68	0.921	0.914	0.917	0.863	563.02

r^2: the squared correlation coefficient; r^2_{cv}: the square of the cross-validated correlation coefficient (LOO: leave-one-out method, L50%O: leave-50%-out method, EXT: external validation set of 20 drug-protein systems). The boldfaced letters and values signify the best corelations.

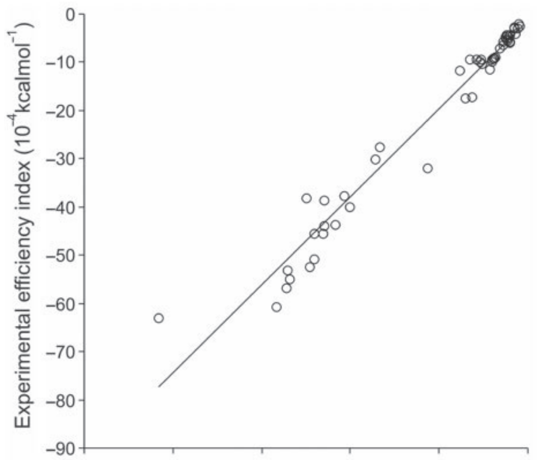


Fig. 2. The correlation of experimental and calculated efficiency indices (EI) using the Wiener index as a size-dependent normalizing factor (AD3 scoring).

the methods of these procedures are described in the Supplementary Material. It was found, that the use of ΔG_C's alone ranked norethindrone to the best 10% of the 1760 compounds. Re-ranking of the best 10% according to W- and $I_AI_BI_C$-based EI's resulted norethindrone in the second and sixth best position (< top 0.5%) on the list of the 1760 compounds, respectively. This test showed that in a second ranking step these new EI's can improve the quality of selection of a real drug.

4 DISCUSSION

4.1 The background of the thermodynamics of EI's

The binding free energy (ΔG) can be written as the sum of experimental enthalpic (ΔH) and entropic (ΔS) binding contributions [Equation (4)], where T is the thermodynamic temperature.

$$\Delta G = \Delta H - T\Delta S \tag{4}$$

As an additive quantity, ΔS can be further split into translational (ΔS_t), rotational (ΔS_r) and vibrational (ΔS_v) entropy changes [Equation (5)] at the ligand molecule. In some articles

(Noskov and Lim, 2001), further contributions are also considered such as solvation/desolvation free energy (ΔG_s) of the ligand and/or the protein molecules, etc. As the SNF's depend solely on the ligands, involvement of protein effects is not necessary in the forthcoming discussion.

$$\Delta G = \Delta H - T(\Delta S_t + \Delta S_r + \Delta S_v) + \Delta G_s \quad (5)$$

The use of statistical thermodynamics expressions (Carlsson and Åqvist, 2005, Murray and Verdonk, 2002) for estimation of components S_t, S_r (Table 5) and S_v of molecular entropy is quite common. S_v depends on the frequencies of normal modes of the ligand molecule, which cannot be connected with the simple SNF's of this study. ΔG_s includes both enthalpic and entropic contributions (Zou *et al.*, 1999) and partly depends on the molecular size and shape of the ligand via the solvent accessible surface area. Accordingly [Equation (5)], the division of ΔG_E [left side of Equation (3)] with SNF's results in normalized ΔH_E's and ΔS_E's.

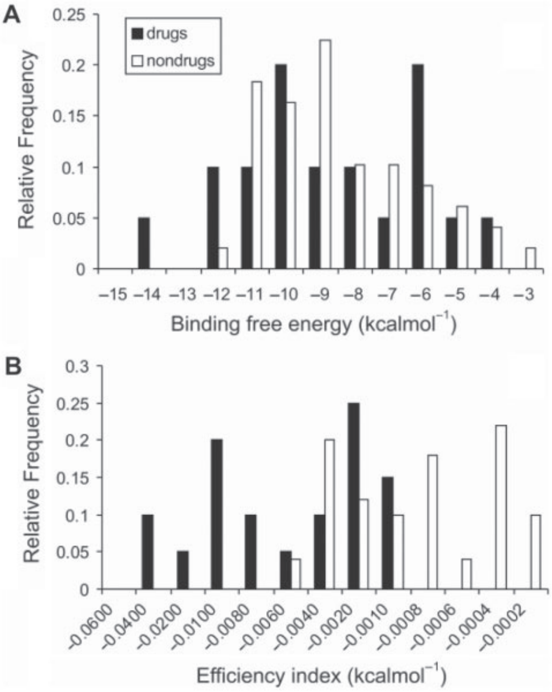


Fig. 3. Histograms showing the distribution of experimental binding free energy (**A**) and Wiener index-based efficiency index (**B**) values for drugs and non-drugs. (The scales cover the full range of values and the same number of bins were applied for both histograms.)

On the right side of Equation (3) there is ΔG_C Equation (6), including three terms [Equation (2), Methods section], which can be assigned (Brooijmans and Kuntz, 2003, Calderone and Williams, 2001) to the enthalpic (ΔH_C) contributions of binding. The fourth term of ΔG_C ($\Delta G_{s,C}$) is an estimate of ΔG_s which represents only a minor portion of ΔG_C [Equation (6)].

$$\Delta G_C = \Delta H_C + \Delta G_{s,C}. \quad (6)$$

Thus, the SNF-normalized ΔG_C [Equations (3) and (6)] contains mostly normalized ΔH_C (and negligible $\Delta G_{s,C}$). Most importantly, there are no terms estimating ΔS_t, ΔS_r and ΔS_v on the right side.

If assuming that experimental entropy (S_E), i.e. S_t and S_r becomes zero after ligand binding, then ΔS_t and ΔS_r will include size-dependent factors, such as M_W or the product of principal moments of inertia ($I_A I_B I_C$), respectively (Table 5). However, it was correctly discussed (Carlsson and Åqvist, 2005), that the assumption of zero final entropy is rather hypothetical as the ligand does fluctuate around its

Table 5. Statistical thermodynamics formulas of molecular entropy

Molecular entropy	Formula
Translational (S_t) (Sackur–Tetrode)	$S_t = Nk \ln\left[\frac{Ve^{5/2}}{N}\left(\frac{2\pi kTM_W}{h^2}\right)^{3/2}\right]$
Rotational (S_r)	$S_r = Nk \ln\left[\frac{8\pi^2}{\sigma}\left(\frac{2\pi ekT}{h^2}\right)^{3/2}(I_A I_B I_C)^{1/2}\right]$

Note, that the Sackur–Tetrode equation used for discussion was originally derived for gas phase. *V*: volume available for the molecule; *N*: number of molecules; M_W: molecular weight; *k*: Boltzmann's constant; *T*: thermodynamic temperature; *h*: Planck's constant; σ: symmetry number; $\boldsymbol{I_A I_B I_C}$: product of principle moments of inertia (see also Table 1).

Table 4. Statistics of the distribution of experimental binding free energy values and efficiency indices based on Wiener index (EI_W) and $I_A I_B I_C$ ($\mathrm{EI}_{I_A I_B I_C}$) for drugs and non-drugs

	Median	Average	Minimum	Maximum
ΔG_E				
Drugs	−9.87	−9.30	−14.75	−4.63
Non-drugs	−9.50	−8.86	−12.94	−3.89
EI_W				
Drugs	-6.281×10^{-3}	-1.226×10^{-2}	-5.930×10^{-2}	-1.014×10^{-3}
Non-drugs	-9.728×10^{-4}	-2.101×10^{-3}	-2.137×10^{-4}	-2.137×10^{-4}
$\mathrm{EI}_{I_A I_B I_C}$				
Drugs	-3.151×10^{-10}	-6.564×10^{-9}	-6.783×10^{-8}	-4.644×10^{-12}
Non-drugs	-4.190×10^{-12}	-3.579×10^{-11}	-3.283×10^{-10}	-1.677×10^{-13}

ΔG_E and EI_W values are in kcalmol^{-1} units. $\mathrm{EI}_{I_A I_B I_C}$ has a dimension of kcalmol^{-1}amu^{-3}Å^{-6}.

binding position. Whereas the formulas of Table 5 can hardly be applied for calculation of binding entropy of ligands in their present forms, they obviously show the dependence of molecular entropy on M_W and $I_AI_BI_C$ of ligands. Thus, normalization of $\Delta S_{\rm E}$ [left side of Equation (3)] with SNF's such as M_W or $I_AI_BI_C$ can be expected to decrease the ligand-dependency of the $\Delta S_{\rm E}$ terms resulting in a constant part of the normalized $\Delta S_{\rm E}$.

The constant part of SNF-normalized $\Delta S_{\rm E}$ does not affect the level of correlation and the remaining SNF-normalized enthalpic terms in Equation (3) correlate well with each other (Table 3).

4.2 New 3D SNF's

To test the prediction of the previous section, i.e. the usefulness of $I_AI_BI_C$ as a 3D SNF, it was employed in Equation (3). The statistical parameters of the corresponding LR (Table 3, Supplementary Material) show an excellent correlation ($r^2 = 0.966$) verifying the expectation. Remarkably, both the 3D $I_AI_BI_C$ and the 2D W involve the calculation of real or topological internal lengths of the ligand molecules, and, therefore their connection is trivial. Their correlation for the 50 ligands is $r^2 = 0.864$. Interestingly, the 2D W performed as well (Table 3) as the obviously more elaborate 3D $I_AI_BI_C$ in case of the 50 systems. It was also found, that $I_AI_BI_C$ works even for smaller subsets of the 50 investigated systems resulting in, e.g. an r^2 of 0.973 for the 10 modeled β-secretase complexes alone (AD3 scoring).

Other internal distance-based 3D SNF's such as the gravitation index (GI), a descriptor successful in prediction of boiling points (Katritzky *et al.*, 1996) also provided good LR results in calculation of EI's (Table 3).

4.3 Methodological aspects of the results

Scoring functions of docking programs are generally based on correlations of $\Delta G_{\rm E}$ with $\Delta G_{\rm C}$. However, during the development of scoring functions, separate fit of experimental $\Delta H_{\rm E}$ and $\Delta S_{\rm E}$ to the corresponding enthalpic and entropic terms (Brooijmans and Kuntz, 2003) of the scoring functions would be an ideal way (Murphy, 1999) to decrease errors coming from overlapping and/or coupled terms. However, most of the experimental thermodynamic BA data available are $\Delta G_{\rm E}$ values or pK's from which $\Delta G_{\rm E}$'s can be calculated (Wang *et al.*, 2004). The amount of enthalpic data is limited as experimental binding enthalpy ($\Delta H_{\rm E}$) can be obtained only by additional measurements with special techniques, e.g. isothermal titration calorimetry (Campoy and Freire, 2005). The LR's of the previous sections showed, that the SNF-normalization of $\Delta G_{\rm E}$ provides excellent correlation with the normalized $\Delta H_{\rm C}$ without additional measurements of $\Delta H_{\rm E}$, due to the high enthalpic content of both sides of Equation (3) (see previous sections for details).

It can also be recognized [Equation (3)], that the reciprocals of the SNF's are actual weights in the weighted least squares fit of the calculated enthalpic terms to the experimental $\Delta G_{\rm E}$'s. By using these weights during development of scoring functions, the degree of correlation and the accuracy of computational docking-scoring methods can be increased.

4.4 Practical applications

The EI's are simple indicators developed to aid rational drug design and hit-to-lead approaches (Keserű and Makara, 2006). In the present study, new EI's involving 2D and 3D SNF's were introduced. It was shown, that precise, structure-based calculation of EI's is a real alternative of time-consuming measurements and that the new EI's can be used in separation of drugs from non-drugs. The calculation of EI's of a large set of available drugs will allow the determination of reference EI-limits for selection of drug-like candidates in the future. The building of an EI database for the precise determination of EI-limits has already been started in our laboratory. As the proposed EI-calculators are fast and cost-effective, they will help to reduce the number of experimental measurements and can easily be combined with available methods in high-throughput computational docking and scoring (Fig. 1).

ACKNOWLEDGEMENTS

The authors are thankful to Foundation Innove (www.innove.ee) project No. 1.0101-0310 for the financial support. The referees of the manuscript are acknowledged for their constructive suggestions.

Conflict of Interest: none declared.

REFERENCES

Abad-Zapatero,C. and Metz,J.T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today*, **10**, 464–469.

Balaban,A.T. (1982) Highly discriminating distance based topological index. *Chem. Phys. Lett.*, **89**, 399–404.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Brooijmans,N. and Kuntz,I.D. (2003) Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 335–373.

Calderone,C.T. and Williams,D.H. (2001) An enthalpic component in cooperativity: the relationship between enthalpy, entropy, and noncovalent structure in weak associations. *J. Am. Chem. Soc.*, **123**, 6262–6267.

Campoy,A.V. and Freire,E. (2005) ITC in the post-genomic era…? Priceless. *Biophys. Chem.*, **115**, 115–124.

Carlsson,J. and Åqvist,J. (2005) Absolute and relative entropies from computer simulation with applications to ligand binding. *J. Phys. Chem. B*, **109**, 6448–6456.

Chen,W. *et al.* (2006) Concepts in receptor optimization: targeting the RGD peptide. *J. Am. Chem. Soc.*, **128**, 4675–4684.

DeLano,W.L. (2006) *PyMol Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Devillers,J. and Balaban,A.T. (eds.) (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Beach Science Publishers, Amsterdam.

Hetényi,C. *et al.* (2006) Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc.*, **128**, 1233–1239.

Hong,L. *et al.* (2000) Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science*, **290**, 150–153.

Hopkins,A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today*, **9**, 430–431.

Huey,R. *et al.* (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, **28**, 1145–1152.

Karelson,M. (2000) *Molecular Descriptors in QSAR/QSPR*. J. Wiley & Sons, New York.

Karelson,M. *et al.* (1996) Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.*, **96**, 1027–1043.

Katritzky,A.R. *et al.* (1995) QSPR:the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.*, **24**, 279–287.

Katritzky,A.R. *et al.* (1996) Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.*, **100**, 10400–10407.

Keserű,G.M. and Makara,G.M. (2006) Hit discovery and hit-to-lead approaches. *Drug Discov. Today*, **11**, 741–748.

Kier,L.B. (1990) Rouvray,D.H. (ed.) *Computational Chemical Graph Theory*. Nova Science Publishers, New York, pp. 151–174.

Kier,L.B. and Hall,L.H. (1976) *Molecular Connectivity in Chemistry and Drug Reasearch*. Academic Press, New York.

Kuntz,I.D. *et al.* (1999) The maximal affinity of ligands. *Proc. Natl Acad. Sci. USA*, **96**, 9997–10002.

Lindstrom,W.H. *et al.* (2003) The NCI Diversity Set for AutoDock. http://autodock.scripps.edu/resources/databases

Lipinski,C. and Hopkins,A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.

Lipinski,C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.

Morris,G.M. *et al.* (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.*, **10**, 293–304.

Morris,G.M. *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.

Murphy,K.P. (1999) Predicting binding energetics from structure: looking beyond DeltaG degrees. *Med. Res. Rev.*, **19**, 333–339.

Murray,C.W. and Verdonk,M.L. (2002) The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.*, **16**, 741–753.

NCI/NIH, National Cancer Institute, Developmental Therapeutics Program. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html.

Noskov,S.Y. and Lim,C. (2001) Free energy decomposition of protein-protein interactions. *Biophys. J.*, **81**, 737–750.

Randić,M. (1975) On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.

Ren,P. and Ponder,J.W. (2003) Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B*, **107**, 5933–5947.

Rishton,G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today*, **8**, 86–96.

Schuffenhauer,A. *et al.* (2006) Relationships between molecular complexity, biological activity, and structural diversity. *J. Chem. Inf. Model.*, **46**, 525–535.

Stouten,P.F.W. *et al.* (1993) An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.*, **10**, 97–120.

Swinney,D.C. (2004) Biochemical mechanisms of drug action: what does it take for success? *Nat. Rev. Drug Discov.*, **3**, 801–808.

Swinney,D.C. (2006) Biochemical mechanisms of new molecular entities (NMEs) approved by United States FDA during 2001–2004: mechanisms leading to optimal efficacy and safety. *Curr. Top. Med. Chem.*, **6**, 461–478.

Wang,R. *et al.* (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.

Wiener,H. (1947) Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.

Yang,J.-M. and Tung,C.-H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.*, **34**, 3646–3659.

Zou,X. *et al.* (1999) Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. *J. Am. Chem. Soc.*, **121**, 8033–8043.

D29

Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions

ALFONSO T. GARCÍA-SOSA, CSABA HETÉNYI, UKO MARAN
Institute of Chemistry, University of Tartu, Jakobi 2, Tartu 51014, Estonia

Received 3 January 2009; Revised 13 February 2009; Accepted 20 March 2009
DOI 10.1002/jcc.21306
Published online 6 May 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: A dataset of protein-drug complexes with experimental binding energy and crystal structure were analyzed and the performance of different docking engines and scoring functions (as well as components of these) for predicting the free energy of binding and several ligand efficiency indices were compared. The aim was not to evaluate the best docking method, but to determine the effect of different efficiency indices on the experimental and predicted free energy. Some ligand efficiency indices, such as $\Delta G/W$ (Wiener index), $\Delta G/\mathrm{NoC}$ (number of carbons), and $\Delta G/P$ (partition coefficient), improve the correlation between experimental and calculated values. This effect was shown to be valid across the different scoring functions and docking programs. It also removes the common bias of scoring functions in favor of larger ligands. For all scoring functions, the efficiency indices effectively normalize the free energy derived indices, to give values closer to experiment. Compound collection filtering can be done prior or after docking, using pharmacokinetic as well as pharmacodynamic profiles. Achieving these better correlations with experiment can improve the ability of docking scoring functions to predict active molecules in virtual screening.

© 2009 Wiley Periodicals, Inc. J Comput Chem 31: 174–184, 2010

Key words: virtual screening; scoring function; drug design; docking; free energy of binding

Introduction

Many drugs have been developed with the use of structure-based drug design and molecular docking.[1–5] When used correctly, docking can be an invaluable tool for drug discovery and design. Commonly, docking is used as a complement to other techniques such as high-throughput screening (HTS). However, as an example, Pierce et al. show that it can also be the primary technique, predicting 4 kinase inhibitors with a 14-fold increase in enrichment over HTS.[6] The active molecules' binding modes predicted by docking were experimentally confirmed by X-ray crystallography.

Docking scoring functions perform generally well for predicting protein-ligand binding modes,[1–5] although they are less accurate for predicting binding free energy.[1–5,7] Docking programs employ at least one scoring function for calculating the fit or energy of a protein-ligand association. Scoring functions are usually derived from atomic parameters generated from empirical or knowledge-based approximations to the experimental binding energy of protein-ligand complexes. Most scoring functions are additive in nature, in the sense that the more functional groups a ligand has, the more interactions it can have with the protein and the greater the intermolecular energy is thus calculated. In the case of polar functional groups, this would normally be offset by higher desolvation energies, which are unfavorable to the overall binding free energy. However, these desolvation energies, if included in the scoring function or docking program at all, do not tend to reflect the real trends, and so the scoring functions end up overestimating the binding energy for larger ligands at the expense of smaller ligands.[1] A similar situation arises for large hydrophobic ligands because the larger the molecule, the more van der Waals contacts are calculated. Again, large molecules would also incur in entropy penalties when binding and even if some scoring functions attempt to estimate this entropy loss by a measure of the number of rotatable bonds of a ligand, they are not accurate and end-up still favoring larger molecules. The proper calculation of entropies of binding is also a complex issue for scoring functions, unlikely to be solved by simple rotatable bond counts.[8]

Additional Supporting Information may be found in the online version of this article.

Correspondence to: A. T. García-Sosa; e-mail: alfonsog@ut.ee

Contract/grant sponsor: Estonian Science Foundation; contract/grant number: JD80

Contract/grant sponsor: Estonian Ministry for Education and Research; contract/grant number: SF0140031Bs09

An inaccuracy of only 1–2 kcal/mol represents already a difference of one or more orders of magnitude in the calculated affinities of proteins for ligands. However, even within this level of inaccuracy, docking should be able to classify ligands as having milli-, micro-, or nanomolar affinity in order to have predictive ability (a difference of 3 orders of magnitude is around 4 kcal/mol). In drug discovery and design, accuracy is arguably a more critical value to achieve than extreme precision. In other words, accurate relative ranking of diverse and unrelated active and inactive compounds is more sought after than less accurate but precise absolute binding energies. In any case, there is a need for computed values with higher accuracy that can be compared with experimental data. This would reduce the number of false positive and false negatives that a virtual screen can produce. The development of better scoring functions and docking methods is an active field of research,[1,2] with improvements likely to come from better descriptions and parameterizations of binding,[9–12] solvation interactions,[13–17] as well as flexibility[18–20] and entropy effects.[8,21]

Another method of improving the result from docking experiments is postprocessing the results, such as combining the result of several scoring functions, called "consensus scoring,"[7,22] or rescoring the energy for docked poses with a different method, such as molecular mechanics Poisson-Boltzmann (MM/PBSA)[23] or generalized Born/Surface Area (MM/GBSA).[24,25] Comparisons between scoring functions and related challenges have been performed elsewhere,[26,27] and it is not the objective of this article.

Recently, ligand efficiency indices (E.I. = ΔG/Measure, where ΔG is the binding free energy) have been proposed as a method to normalize the experimental,[28–31] as well as computational binding free energies of ligands.[32–35] An efficiency index measure can be any molecular measure of comparison between ligands and can be related to the molecular size such as molecular weight (MW), number of heavy atoms (NHA), number of carbons (NoC), or molecular or polar surface area. They can also be related to the solubility and permeability of a ligand by incorporating the logarithm of the octanol-water partition coefficient log P.[34] They can provide a measure of how efficiently a ligand binds to a biomolecule, even being able to determine the compounds that may disrupt a protein–protein interaction if their number of heavy atoms efficiency index, ΔG/NHA, is deeper than -0.24 kcal/molNHA.[35] The reason for this is that those small molecules have a higher efficiency of binding per heavy atom than the protein or peptide they displace, even with a surface area as low as half that of the peptide or protein.[35] There is a well-known tendency of lead molecules to increase in size and lipophilicity during optimization in search of higher affinity.[36] But this increase in lipophilicity can also carry more risks in associated side-effects and toxicity.[36] The related measure of pIC50 − cLogP has also been introduced to try to define the lipophilic space available to drug candidates.[36]

In this work, we explore how different docking programs and scoring functions can correlate with experimental values, both for free energy of binding, as well as to five different efficiency indices. These efficiency indices are free energy of binding/molecular weight (ΔG/MW), free energy of binding/number of heavy atoms (ΔG/NHA), free energy of binding/number of carbons (ΔG/NoC), logarithm of −free energy of binding/partition coefficient ($\log(-\Delta G/P)$), and free energy of binding/Wiener index ($\Delta G/W$).[33,34] Achieving better correlations of scoring functions with experimental values can increase the accuracy of scoring functions, and therefore the reliability of docking programs to predict active molecules in screening procedures.

Results

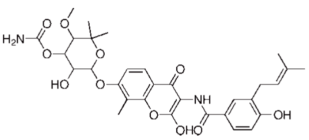
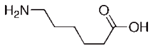
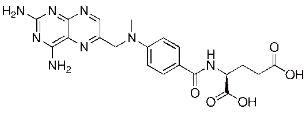
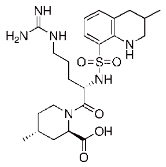
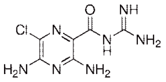
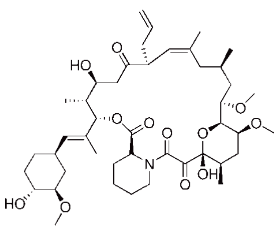
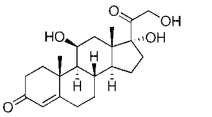
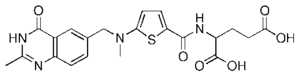
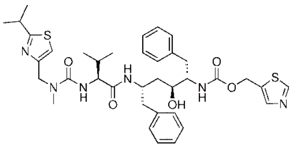
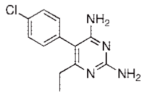
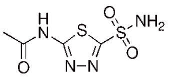
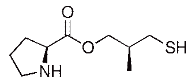
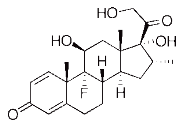
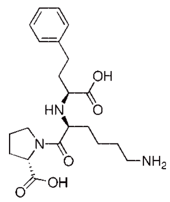
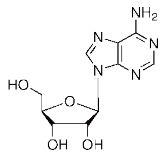
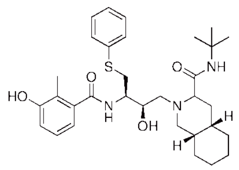
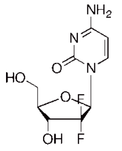
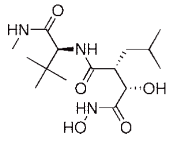
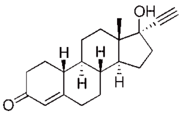
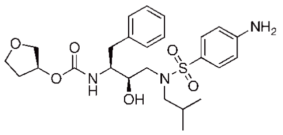
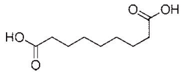
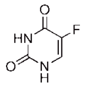
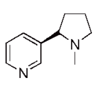
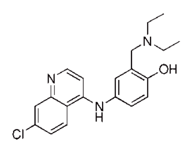
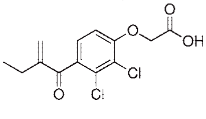
It is important to use drug compounds as test systems, because their complexity as compared with standard compounds can be challenging for scoring functions. Twenty-six protein-drug complexes with known experimental free energy of binding were obtained from comparing the PDBbind and DrugBank databases, yielding a wide variety of drugs with different shapes, sizes and chemical features. These complexes are given in the Supporting Information Table S1, while the structures of the drugs are shown in Table 1. Efficiency indices were also determined for all cases. Molecular surface area and polar surface area were not used because they can be sensitive to conformation.

The experimental free energies of binding were collected and the calculated free energies of binding were computed for each complex, using all the scoring functions as well as a selection of their components. Care was taken to use the experimental structure for calculating the docking score and only relaxations of this structure, or "docking in place" was performed, to maintain the same binding mode and pose for all programs (achieving complexes with electrostatic and van der Waals energies such as that in Fig. 1).

The results for several scoring functions and components for the 26 resulting protein-ligand complexes are shown in Table 2, and plotted results are shown in Figure 2a. Table 2 and Figure 2a show the difference in values for the experimental and calculated free energy of binding for each protein-drug complex. Chemscore and Goldscore are included in Figure 2a, even though they have positive scales (they return a positive value instead of a negative free energy). They were included as the negative of their value, i.e., −Goldscore (GS) and −Chemscore (CS), for the sake of comparison. XPc and SPc correspond to the Coulomb + van der Waals components of XP and SP, respectively. XP and SP correspond to XP and SP "refine" treatment, whereas SPi refers to SP "in place." ABE corresponds to autodock binding free energy, AIE to autodock intermolecular energy. DGe is the experimental calculated energy, DGb is a component of CS called ΔGbindGOLD. Some scoring functions have values that are closer to the experimental ones, and some follow the trend of the experimental values better.

The efficiency indices were then calculated for each scoring function value (as well as the selected components of the scoring functions) substituting the value for ΔG in ΔG/MW, ΔG/NHA, ΔG/NoC, $\log(-\Delta G/P)$, and $\Delta G/W$. The same efficiency indices were calculated for the experimentally determined ΔG. The means, medians, and ± standard deviations for all systems, as well as the complete tables are available in the Supporting Information Tables S2–S6. The plotted results for the molecular weight efficiency index for all scoring functions and experiment

Table 1. Drug Structure Dataset.

 Novobiocin, **1**	 Aminocaproic acid, **2**
 Methotrexate, **3**	 Argatroban, **4**
 Amiloride, **5**	 Tacrolimus, **6**
 Hydrocortisone, **7**	 Raltitrexed, **8**
 Ritonavir, **9**	 Pyrimethamine, **10**
 Acetazolamide, **11**	 Captopril, **12**
 Dexamethasone, **13**	 Lisinopril, **14**
 Adenosine, **15**	 Nelfinavir, **16**
 Gemcitabine, **17**	 Marimastat, **18**
 Norethindrone, **19**	 Amprenavir, **20**
 Azelaic acid, **21**	 5-Flurouracil, **22**
 Nicotine, **23**	 Amodiaquine, **24**
 Ethacrynic acid, **25**	

are shown in Figure 2b, while the Wiener index efficiency index is shown in Figure 2c.

As can be seen from Supporting Information Tables S1–S5 and Figure 2b (ΔG/MW), there is still some variation between the experimental efficiency indices and the calculated efficiency indices. However, Figure 2c (ΔG/W) shows how the experimental and calculated efficiency indices are now quite close. The linear regression correlation coefficients between the experimental and calculated binding energies, as well as between experimental and calculated efficiency indices were computed for all cases. They are shown in Table 3.

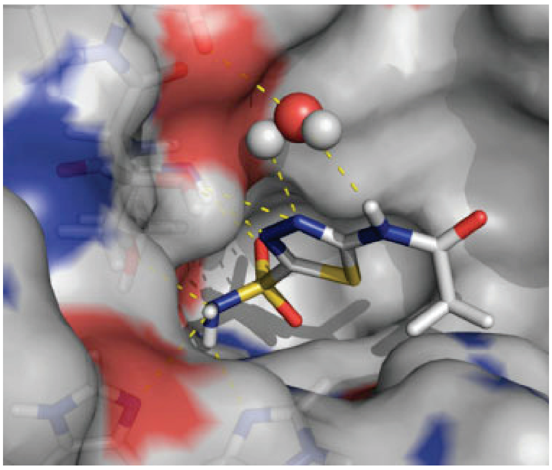


Figure 1. Complex of acetazolamide (11, sticks) with carbonic anhydrase X11 (surface and sticks) obtained by docking on crystal structure 1JD0. Nitrogen atoms in blue, oxygen in red, hydrogen in white, sulfur in yellow, hydrogen bonds as yellow dashes.

Some efficiency indices appear to be better than others for correlating experimental and calculated values. From Table 3, it can be seen that for some scoring functions, MW and NHA either do not improve the results or provide only a modest improvement over the correlations with experimental values. The simple measure NoC (number of carbons) provides a good correlation for some of the scoring functions. This efficiency index is related to the nonpolar surface area, because the larger NoC a compound has, the larger its nonpolar surface is likely to be. Therefore, it may be providing an indirect measure of the desolvation energy for a molecule. The efficiency index $\log(-\Delta G/P)$ provides good correlations for all scoring functions between experimental and calculated values. This index is directly related to the permeability of a molecule. The efficiency index $\Delta G/W$ also improves all of the correlations. The p values in Table 3 show the probability that the corresponding F-statistic could have occurred by chance. All of them are below α = 0.05, indicating that the regression models are useful in predicting the linear relationship with the experimental values (at a 95% confidence level). Efficiency indices, therefore, also appear to be able to introduce useful extra information in addition to the free energy of binding into a derived measurement.

As examples of the good linear correlations between experimental and calculated values, the plot of the experimental ΔG/NoC versus calculated ΔG/NoC for DGb (ΔGbindGOLD, a component of Chemscore) is shown in Figure 3a; experimental $\log(-\Delta G/P)$ versus calculated $\log(-\Delta G/P)$ for the same DGb is shown in Figure 3b; and the plot of the experimental $\Delta G/W$ versus calculated $\Delta G/W$ for DGb is shown in Figure 3c.

Table 2. Experimental and Calculated Free Energies of Binding (kcal mol^{-1}).[a]

PDB code	DGe	ABE	AIE	GS	CS	DGb	XP	SP	SPi
1aj6	−8.07	−9.97	−8.85	−44.84	−14.52	−17.88	−7.07	−7.69	−5.93
1cea	−6.76	−6.82	−6.82	−40.12	−20.72	−22.17	−8.18	−5.58	−6.86
1dhi	−9.90	−9.54	−9.54	−67.31	−22.29	−24.97	−9.46	−9.60	−9.00
1dhj	−8.93	−7.86	−10.56	−72.54	−25.83	−27.29	−8.72	−8.72	−7.60
1dwc	−10.10	−10.29	−10.29	−12.28	−27.79	−34.23	−11.08	−7.65	−6.07
1f5l	−7.19	−7.12	−7.52	−35.58	−17.86	−18.48	−6.92	−7.16	−6.57
1fkf	−12.81	−10.19	−12.04	−52.82	−35.20	−37.38	−11.33	−7.13	−6.26
1h61	−6.66	−6.81	−6.81	−24.87	−23.01	−23.99	−9.13	−5.90	−3.71
1hvy	−8.42	−8.33	−7.24	−46.14	−16.59	−17.50	−5.77	−6.87	−6.04
1hxw	−14.75	−15.12	−15.12	−90.22	−43.46	−47.69	−14.57	−12.24	−11.20
1j3j	−10.92	−7.50	−7.50	−52.20	−21.48	−25.36	−8.15	−6.55	−5.83
1jd0	−11.24	−5.87	−6.34	−41.68	−18.77	−22.77	−4.55	−4.22	−4.18
1m2x	−5.66	−14.66	−15.50	−53.91	−22.66	−24.30	−7.05	−9.82	−9.30
1m2z	−9.84	−10.24	−11.15	−46.61	−36.01	−38.04	−13.97	−9.44	−8.86
1o86	−13.04	−17.32	−21.14	−59.41	−34.82	−43.04	−12.90	−11.70	−10.69
1odi	−5.73	−5.34	−6.32	−48.17	−15.83	−16.93	−8.05	−7.90	−6.26
1ohr	−11.86	−12.57	−13.93	−52.87	−36.34	−39.09	−11.16	−9.43	−9.48
1p62	−6.35	−5.73	−5.73	−46.16	−18.62	−21.90	−12.56	−7.33	−5.76
1r55	−9.26	−9.37	−11.69	−52.77	−22.75	−35.60	−10.56	−9.47	−9.43
1sqn	−12.81	−10.07	−10.07	−60.74	−32.60	−35.49	−11.01	−8.51	−8.36
1t7j	−11.86	−10.34	−13.05	−72.31	−25.23	−29.21	−9.35	−7.68	−6.50
1tuf	−5.52	−7.11	−9.27	−23.67	−8.07	−11.63	−2.97	−4.36	−4.42
1upf	−6.27	−3.83	−3.83	−17.72	−10.25	−11.11	−5.65	−5.98	−5.30
1uw6	−10.01	−6.53	−6.78	−41.36	−25.63	−28.06	−4.97	−5.85	−5.94
2aou	−10.54	−10.35	−10.35	−24.79	−36.79	−39.81	−10.22	−9.00	−8.68
2gss	−6.73	−6.39	−7.63	−28.72	−18.02	−19.77	−6.45	−5.40	−5.57

[a] DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindGOLD; XP, XPrefine; SP, SPrefine; SPi, SP in place.

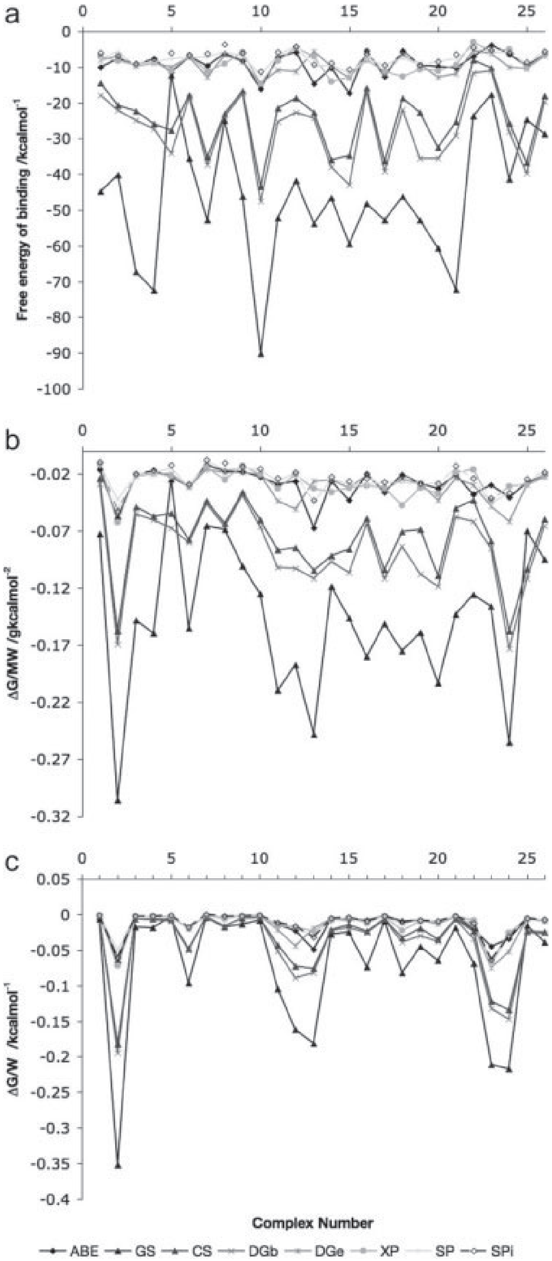


Figure 2. (a) Free energy of binding (ΔG) for each complex and several docking experiments, as well as determined by experiment. (b) Comparison of free energy of binding/molecular weight (ΔG/MW) efficiency indices for experiment and several scoring functions. (c) Comparison of free energy of binding/Wiener index (ΔG/W) efficiency indices for experiment and several scoring functions.

Since DGb is a component of CS, and therefore also of SP and XP, it is interesting to note that a component can have better correlation with experimental values than the full scoring function. This can be due to the need of improvement in the extra features of the scoring function, such as the desolvation penalties and entropy corrections. As controls, the experimental free energy of binding was correlated against the MW, NHA, NoC, log P, and Wiener values. The results did not show any strong linear correlation, with the R^2 values being 0.323, 0.364, 0.404, 0.205, and 0.315, respectively. This means that the good correlations found between experimental and calculated efficiency indices are not spurious or redundant.

Linear regressions were also carried out between the simple DGe and all of the calculated efficiency indices, and they showed no linear correlation stronger than 0.1. Linear regressions were also calculated for all the efficiency indices against the molecular properties (MW, NHA, NoC, etc.) to test the dominance of these in the derived efficiency index, showing no strong linear correlation either, with most beneath 0.5, except SPi and SP (most SPi and SP correlation R^2s around 0.6, except SPi/W and SP/W vs. W, $R^2 = 0.2$). An exception for all the scoring functions was log P, which showed strong correlation between log($-\Delta G/P$) vs. log P of circa 0.99 in R^2. However, this effect was created by the logarithm function. If the simple $\Delta G/P$ was calculated instead, then all of the scoring functions had correlations between $\Delta G/P$ vs. P lower than 0.1. Indeed, this efficiency index is better suited than log($-\Delta G/P$), and also shows strong correlations between calculated and experimental efficiency indices as seen in Table 4.

The improvements in going from ΔG to the different efficiency indexes are shown in Figure 4, where it can be seen that some of the efficiency indices (ΔG/NoC, log($-\Delta G/P$), $\Delta G/W$, and $\Delta G/P$) produce better improvements than others.

Figures 5a–5g show box plots for all of the distributions studied, and compares experimental and calculated efficiency indices where the spread between and within each series of data can be observed. The horizontal dark lines represent the median of the distributions, while the dark diamonds represent outliers. The plots of the free energy of binding have quite different spreads between the experimental and the calculated values, except for ABE, AIE, XP, SP, and SPi (Fig. 5a). ΔG/MW (Fig. 5b) and ΔG/NHA (Fig. 5c) indices do not change these spreads very much, while ΔG/NoC (Fig. 5d) already provides closer spreads between calculated and experimental values. Log($-\Delta G/P$) (Fig. 5e), ΔG/W (Fig. 5f), and $\Delta G/P$ (Fig. 5g) all show spreads that are now quite comparable between the experimental and calculated efficiency indices.

Shapiro normality tests were conducted for all the distributions studied, and they are shown in Supporting Information Table S8. Some of the distributions did not differ from a normal distribution with a 95% confidence limit and for these, Welch, independent, two-sided, t-tests were carried out between the experimental and calculated values (Table S9 in the Supporting Information). In the case of free energy of binding, for DGe, ABE, AIE, and XP, the test showed that the null hypothesis was true, i.e., that the true difference in means between the calculated and the experimental distributions is equal to zero, and they are comparable distributions. This was also the case for all the experimental and calculated log($-\Delta G/P$) efficiency indices.

For all of the distributions, Mann-Whitney U tests (a nonparametric test) were carried out to compare the experimental

Table 3. Linear Regression Correlation Coefficients and Statistics Between Experimental and Calculated Values for Binding Free Energy, ($y = ax + b$) as well as Five Efficiency Indices.[a]

ScorF	ΔG	ΔG/MW	ΔG/NHA	ΔG/NoC	log($-\Delta G/P$)	$\Delta G/W$
DGe	1	1	1	1	1	1
DGb	0.676, 50.2, $p < 0.001$	0.684, 51.9, $p < 0.001$	0.673, 49.4, $p < 0.001$	0.842, 127.7, $p < 0.001$	0.997, 9065, $p < 0.001$	0.885, 184.5, $p < 0.001$
CS	0.634, 41.6, $p < 0.001$	0.644, 43.4, $p < 0.001$	0.626, 40.3, $p < 0.001$	0.798, 94.7, $p < 0.001$	0.997, 7216, $p < 0.001$	0.870, 161.4, $p < 0.001$
GS	0.310, 10.8, 0.003	0.473, 21.6, $p < 0.001$	0.490, 23.0, $p < 0.001$	0.712, 59.4, $p < 0.001$	0.992, 2916, $p < 0.001$	0.822, 110.9, $p < 0.001$
XP	0.315, 11.0, 0.0029	0.319, 11.3, 0.0026	0.299, 10.2, 0.0038	0.450, 19.6, $p < 0.001$	0.994, 4341, $p < 0.001$	0.819, 108.6, $p < 0.001$
XPc	0.357, 13.3, 0.0012	0.362, 13.6, 0.0011	0.416, 17.1, $p < 0.001$	0.769, 79.8, $p < 0.001$	0.996, 5545, $p < 0.001$	0.877, 171.6, $p < 0.001$
SP	0.246, 7.82, 0.010	0.415, 17.0, $p < 0.001$	0.390, 15.3, $p < 0.001$	0.511, 25.1, $p < 0.001$	0.996, 6009.8, $p < 0.001$	0.856, 142.7, $p < 0.001$
SPc	0.387, 15.1, $p < 0.001$	0.363, 13.6, 0.0011	0.389, 15.3, $p < 0.001$	0.696, 55.1, $p < 0.001$	0.996, 5417, $p < 0.001$	0.912, 248.9, $p < 0.001$
SPi	0.261, 8.5, 0.0077	0.497, 23.7, $p < 0.001$	0.466, 20.9, $p < 0.001$	0.550, 29.3, $p < 0.001$	0.996, 5817, $p < 0.001$	0.864, 152.2, $p < 0.001$
ABE	0.347, 12.8, 0.0015	0.290, 9.8, 0.0046	0.273, 9.0, 0.0061	0.512, 25.2, $p < 0.001$	0.996, 5373, $p < 0.001$	0.743, 69.5, $p < 0.001$
AIE	0.318, 11.2, 0.0027	0.228, 7.1, 0.013	0.219, 6.7, 0.016	0.474, 21.6, $p < 0.001$	0.995, 4542, $p < 0.001$	0.718, 61.2, $p < 0.001$

R^2, F-statistic, and p values are given in the table.

[a] ScorF, scoring function; DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindGOLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.

and calculated distributions to assess whether two samples of observations come from the same distribution, including for all cases where the values were not normally distributed. The results are shown in Supporting Information Table S10. For free energy of binding, the test statistics W and p-values (p higher than 0.05, 95% confidence level) showed that there was no statistically significant difference between the experimental ΔG values and each of the calculated ABE, AIE, XP, and SP distributions. For ΔG/MW, there was no statistically significant difference between the experimental and each of the ABE, AIE, XP, SP, and SPi calculated distributions. This was also true for ΔG/NHA, ΔG/NoC, and $\Delta G/W$ index. For log($-\Delta G/P$) and $\Delta G/P$, all of the calculated distributions had no statistically significant difference to the experimental one, for all of the scoring functions studied.

The equations between experimental and calculated values were then Y-scrambled with random numbers in the same range of values. Nearly all the R^2 values were markedly lower than for the unscrambled models (below 0.6). The only exception were the values of log($-\Delta G/P$) which remained high even in the scrambled models. This indicates that this efficiency index is not particularly good for improving the correlations, since it cannot distinguish a true correlation from a random one, although the logarithm function was responsible for that behavior. Importantly, the efficiency index $\Delta G/P$ had low correlation values for the scrambled models, which indicates that it has reliability. From these scrambling results, we can see that there is a small component in the efficiency indices which improves the correlations with experimental values due to mathematical correction (that is, it is beneficial to have the values on the same scale), but it does not account for all of the improvement. This suggests that there may be physical underlying causes to the improvements, which depend on the normalizing measure incorporated into the efficiency index. The improvement effect may be due to description of the entropic part of the free energy of binding, through efficiency indices that describe the topology of a molecule (such as W).[32] Other efficiency indices such as NoC and P, may provide improvement through a description of the desolvation and of the permeability of a compound. For all scoring functions, the best efficiency indices effectively normalize the free energy derived indices, to give values closer to experiment.

Discussion

Efficiency indices can improve the outcome of docking scoring functions because they provide a closer agreement with experimental values. In addition, useful information related to the molecular properties of a molecule such as its lipophilicity P, or topology (described by W), can be incorporated into a single indicator. Some efficiency indices appeared to be better than others at improving the correlations. ΔG/NoC, $\Delta G/W$, and $\Delta G/P$ are better than ΔG/MW or ΔG/NHA, and this effect was observed for all scoring functions.

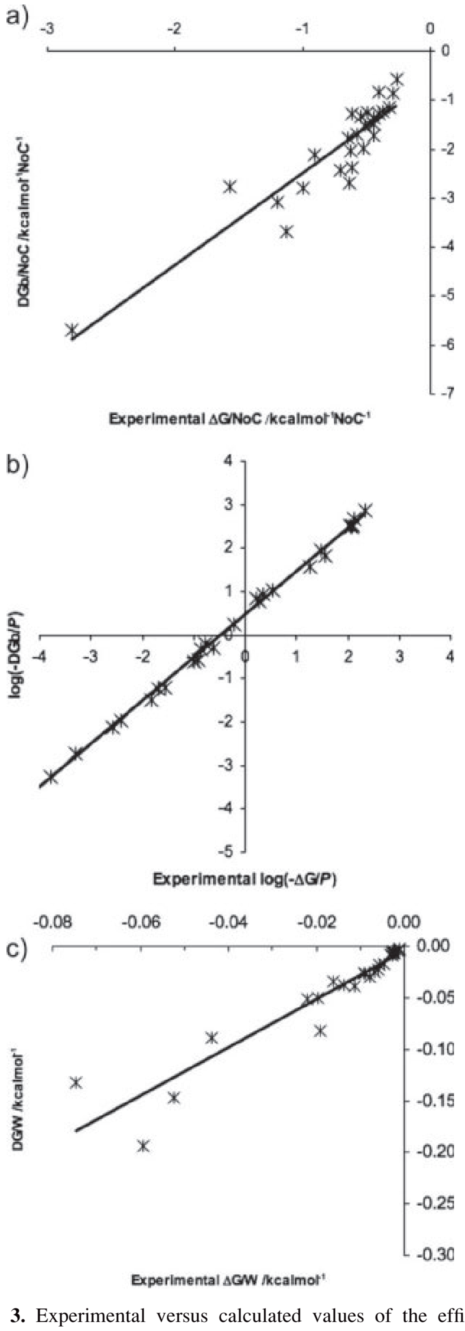


Figure 3. Experimental versus calculated values of the efficiency indices: (a) ΔG/NoC (free energy of binding/number of carbons) for DGb (ΔGbindGOLD) for 26 protein-drug complexes. $R^2 = 0.842$. (b) log($-\Delta G/P$) (logarithm of (−)free energy of binding/octanol-water partition coefficient) for DGb for 26 protein-drug complexes. $R^2 = 0.997$. (c) $\Delta G/W$ (free energy of binding/Wiener index) for DGb for 26 protein-drug complexes. $R^2 = 0.885$.

To test the performance of efficiency indices with different types of compounds, the 25 ligands (in 26 protein-ligand complexes) were separated into two groups, small and large ligands if they were below or above the average MW and also by taking the 1st quartile (lowest 25%) and 3rd quartile (highest 25%). The same separation was conducted for polar and nonpolar ligands only now considering polar surface areas (PSA, in Å^2).[37] The sum of squares of the residuals (a measure of fitting error) were then recorded for each ligand complex for the differences between the calculated and the experimental value as: RSS = Σ(Experimental value − Predicted value)^2/n, where n is the number of ligands, and the summation is over all the members in that group. The effect of molecular size on the efficiency indices were most marked for ABE, AIE, XP, and SP, where there was a large reduction of the difference between the errors for the small ligands compared to the large ones, using both separation methods. On average, small ligands had RSS errors of 10.19 kcal^2/mol^2 for binding free energy compared to 4.89 for large ligands. The efficiency indices markedly reduced this disparity till having equal differences between the errors for small and large ligands (average differences in RSS between small and large ligands: 0.0002 gkcal/mol^2 for ΔG/MW, 0.050 kcal/molNHA for ΔG/NHA, 0.234 kcal/molNoC for ΔG/NoC, 0.021 for log($-\Delta G/P$), and 0.0001 kcal/mol for $\Delta G/W$). This applied to all efficiency indices except $\Delta G/P$ where only SP/P produced the smallest differences (0.91). Nonpolar ligands (i.e., those with a small polar surface area) were also at a disadvantage compared to polar ones (large polar surface area). Using the 1st and 3rd quartiles, the efficiency indices (except $\Delta G/P$) for ABE, AIE, XP, and SP corrected this bias by reducing the differences in errors from averages of 9.59 kcal^2/mol^2 in binding free energy for nonpolar ligands and 4.27 for polar ligands, so that the differences between

Table 4. Linear Regression Correlation Coefficients and Statistics Between Experimental and Calculated Values for Binding Free Energy/Octanol-Water Partition Coefficient ($\Delta G/P$) Efficiency Index.[a]

Scoring function or component	$\Delta G/P$	Scoring function or component	$\Delta G/P$
DGe	1	DGe	1
DGb	0.981, 1247.9, 3.31e^{-22}	SPc	0.936, 350.3, 8.9e^{-16}
CS	0.989, 2206.8, 3.9e^{-25}	SP	0.986, 1686.1, 9.47e^{-24}
GS	0.926, 299.7, 4.61e^{-15}	SPi	0.995, 4547, 7.09e^{-29}
XP	0.879, 174.9, 1.62e^{-12}	ABE	0.960, 578.1, 2.6e^{-18}
XPc	0.924, 292.8, 6.0e^{-14}	AIE	0.930, 317.2, 2.4e^{-17}

R^2, F-statistic, and p values are given in the table.

[a] DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔG_{bind}GOLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.

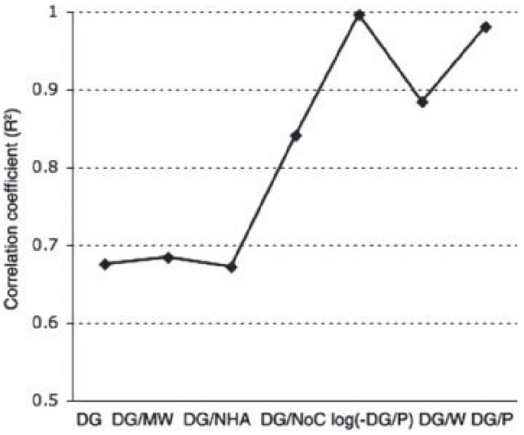


Figure 4. Correlation between experimental and calculated efficiency indices for ΔG, ΔG/MW, ΔG/NHA, ΔG/NoC, log($-\Delta G/P$), $\Delta G/W$, and $\Delta G/P$ for the scoring function component DGb (ΔGbindGOLD) for 26 protein-drug complexes.

the errors between experimental and predicted values were small and similar for both classes of ligands (average differences in RSS between nonpolar and polar ligands: 0.0003 gkcal/mol^2 for ΔG/MW, 0.027 kcal/molNHA for ΔG/NHA, 0.064 kcal/molNoC for ΔG/NoC, 0.009 for log($-\Delta G/P$), and 0.0001 kcal/mol for ΔG/W). Using below and above average PSA to divide the groups, only ABE, AIE and XP showed this effect. The original bias may have risen due to possible overestimation of ligand-protein hydrogen bonding interactions by the scoring functions, or due to inadequate desolvation energy calculation of the nonpolar ligands by the scoring functions.

The drugs shown in Table 1 include large and small size ligands. It is typical for scoring functions to overestimate the binding energy of a compound because they are additive in nature: the larger the ligand, the more protein-ligand interactions it will have. However, the introduction of normalizing ligand efficiency measures allow for the smaller size ligands to be compared positively with larger size ligands because the binding energy is divided by a value which can be related to the molecular size. In a docking or virtual screening experiment, molecules with a large number of carbons are no longer favored. In this way, efficiency indices can repair the errors introduced by the bias of scoring functions toward large size ligands due to errors in the calculation of entropy and desolvation energies. It is promising that the effect was seen on all the scoring functions and programs.

Molecules with extremely high or low values of hydrophobicity or hydrophilicity, that is, with extreme values of P or log P, can be removed through filters before docking. The $\Delta G/P$ efficiency index will also penalize those with borderline values, in addition to having low calculated free energy of binding. Thus, molecules with unfavorable permeability values can be readily detected, in combination with the binding free energy. If ranges for values are established for the different efficiency indices (see for example Hetényi et al.,[33] and also in this present work), these can tell whether a compound's calculated efficiency index is in a favorable range. New efficiency indices can be compared and tested in a manner analogous to the present work.

Conclusions

We have shown that simple ligand efficiency indices can aid the drug design process by providing better comparisons of calculated and experimental values of binding energy. This may increase the accuracy and reliability of docking programs. We observed that efficiency indices also add information to the binding free energy into a single indicator. Permeability of a compound, for example, can be assessed at the same time as the binding affinity in an efficiency index such as $\Delta G/P$, especially if filters are applied to remove compounds with extreme values. Entropy of a compound may be assessed by $\Delta G/W$ values. ΔG/NoC, although simple, is also an effective efficiency index for improving the trend between experimental and calculated values. $\Delta G/P$ and $\Delta G/W$ produced the best results, together with ΔG/NoC. These efficiency indices can be applied across different docking programs, scoring functions, or even components of these. They can also be calculated quickly, likely on the fly. Compounds can be ranked based on efficiency indices that may include data such as absorption and metabolic properties in addition to the free energy of binding, and in this way allow for the selection of molecules that satisfy several criteria in parallel.

Computational Methods

The structures of protein-drug complexes and their experimental inhibition constants (Ki) were collected from the PDBbind database v2005,[38,39] which contains protein-ligand complex structural data from the Protein Data Bank (PDB)[40] as well as experimental Ki determined for those systems. The collection of all small-molecule approved drugs was obtained from the DrugBank database,[41] which contains data on drugs approved by the FDA (U.S. Food and Drug Administration agency). Programs written in Python were used to extract the ligand names (HET-ID) from the PDBbind database and to query them in the DrugBank collection to identify those ligands that are approved drugs. All results were verified visually. The experimental ΔG was computed with $\Delta G = -RT\ln K$, using $T = 25°C$ (298.15 K), and $R = 1.987$ cal/Kmol. The program XLOGP v2.0[42] was used for calculating the octanol/water partition coefficient (log P) by an atom-additive method including correction factors.

Docking programs differ by the scoring functions they contain, as well as the way of minimizing the function values. In our present study, we focused on three main programs that are widely available and used by computational and medicinal chemists: GOLD v.3.1,[43] Glide v.4.5,[44] and Autodock4.[45] Their scoring functions and docking methods are shown in Supporting Information Table S11. GOLD v.3.1[43] uses a genetic algorithm to find the best ligand positioning in a binding site. It can use two scoring functions: Goldscore[43] and Chemscore.[46] Chemscore has a component called ΔGbinding (DGb), which was also used for our correlations. Parameters for runs were: run_flag =

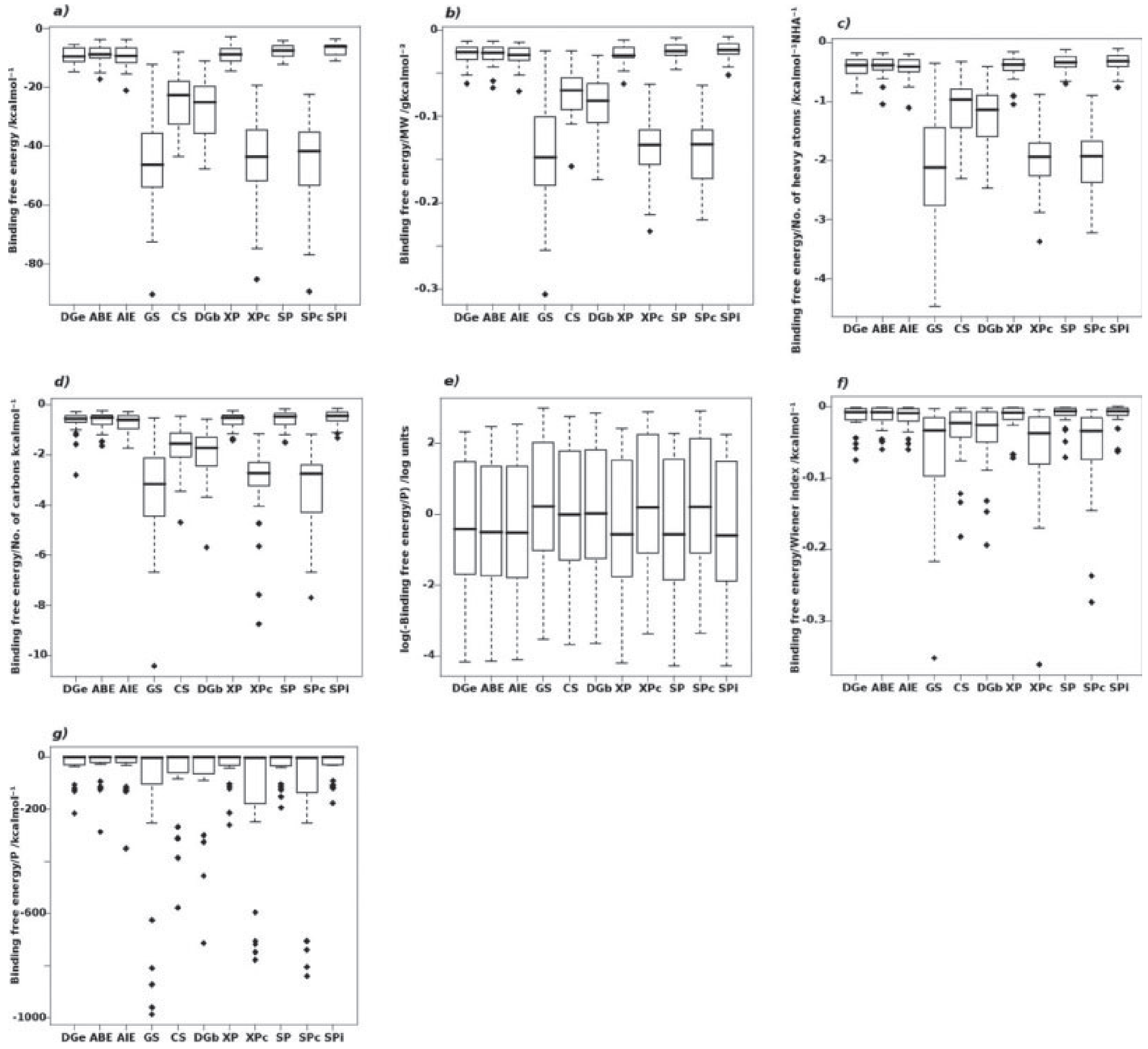


Figure 5. Box plot comparisons of free energies and efficiency indices for experiment and several scoring functions: (a) Free energy of binding (ΔG). (b) Free energy of binding/molecular weight (ΔG/MW). (c) Free energy of binding/number of heavy atoms (ΔG/NHA) efficiency index. (d) Free energy of binding/number of carbons (ΔG/NoC) efficiency index. (e) Logarithm of the (changed sign) free energy of binding/octanol-water partition coefficient ($\log(-\Delta G/P)$) efficiency index. (f) Free energy of binding/Wiener index ($\Delta G/W$) efficiency index. (g) Free energy of binding/octanol-water partition coefficient ($\Delta G/P$) efficiency index. DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindG-OLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.

RESCORE, in addition to default parameters for the genetic algorithm. Waters were switched to ON.

Glide v4.5 (2007) uses a hierarchical search, and has the scoring functions XP and SP,[44] which are a proprietary modification of Chemscore.[46] In addition, we also employed the component C_vdW (a combination of Coulomb and van der Waals terms). Default parameters for runs were used.

Autodock v4.0 also uses a genetic algorithm to find for the best solutions for docked ligands. It uses one scoring function, which produces a binding free energy (ABE).[47] We also employed the component of intermolecular energy (AIE). Parameters used that were different than default values: spacing = 0.375 Å, npts = 40 40 40, ga_pop_size = 150, ga_num_evals = 20,000,000, ga_num_generations = 27,000, tran0 coordinate equal to the "about coordinates", quat0 = 1. 0. 0. 0., and dihe = 0.

Protein and ligand structures already contained hydrogens from the PDBBind dataset. Protein structures were used including the metal atoms and select water molecules that were interacting with protein and ligand in the binding site. The "toggle" setting was used for these special bridge water molecules in GOLD. Docking runs were calculated both including and excluding select crystallographic water molecules. The case which produced a binding energy closest to the experimental was kept. Most of the complexes which included select water molecules had a small effect on the binding energy and efficiency indices as they differed by less than 1 kcal/mol from the "dry" cases in binding free energy, as well as being in the same range and evenly distributed for binding free energy and efficiency indices as the cases without water molecules. The complete list of water molecules is shown in Supporting Information Table S7. Water molecules were included only if they had medium to low crystallographic B-factors, made contacts with the protein, were within 4.5 Å of the ligand, and were at least partially occluded from bulk solvent since these tightly bound water molecules have a higher chance of remaining bound to the protein (remaining conserved in several protein structures),[13–15] and can be considered an integral part of the protein-ligand complex. As such, these specially selected crystallographic water molecules are included in the binding free energy, as well as in the efficiency indices. There was no additional water inclusion or removal when calculating the efficiency indices, which take their binding energy direct from the complex. Efficiency indices are unique to each protein and each ligand in a biomolecular complex, although general trends and ranges can be observed across complexes. Methotrexate (**3**), for example, forms two complexes with different proteins in the dataset, consequently with different binding free energies and efficiency indices. Most of the protein-drug structures which included bridge water molecules mediating their interaction had high crystal structure resolutions, from 1.4 Å and on average lower than 2 Å (median of 1.95 Å) which may increase the probability of detecting reliable water molecule electron density.[48] They included a wide diversity of ligands, though the exposed, shallow complex of the small, relatively nonpolar aminocaproic ligand (**2**) did not have bridging waters, nor did the completely buried dexamethasone (**13**).

Complexes were prepared for the dockings by minimizing in water with generalized Born (GB) implicit solvation and a steepest descent method to a gradient threshold of 239 kcal/molnm, followed by a minimization in water (GB) with a truncated-Newton conjugated gradient method to a gradient threshold of 143.4 kcal/molnm using MacroModel.[49] Statistical tests and box plots were performed using the package R for statistical computing.[50] Marvin Calculator Plug-ins were used for the calculation of ligand molecular formulas and molecular mass (MW).[37]

References

1. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Nat Rev Drug Discov 2004, 3, 935.
2. Coupez, B.; Lewis, R. A. Curr Med Chem 2006, 13, 2995.
3. Klebe, G. Drug Discov Today 2006, 11, 580.
4. Huang, N.; Jacobson, M. P. Curr Opin Drug Discov Dev 2007, 10, 325.
5. Waszkowycz, B. Drug Discov Today 2008, 13, 219.
6. Pierce, A. C.; Jacobs, M.; Stuver-Moody, C. J Med Chem 2008, 51, 1972.
7. Vigers, G. P. A.; Rizzi, J. P. J Med Chem 2004, 47, 80.
8. Chang, C. E. A.; Chen, W.; Gilson, M. K. Proc Natl Acad Sci USA 2007, 104, 1534.
9. Velec, H. F. G.; Gohlke, H.; Klebe, G. J Med Chem 2005, 48, 6296.
10. Pfeffer, P.; Gohlke, H. J Chem Inf Model 2007, 47, 1868.
11. Raub, S.; Steffen, A.; Klamper, A.; Marian, C. M. J Chem Inf Model 2008, 48, 1492.
12. Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. Proteins 2008, 73, 395.
13. García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. J Mol Model 2003, 9, 172.
14. Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. J Mol Biol 2003, 326, 621.
15. García-Sosa, A. T.; Firth-Clark, S.; Mancera, R. L. J Chem Inf Model 2005, 45, 624.
16. Li, Z.; Lazaridis, T. J Phys Chem B 2006, 110, 1464.
17. Li, Z.; Lazaridis, T. Phys Chem Chem Phys 2007, 9, 573.
18. Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Med Res Rev 2006, 26, 531.
19. Amaro, R. E.; Baron, R.; McCammon, J. A. J Comput Aided Mol Des 2008, 22, 693.
20. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Constantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. J Med Chem 2008, 51, 6237.
21. Chen, W.; Chang, C. E.; Gilson, M. K. Biophys J 2004, 87, 3035.
22. Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. J Chem Inf Model 2005, 45, 1134.
23. Wang, J. M.; Kang, X. S.; Kuntz, I. D.; Kollman, P. A. J Med Chem 2005, 48, 2432.
24. Lyne, P. D.; Lamb, M. L.; Saeh, J. C. J Med Chem 2006, 49, 4805.
25. Guimarães, C. R. W.; Cardozo, M. J Chem Inf Model 2008, 48, 958.
26. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Proteins 2005, 60, 325.
27. Warren, G. L.; Andrew, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lamber, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. J Med Chem 2006, 49, 5912.
28. Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. Proc Natl Acad Sci USA 1999, 96, 9997.
29. Hopkins, A. L.; Groom, C. R.; Alex, A. Drug Discov Today 2004, 9, 430.
30. Abad-Zapatero, C.; Metz, J. T. Drug Discov Today 2005, 10, 464.
31. Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. J Med Chem 2008, 51, 2432.
32. Hetényi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. J Am Chem Soc 2006, 128, 1233.
33. Hetényi, C.; Maran, U.; García-Sosa, A. T.; Karelson, M. Bioinformatics 2007, 23, 2678.
34. García-Sosa, A. T.; Sild, S.; Maran, U. J Chem Inf Model 2008, 40, 2074.
35. Wells, J. A.; McClendon, C. L. Nature 2007, 450, 1001.

36. Leeson, P. D.; Springthorpe, B. Nat Rev Drug Discov 2007, 6, 881.
37. Marvin v4.8.1. 2007. ChemAxon. Available at: http://www.chemaxon.com (accessed on October 30, 2008).
38. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. J Med Chem 2004, 47, 2977.
39. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C.-Y.; Wang, S. M. J Med Chem 2005, 48, 4111.
40. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
41. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Nucleic Acids Res 2006, 34, D668.
42. Wang, R. X.; Gao, Y.; Lai, L. H. Perspect Drug Discov 2000, 19, 47.
43. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Mol Biol 1997, 267, 727.
44. Schrödinger, LLC. Glide Version 4.5; Schrödinger, LLC: New York, 2007.
45. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. J Comput Chem 1998, 19, 1639.
46. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. J Comput Aided Mol Des 1997, 11, 425.
47. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. J Comput Chem 2007, 28, 1145.
48. Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Angew Chem Int Ed Engl 2003, 42, 2718.
49. Schrödinger, LLC. Macromodel v9.5; Schrödinger, LLC: New York, 2007.
50. The R Project for Statistical Computing. Available at: http://www.r-project.org (accessed on October 30, 2008).

Molecular Property Filters Describing Pharmacokinetics and Drug Binding

A.T. García-Sosa[1], U. Maran[1] and C. Hetényi*[,2]

[1]*Institute of Chemistry, University of Tartu, 14A Ravila, 50411 Tartu, Estonia*

[2]*Departments of Biochemistry and Genetics, Eötvös University, Pázmány sétány 1/C, 1117 Budapest, Hungary*

Abstract: Drug-target binding affinity and pharmacokinetics are equally important factors of drug design. Simple molecular properties such as molecular size have been used as pharmacokinetic and/or drug-likeness filters during chemical library design and also correlated with binding affinity. In the present study, current property filters are reviewed, a collection of their optimal values is provided, and a statistical framework is introduced allowing calibration of their selectivity and sensitivity for drugs. The role of ligand efficiency indices in drug design is also described. It is concluded that the usefulness of property filters of molecular size and lipophilicity is limited as predictors of general drug-likeness. However, they demonstrate increased performance in specific cases, e.g. in central nervous system diseases, emphasizing their future importance in specific, disease-focused library design instead of general drug-likeness filtering.

Keywords: Binding site, entropy, free energy, molecule, pocket, protein, structure, target, logP, Wiener index.

1. INTRODUCTION

The effects of drug molecules are produced by their interactions with one or more macromolecular targets [1, 2], constituents of the human body. Therefore, small molecule drug design strategies involve multiple screening steps [3, 4] using the structure [5] of drug candidates (ligands) in complex with targets and also the corresponding thermodynamic measures of equilibrium binding affinities [6], the free energy changes (ΔG). In general, an appropriate ΔG is a necessary but not a sufficient property of a successful candidate as pharmacokinetic, toxicological, etc. characteristics also influence drug-likeness [7].

Molecular properties of small compounds have been extensively used as descriptors in structure-activity relationships [8, 9]. For example, molecular weight (MW) is atom-type sensitive and related to the molecular size; logP is a measure for partitioning of compounds between lipophylic and aqueous phase; number of heavy atoms (NHA) is the simplest molecular property providing a crude estimate of the size of a molecule; Wiener index, a topological descriptor characterizes the compactness of a molecule and is proportional to the molecular surface area [10, 11]. Such molecular properties were also adopted for the prediction of complex physiological properties and pharmacokinetics: absorption [12], or blood brain barrier penetration [13, 14], and their use culminated in the definition of general drug-likeness ranges. These empirical ranges of the properties were proved to be useful as property filters in the design of compound libraries of drug screening [15-20]. Notably, the selection of high quality (drug-like) compound libraries [3, 21-23] is a primary and key step of the screening process.

Besides their connection to pharmacokinetics, it has been shown in numerous studies that the above size-dependent filters (MW, NHA) are also coupled to ΔG as they correlate with the (maximal) binding affinity achievable by a ligand. To decouple ΔG from ligand size, efficiency indices (EI, also called ligand efficiencies or binding efficiencies) have been defined dividing ΔG by NHA or MW [24].

The present review sketches how complex phenomena of pharmacokinetics and equilibrium binding are coupled with the above molecular properties. An overview of their use is provided, and a summary of available correlations of ligand-based properties with ΔG is assembled. The role of EIs is discussed, and limitations of the general drug-likeness concept are analyzed. Selectivity and sensitivity of the property filters are defined, and a statistical decoupling of ΔG from the properties is suggested for pharmacokinetics-focused analyses. Besides general drug-likeness, disease- and target-specificity is discussed and future perspectives are outlined.

2. MOLECULAR PROPERTY FILTERS DESCRIBING DRUG-LIKENESS

Filtering of large compound sets generated by combinatorial or other techniques [25, 26] is a central issue of library design. As Martin and Critchlow showed [27], merely random selection of compounds for high throughput screening (HTS) is poor both in structural diversity and in distribution of physicochemical properties. Random libraries are systematically biased toward heavy, flexible compounds that have very high or very low lipophilicity and possess inappropriate bioavailability. Thus, the need for effective filtering to produce 'drug-like' libraries was early recognized and several groups have developed filters based on the analysis of molecular property distribution in available drug databases. The present paper is focused on the analysis of simple molecular properties such as MW or logP coupled to both pharmacokinetics and ΔG (Introduction). Other filters including information on e.g. functional groups [28] are beyond the scope of this study.

2.1. Definition of Drug-Likeness

The first drug-likeness studies dealt with pharmacokinetic properties of drug candidates. Lipinski *et al*. [29] found that poor absorption or permeation is more likely if ligand properties such as MW or logP fulfill the 'rule of 5' (Ro5, Table **1**) criterion. Fecik *et al*. [30] also analyzed the relationship between MW and oral bioavailability. Clark and Pickett [31] describe the term general drug-likeness filtering. According to their definition, such filters incorporate substructure searches for toxic or reactive groups and/or include limits on molecular properties which may be generally useful in drug design, i.e. non-specific for disease types. Other early reviews [28] also use the phrase drug-likeness for "*molecules which contain functional groups and/or have physical properties consistent with the majority of known drugs*". Muegge [19] remarked that "*Drug-likeness is mostly a statistical descriptor derived from databases of other compounds. It should, therefore, be used to evaluate the drug-likeness of other compound selections such as screening libraries, combinatorial libraries, or virtual libraries rather than that of a single compound.*" Taking into account the general opinion formulated by the above studies the drug-likeness paradigm in the present review can be classified as (i) general drug-likeness (all diseases and mostly oral drug administration); and (ii) specific drug-likeness (classified by disease, administration, target, etc.).

*Address correspondence to this author at the Departments of Biochemistry and Genetics, Eötvös University, Pázmány sétány 1/C, 1117 Budapest, Hungary; Tel: +36-13812173; Fax: +36-13722641; E-mail: csabahete@yahoo.com

1875-533X/12 $58.00+.00

Table 1. General Drug-Likeness Values of Property Filters

Source			Statistics	Property									Database	
Year	Author	Ref		NCC	HBD	HBA	logP	MW	NHA	NR	NRB	PSA	Description	N
1997	Lipinski *et al.* (Ro5)	[12]	~90P		5	10	5(c) 4.15(m)	500					World Drug Index (WDI) filtered by USAN, INN names, etc.	2245
1999	Ghose	[33]	MEAN				2.3(a)	357					CMC	6304
			SD				2.6(a)	174						
2001	Sakaeda *et al.*	[34]	MEAN				1.73(c) 1.94(m)	332					Oral drugs	222
			SD				2.21(c) 2.03(m)	140						
2003	Feher and Schmidt	[39]	MEAN	2.3	1.9	5.7	2.2(s)	340	23.5	2.6	5.6		Chapman and Hall Dictionary of Drugs, The Merck Index	10968
			MED	1	1	5	2.3(s)	312	22	2	5			
2003	Wenlock *et al.*	[40]	MEAN		2.1	4.9	2.5(c)	337			5.9		The Physicians' Desk Reference 1999	594
			SD		2.4	3.6	2.5(c)	157			4.5			
			90P		4	8	5.5(c)	473			11			
2004	Leeson and Davis	[42]	MEAN		1.81	5.14	2.27(c)	331		2.56	4.97	21.1%	Oral drugs pre-1982	864
			MED		1	4	2.31(c)	310		3	4	18.5%		
2004	Leeson and Davis	[42]	MEAN		1.77	6.33	2.50(c)	377		2.88	6.42	21.0%	Oral drugs 1983-2002	329
			MED		1	6	2.36(c)	357		3	6	19.4%		
2004	Vieth *et al.*	[43]	MEAN		1.8	5.5	2.3(c)	343.7		2.6	5.4	78	FDA Orange Book	1193
			90P		3	9	5.2(c)	475		4	10	134		
2004	Vieth *et al.*	[43]	MEAN		3	4.5	2.5(c)	300					Lipinski *et al.* recomputed based on the 2001 edition of the WDI	1791
			90P		3	8	5.3(c)	427.5						
2005	Proudfoot	[44]	MEAN		1.5	5.1	2.5	333					Oral drugs 1937-1997	1791
			90P		3	9	4.8	469						
2006	Vieth and Sutherland	[48]	MEAN		1.8	5.5	2.3(c)	345					Vieth *et al.* 2004 updated with FDA release after 2003	1210
			90P		4	9	5.3(c)	478.4						
2009	Tyrchan *et al.*	[49]	MEAN		1.5	3.9	2.74(c)	335.5			5.6	64.7	GVKBIO, IBEX	976
			SD		1.5	2	2.22(c)	109.2			3.6	39.7		
			MED		1	4	2.83(c)	318.5			5	59.1		

Abbreviations. 90P: 90th percentile; HBA: number of H-bond acceptors (O+N); HBD: number of H-bond donors (OH+NH); logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different methods of logP calculation); MED: median; MW: molecular weight; N: number of drugs in database; NCC: number of chiral centers; NHA: number of heavy atoms; NR: number of rings; NRB: number of rotatable bonds; PSA: polar surface area; SD: standard deviation.

2.2. General Drug-Likeness

Ajay *et al.* [32] investigated the possibility of distinction between general drug-likeness and non-drug-likeness by one- or two-dimensional descriptors within neural network-based models. They used the Comprehensive Medicinal Chemistry (CMC) and the MACCS-II Drug Data Report (MDDR) as drug-like data sets and the Available Chemicals Directory (ACD) as a surrogate for non-drugs. It was correctly remarked that using the above databases as drug/non-drug collections is an assumption as "the characteristics of drug molecules today may change in the future". Therefore, the conclusions of dataset-based drug-likeness studies may always reflect the actual state of the common knowledge on drug-likeness and *a priori* include errors. This study can be regarded as a key analysis, which included not only drugs, but also quasi non-drugs in a truly comparative manner.

However, most of the studies reporting drug-likeness thresholds (Table **1**) deal only with drug (lead or bioactive) databases. Ghose *et al.* [33] based their analysis on the CMC database and provided drug-likeness thresholds for MW and logP. They also concluded the priority of some fragments (e.g. benzene ring) occurring in drug structures. The analysis of 222 commercially available oral drugs by Sakaeda *et al.* [34] supported the Ro5. However, the authors also remark that compounds with a sugar moiety, high atomic weight, and/or large cyclic structure were exceptions to the MW=500 upper threshold. Veber *et al*. [35] also found that molecular weight cutoff at 500 does not itself significantly separate compounds with poor oral bioavailability from those with acceptable values. They analyzed the oral bioavailability of a large data set in rats containing more than 1000 compounds. It was also concluded that compounds of possibly good oral bioavailability possess 10 or fewer rotatable bonds (NRB) and polar surface area (PSA) equal to or less than 140 Å^2 or 12 or fewer H-bond donors (HBD) and acceptors (HBA). Their analysis on artificial membrane permeation rates showed that reduced PSA correlated better with increased permeation rate than did ClogP, and an increased NRB had a negative effect on the permeation rate. Lu *et al*. [36] also investigated the predictive power of NRB and PSA on 434 Pharmacia compounds and found that their correlations with bioavailability depended on the therapeutic class.

Hann *et al*. [37] studied the differences in the properties of drug leads and optimized compounds. The data indicates that, on average, drug leads have lower MW, lower ClogP, fewer aromatic rings (NR), fewer HBA than the corresponding drugs. On the contrary, Proudfoot [38] found that most drugs are within 25% of the lead values with regard to MW, and nearly all are within one calculated MLogP unit.

In another interesting comparative analysis of drugs, natural products and combinatorial libraries Feher and Schmidt [39] also emphasized the importance of properties beyond the often used MW and logP. For example, it was shown that the 'number of chiral centers' in a molecule has a great impact on its drug-likeness. They found that while chiral centers are normally present in drug and natural product molecules, they tend to diminish in combinatorial compounds, which is most probably a consequence of the oversimplified synthetic/construction steps in the generation of combinatorial libraries.

Wenlock *et al*. [40] compared distributions of physico-chemical properties such as MW and logP of marketed oral drugs and of compounds in development. In their analysis, the mean MW of orally administered drugs in development decreased on passing through each of the different clinical phases and gradually converged towards the mean molecular weight of marketed oral drugs. In addition, the most lipophilic compounds diminished during development. They compared upper property thresholds below which 90 % of oral drugs in their data set with the results of the Ro5, and good agreement was found (Table **1**).

Besides the thresholds values, the historical trends of, e.g. MW of drug candidates, may be also useful as collected by Lipinski [41] for the period 1960–2004. It was demonstrated that advanced clinical candidates produced by a "rational drug design" approach of Merck had a time-dependent higher MW, higher H-bonding properties, unchanged logP, and poorer permeability. Early candidates from a HTS-based approach of Pfizer (Groton, CT) had higher molecular weight, unchanged H-bonding properties, and higher logP, i.e. poorer aqueous solubility. In another retrospective study, Leeson and Davis [42] showed that mean values of lipophilicity, percent of PSA and HBD had not changed in the period of 1983-2002. In contrast, mean values of MW and the numbers of O + N atoms, HBA, NRB, and number of rings have increased by 13-29%. Similarly, Vieth *et al*. [43] demonstrated that the mean property values for oral drugs do not vary substantially with respect to launch date. The limited change in the most important oral drug-like property values lead the authors to suggest that the range of acceptable oral properties is independent of the synthetic complexity or targeted receptor. Proudfoot [44] analyzed the very long period of 1937-1997. During this period a steady increase was observable in mean and median MW. Only seven marketed drugs with MW>500 were designed in the 15 year period 1937–1951, and thirty two in the comparable period 1983–1997. Mean and median logP was unchanged in the 60 year period examined. Fewer than 5% of oral marketed drugs had more than 4 H-bond donors and just 2% had MW>500 and >3 H-bond donors. An analysis by Leeson and Springthorpe [45] suggested that clogP is the most important molecular property, as it is changing less over decades in launched oral drugs than other properties. As ClogP plays a dominant role in promoting binding to unwanted drug targets, a high logP therefore carries increased risks of developmental attrition. They conclude that a 5% improvement in attrition would double the output of new medicines and that this might be achieved simply by lowering logP. Comparing sets of drugs and their originating leads, Perola [46] also found that on average, the two sets have similar logP, suggesting that the ability to maintain low levels of logP while increasing MW is one of the keys to a successful drug discovery program.

Schneider *et al*. [47] investigated the combined use of drug-likeness property filters in gradual filtering by decision trees. With rapidly computable properties such as MW, XlogP, molar refractivity, and several drug-likeness indices, up to 76% of all non-drugs could be sorted out in the first filtering step. With the aid of sophisticated (quantum chemical) properties in the succeeding steps up to 92% of the initial non-drugs were filtered out, while less than 19% of the actual drugs were lost. In addition to the above examples, Table **1** also lists threshold values given by Vieth and Sutherland [48] and Tyrchan *et al*. [49].

2.3. Limitations of the General Drug-Likeness Concept

Although physicochemical properties are widely used as general drug-likeness filters (Section 2.2), there are several articles pointing to their limitations. As Walters *et al*. [28] envisioned, instead of dealing with the complex problem of drug-likeness, a viable alternative is the prediction of the various pharmacokinetic properties (logP, half-life, plasma protein binding, etc.) that contribute to a drug's success. Remarkably, even the calculation and modeling of these properties themselves is rather complex [50] and extremely difficult in many cases.

The lack of validated sets of drugs and decoy sets of non-drugs [51] also limits the usefulness of any drug-likeness filters as there are compounds, e.g., that can easily fall into either category. Moreover, the filters can only recognize those compounds that resemble existing drugs as drug-like – compounds from completely new classes could be misclassified [31]. Remarkably, the original publication of Lipinski [29], root of many others in this field, addressed the prediction of *only* pharmacokinetic properties (absorption and permeation) and *not* general drug-likeness.

However, collecting sets of good and bad pharmacokinetic properties remains a challenge for property filters due to the above-mentioned complexity of the properties themselves. In addition, the final decision on drug-likeness is just further postponed if a filter can provide information only on one drug-likeness property. In fact, there are several properties to be predicted which can easily give controversial results in ranking of a compound or a library and it is still unclear which property should be prioritized for the final decision, etc. For example, Kubinyi [52] finds that "*inappropriate ADME (Absorption, Distribution, Metabolism, Excretion) characteristics have clearly made far less of a contribution to clinical failures than is widely supposed!*". At the same time, he also accepts that the application of the Ro5 aimed at prediction of "A" of ADME significantly aided improving early combinatorial libraries which had included "*many large and greasy, biologically inactive molecules*". This example of the controversial judgment of the fairly well-studied ADME properties illustrates that it would be indeed very difficult to set the above-mentioned priority order of properties in a decision tree. The questions on the appropriate use of a property, i.e., "*where and to which extent*" seem to remain unanswered in general.

Similarly, an important study by Feher and Schmidt analyzing properties of natural products [39] concluded that: "Drug-like filters, such as the Lipinski rules, are very helpful in isolating likely problem molecules. However, overly strict adherence to it can have the adverse effect of restricting diversity … and hence also reducing similarity to natural products. … A large proportion of natural products is biologically active and has favorable ADME/T properties, despite the fact that they often do not satisfy 'drug-likeness' criteria." Furthermore, Ganesan [53] analyzed a total of 24 unique natural products that led to an approved drug in the period 1970–2006. They found an identical success rate of 50% both for the classes conforming or violating the Ro5. It was also found that natural products are successful in maintaining favorable logP and intermolecular H-bond donating potential even with high MW and large numbers of rotatable bonds.

Lajiness *et al* [54] raise additional concerns regarding drug-likeness studies. They claimed that there are very few studies accompanied by the data sets used for analysis, and therefore, reproducibility of the results is questionable. During collection of data in Table **1**, we also found that in many cases authors refer to, e.g. in-house, company-owned data sets or other resources with no or reduced public availability or a non-defined sub-set of an available database. However there is no guarantee that proprietary collections are adequate for the analysis of general drug-likeness. For example, Lajiness *et al*. [54] mentioned that proprietary collections may be biased due to historical lead optimization efforts focused at particular chemical classes, such as steroids or benzodiazepines. They also concluded that comparing drug-likeness of groups instead of individual compounds was appropriate to achieve significant results.

There are also methodological problems with the properties 'traditionally' used as filters. For example, Bhal *et al*. [55] suggest the cautious use of logP in drug design due to its inability to account for the ionization of compounds under physiological conditions. They conclude that the pH-dependent logD is a more realistic descriptor of lipophilicity under physiological pH's and, therefore, logD should be used preferentially over logP as the descriptor for lipophilicity, especially when working with ionizable compounds. Vistoli *et al*. [51] also mention the problems of pH-dependent properties.

In their seminal paper, Lipinski *et al*. [29] already claimed that antibiotics, antifungals, vitamins, and cardiac glycosides fell outside their Ro5, possibly due to transporter effects. The results of the study of Good and Hermsmeier [56] suggest further discontinuities in drug-like space, beyond those claimed by Lipinski *et al*. [29], in the context of classification. Giménez *et al*. [57] also concluded that Ro5 is very useful to select better compounds in chemical libraries, but it must be used carefully to avoid a possible exclusion of promising compounds. They evaluated the top pharmaceutical products in 2007. Among 60 drugs, 7 (atorvastatin, montelukast, docetaxel, telmisartan, tacrolimus, leuprolide and olmesartan) did not fit the Ro5, and 5 failed one of the threshold values.

Zhang and Wilkinson [58] summarized their criticism of the overemphasis of Ro5 of drug-likeness from two points of view. Firstly, they claim that only 51% of all FDA-approved small molecule drugs are both used orally and comply with the Ro5. This does not even include the increasing number of biologicals of which several have reached 'blockbuster' status. Secondly, the Ro5 does not cover natural product and semisynthetic natural product drugs, which constitute over one-third of all marketed small-molecule drugs (see also Feher and Schmidt [39]).

A further doubt arises from the finding (Dobson and Kell [7]) that general drug-likeness properties such as MW or logP, adequate for passive diffusion, have decreased ability for prediction of carrier-mediated and active uptake of drugs that are more common forms of transport than is usually assumed. For drugs transported by carriers, general property filters are not normally effective in individual cases, and specific data on interactions of drugs and transporters would therefore accelerate research in this field. Similarly to drugs, naturally occurring intermediary metabolites may also require solute carriers to enter cells. Thus, an evaluation of metabolite-likeness (Dobson *et al*.) [59] would be essential to understand the true physiological processes. However, estimation of metabolite-likeness is missing from most of the present drug-likeness studies.

2.4. Specific Drug-Likeness

Considering the diversity of drug profiles, specific approaches of drug-likeness may become an alternative to the limited general concept reviewed in the previous Sections. Drugs achieve their effects through different mechanisms in the body, targeting different proteins, organs or even organisms, as in the case of anti-infective agents. Moreover, dermatological agents used topically may require completely different pharmacokinetic properties than drugs which are inhaled, injected or administered orally. In addition, drugs that affect the central nervous system have to pass yet another obstacle, the blood-brain barrier (BBB).

Besides their general analysis, (Section 2.2) Ghose *et al*. [33] also investigated the property profile (MW, logP, etc.) of seven different classes of drug molecules in the CMC such as central nervous system (CNS), cardiovascular, cancer, inflammation, and infectious diseases (Table **2**). They provided drug-likeness ranges for the different classes and found considerable outliers from the general drug-likeness trend. For example, the antibacterial compounds formed a special class of biologically active compounds very different from regular drugs. The logP of anticancer drugs showed a high standard deviation possibly due to the complexity of cancer, which affects different parts of the body and tissues. On the other hand, the standard deviation of logP of CNS drugs was relatively small due to the requirement that they should cross the BBB. They concluded that for different drug classes the ranges may be considerably tighter than the general drug-likeness ranges. Leeson and Davis [42] also found that significant differences exist between the property distributions of different therapeutic areas of oral drugs of the 1983-2002 period. The distributions of MW and logP among antiinfectives show different trends from the other drug classes probably related to the need for their activity in a non-human organism, and cell wall penetration in the case of antibiotic drugs.

Vieth *et al*. [43] analyzed the differences between routes of administration (Table **2**). It was observed that oral drugs tend to be lighter and have fewer H-bond donors, acceptors, and rotatable bonds than drugs with other routes of administration. These differences are particularly pronounced for oral vs. injectable drugs. However, they concluded that due to the substantial overlap in the range of properties found between the different drug classes, a particular drug cannot be adequately classified as either oral or injectable on the basis of simple physical property calculations. Tronde *et al*. [60] have studied the physicochemical properties and absorption qualities of inhaled drugs, finding that the pulmonary epithelium allows for higher PSA (up to 479 Å^2) in compounds, as compared to the intestinal mucosa and BBB. They propose the lung route as an alternative to drugs poorly absorbed through the oral route. Ritchie *et al*. [61] also studied respiratory drugs administered through intranasal/inhaled routes, and found their calculated physicochemical properties to have lower lipophilicity, higher molecular weight, and higher PSA, when compared to drugs administered orally.

Table 2. Specific Drug-Likeness Values of Property Filters

Source			Disease/administration /target family	Statistics	Property							Database	
Year	Author	Ref			HBD	HBA	logP	MW	NR	NRB	PSA	Description	N
1999	Ghose *et al.*	[33]	cancer	MEAN			1.59(a)	332				CMC	349
				SD			2.5(a)	129					
			cardiovascular/ antiphypertensive	MEAN			1.97(a)	361					269
				SD			2.1(a)	123					
			CNS/antidepressant	MEAN			3.05(a)	291					208
				SD			1.5(a)	69					
			CNS/antipsychotic	MEAN			4.10(a)	380					105
				SD			1.5(a)	83					
			CNS/hypnotic	MEAN			2.20(a)	277					74
				SD			1.5(a)	99					
			infection	MEAN			2.38(a)	339					39
				SD			2.7(a)	139					
			inflammation	MEAN			3.09(a)	335					290
				SD			1.5(a)	122					
2004	Leeson and Davis	[42]	Cancer	MEAN	1.00	4.5	3.02(c)	313	2.36	5.00	20.8 %	Oral drugs 1983-2002	14
				MED	1	4.5	3.01(c)	299	2	3.5	18.3 %		
			cardiovascular	MEAN	1.46	6.73	3.05(c)	389	2.84	8.23	19.8 %		79
				MED	1	7	3.00(c)	396	3	8	18.6 %		
			gastrointestinal and metabolism	MEAN	2.71	6.84	1.90(c)	378	2.32	7.63	26.7 %		38
				MED	2	6	2.28(c)	357	2.5	7	20.7 %		
			infection	MEAN	2.41	8.78	1.56(c)	456	3.45	6.83	24.6 %		64
				MED	2	7	0.94(c)	389	3	5	21.5 %		
			nervous system	MEAN	1.50	4.32	2.50(c)	310	2.85	4.70	16.3 %		74
				MED	1	4	2.55(c)	307	3	4.5	14.3 %		
			respiratory and inflammation	MEAN	1.37	4.24	3.34(c)	396	3.02	5.52	20.5 %		46
				MED	1	4	2.90(c)	353	3	4.5	19.3 %		
2004	Vieth *et al.*	[43]	absorbent	MEAN	3	6.5	1.6(c)	392.3	2.5	7.9	100.5	FDA Orange Book. Drugdex	116
				10-90P	0-7	2-14	-2.3 to 4.8(c)	172-666	0-4	2-16	20-219		
			injectable	MEAN	4.7	11.3	0.6(c)	558.2	3.2	12.7	143.6		308
				10-90P	0-11	3-23	-3.3 to 4.9(c)	196-1085	1-6	2-27	28-311		
			topical	MEAN	1.9	5	2.9(c)	368.5	2.9	5.3	75.4		112
				10-90P	0-3	2-8	-0.6 to 6.0(c)	188-495	1-5	1-9	21-114		
2006	Vieth and Sutherland	[48]	CYP450	MEAN	0.7	2.9	3.4	300.5				Vieth *et al.* 2004 updated with FDA release after 2003	12
				90 %	2	5	8.8	399.4					
			GPCR-bio	MEAN	1.3	4.2	2.8	326.8					216
				90 %	3	7	5.1	435.4					
			GPCR-lipid	MEAN	1.8	5.0	5.5	414.9					8
				90 %	3	9	8.5	586.2					
			GPCR-pep	MEAN	1.6	8.5	5.0	484.8					11
				90 %	2	12	7.5	600.2					
			ion channel	MEAN	1.3	4.9	2.5	305.5					115
				90 %	2	9	5.0	443.2					
			kinase	MEAN	2	7.0	4.6	439.4					5
				90 %	3	8	5.6	493.6					
			NHR	MEAN	1.4	3.8	4.1	381.8					58
				90 %	3	6	7.2	445.8					
			PDE	MEAN	0.9	6.9	1.7	331.9					15

(Table 2). Contd…..

Source			Disease/administration /target family	Statistics	Property							Database	
Year	**Author**	**Ref**			**HBD**	**HBA**	**logP**	**MW**	**NR**	**NRB**	**PSA**	**Description**	**N**
				90 %	2	10	4.2	480.2					
			protease	MEAN	4.5	7.2	2.3	430.6					35
				90 %	5	11	5.9	636.6					
			transporter	MEAN	1.3	4.2	3.0	304.7					37
				90 %	3	7	5.5	423.5					
1999	Ajay *et al.*	[64]	CNS	MEAN			2.8	354				CMC and MDDR	1050 + 16785
				MED			2.9	351					
				90 %			0.0-5.2	200-540					
1999	Kelder *et al.*	[65]	CNS	MAX							120 Å^2	Passively transported oral drugs	776
				~MEAN							60-70Å^2		
2009	Chico *et al.*	[69]	CNS	~MAX			4	400			80 Å^2	Brain-penetrant small molecules	448
2001	Sakaeda *et al.*	[34]	CNS	MEAN			2.67(c) 2.80(m)	285					44
				SD			2.03(c) 1.98(m)	91					
			inflammation	MEAN			2.63(c) 2.66(m)	279					17
				SD			1.37(c) 1.47(m)	107					
			microbial	MEAN			-0.18(c) -0.13(m)	371					48
				SD			1.88(c) 1.59(m)	161					

Abbreviations. 90P: 90th percentile; HBA: number of H-bond acceptors (O+N); HBD: number of H-bond donors (OH+NH); logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different methods of logP calculation); MED: median; MW: molecular weight; N: number of drugs in database; NHA: number of heavy atoms; NR: number of rings; NRB: number of rotatable bonds; PSA: polar surface area; SD: standard deviation.

Another study of Vieth and Sutherland [48] investigated the distribution of drug-likeness property filters by targeted proteomic families. For proteases, nuclear hormone receptors, lipid and peptide G-protein-coupled receptors (GPCRs), the corresponding drugs significantly exceed Ro5 limits, while others targeting cytochrome P450s, biogenic amine GPCRs, and transporters had significantly lower values for certain properties. It is also an interesting question whether ligands targeting different proteomic families have statistical difference in their property ranges. According to the results of Morphy [62], the ligands of peptide GPCRs and integrin receptors, possess significantly higher median property values than those for aminergic targets, such as monoamine transporters and GPCRs. Agonists for monoamine GPCRs, opioid receptors and ion channels had smaller MW and clogP than the antagonists, but there was no difference between the agonists and the antagonists for peptide GPCRs and nuclear receptors. Paolini *et al.* [63] also found distinct differences in the distribution of molecular properties between sets of compounds active against different families. For example, they also found that the mean MW of ligands binding to aminergic GPCRs is 378(±93), whereas the mean MW of peptide GPCR ligands is greater at 514(±202).

The design of libraries of CNS-active compounds has been a goal of many research groups since the early applications of drug-likeness property filters (Ajay *et al.* [64]) such as MW or logP. Kelder *et al.* [65] found a significant difference in the polar surface area distribution of 776 CNS and 1590 non-CNS drugs. It was concluded that orally active drugs with passive transcellular transport should not exceed a PSA of 120 Å^2, and a 60-70 Å^2 for appropriate BBB permeability. MW was identified as a good descriptor of BBB penetration [66, 67], and applied in fact as a key property filter together with logP in testing a 3042 compound screening library [68] for CNS-compatibility. Chico *et al.* [69] claimed that kinase inhibitor drugs for CNS indications required a modification of the property limits set by the Ro5. They found that most of the brain-penetrating small molecules had a MW<400, logP<4 and PSA<80Å^2. In addition to above examples Table **2** provides also threshold values fro three disease families by Sakaeda *et al.* [34].

3. MOLECULAR PROPERTY FILTERS DESCRIBING BINDING AFFINITY

Besides the use of molecular properties (MW, NHA, logP, etc.) as filters (see previous Sections), several studies investigated the correlation between these properties and the binding affinity of a ligand to its macromolecular target (Eq. 1). Remarkably, during the formation of the [Ligand:Target] complex some water molecules (k in Eq. 1) may leave the binding interface, whilst others may join it [70, 71]. The binding affinity (also called '*in vitro* potency') can be described in terms of thermodynamic equilibrium constants of association, dissociation or inhibition (K_a, K_d. K_i) which can be related to ΔG (Eq. 2). In some cases, the logarithm of inhibitor concentration at 50 % inhibition (pIC_{50}) is also applied as a measure of binding affinity, but pIC_{50} cannot be directly related to ΔG by Eq. 2.

$$\text{Ligand}(H_2O)_n + \text{Target}(H_2O)_m \rightleftharpoons [\text{Ligand:Target}](H_2O)_{n+m-k} + k\,H_2O \qquad \text{Eq. 1}$$

$$\Delta G = -RT\ln K_a = RT\ln K_{d/i} \qquad \text{Eq. 2}$$

(R is the gas constant, T is the thermodynamic temperature)

In a seminal article Kuntz *et al.* [72] plotted experimental ΔG values of a large set of complexes of macromolecular targets and their strongest-binding ligands against the NHA of the ligand molecules. They found that ΔG increases with NHA with an initial slope of ca. -1.5 kcal/mol (1cal = 4.18 J) per atom. Beyond 15 NHAs the increase dropped dramatically suggesting a logarithmic relationship between ΔG and NHA for large molecules. Reynolds *et al.* [73, 74] found a similar, non-linear relationship when plotting the most potent ligands of the BindingDB database. The 'maximal affinities' as measured by pIC_{50} increased rapidly up to 20 heavy atoms, but a plateau existed beyond 25. A recent study of Ferenczy and Keserű [75] also presented a non-linear plot of pK_d-NHA with a plateau starting from 40 heavy atoms.

Ferrara *et al.* [76] calculated the Pearson R value between the experimental ΔG and the logarithm of the MW for different data sets (Table **3**) and found significant correlations in many cases. The logarithmic function was chosen according to the above detailed logarithmic dependence of ΔG on NHA shown by the study of Kuntz *et al.* [72]. Velec *et al.* [77] also calculated a Spearman's rank order correlation coefficient of 0.56 between experimental ΔGs of 100 complexes and the MWs of participant ligands. Affinity predictions purely based on the ligand's MW gave in fact better results for the 100 complexes than many scoring functions involving other terms on, e.g. interaction with the target. Wells and McClendon [78] collected the ΔG of highest-affinity fragments and small molecules that target seven different protein–protein interfaces, and found an R=0.77 (R^2=0.59) correlation between ΔG and NHA. Kim and Skolnick [79] published correlations between pK and logMW values of various data sets (Table **3**).

Olsson *et al.* [80] measured a considerable correlation of ΔG with apolar surface area burial (including both ligand and protein surface) upon complex formation (R^2=0.65) and the change in ligand apolar solvent accessible surface area (ASA, R^2=0.44) using a diverse set of 254 complexes of the SCORPIO database. Notably, binding pocket ASA was shown [81] to correlate with ligand MW at an R^2=0.77 too. For peptide ligands, estimation of ΔH was considered using a linear combination of ΔASA values [82].

The background of the correlations of ΔG (logK) with ligand-based, size-dependent properties (MW and NHA, Table **3**) has not been elucidated yet. According to Eq. 3, for the analysis of correlation of the properties with ΔG it may be a plausible idea to analyze their correlations with the binding enthalpy (ΔH) and entropy (ΔS) changes, respectively. Using the data set published by Reynolds and Holloway [83], no correlation can be observed between NHA and ΔH or TΔS, respectively, but with ΔG, a slight R^2=0.28 can be calculated. This finding hints that such a dissection of ΔG into ΔH and TΔS may not help in finding the reasons of the correlations of Table **3**.

$$\Delta G = \Delta H - T\Delta S \qquad \text{Eq. 3}$$

Table 3. Correlations Between Binding Affinity and Molecular Properties

Source			Correlated quantities		R^2	Database	
Year	**Author**	**Ref**	**Binding affinity**	**Property**		**Description-target protein**	**N**
2004	Ferrara *et al.*	[76]	pK_i	logMW	0.36	LPDB-all	189
					0.23	LPDB-oxidoreductase	37
					0.81	LPDB-serine protease	25
					0.58	LPDB-metalloprotease	13
					0.50	LPDB-immunoglobulin	10
					0.18	LPDB-lyase	8
					0.16	LPDB-L-arabinose binding protein	9
2005	Velec *et al.*	[77]	pK_d	MW	0.31^a	Wang *et al.*	100
2007	Wells and McClendon	[78]	ΔG	NHA	0.59	Ligands of seven different targets	13
2008	Kim and Skolnick	[79]	pK_i/pK_d	logMW	0.38	CDSa(CDS1-7)	146
					0.24	CDS3-HIV-1 protease	28
					0.53	CDS5-Ribonuclease a	13
					0.76	CDS6-Thermolysin	10
					0.50	CDS7-Beta trypsin	47
					0.59	Protein Ligand Database v1.3 CDS8-Beta trypsin	7
					0.88	CDS9-Carbonic anhydrase II	15
					0.40	CDS11-HIV-1 protease	6
					0.49	CDS12-Thrombolysin	9
2011	Reynolds and Holloway	[83]	ΔG	NHA	0.28^b	BindingDB	102
2012	Present study		ΔG	logMW	0.14	Non-drugs	320
				logW	0.15		
				logP	0.19		

Abbreviations. logP: logarithm of octanol/water partition coefficient; MW: molecular weight; N: number of data; NHA: number of heavy atoms; W: Wiener index.
[a]Spearman's R^2; [b]Calculated using the data in the reference.

$$\Delta G \approx \Delta H_{inter} + \Delta H_{intra} - T\Delta S_{config} + \Delta G_{sol} \quad \text{Eq. 4}$$

$$\Delta H_{inter} \approx \Delta E_{Coulomb} + \Delta E_{LJ} + \ldots \quad \text{Eq. 5}$$

ΔG can be approximated (Brooijmans and Kuntz) [84] further by separating the terms of Eq. 3 into enthalpy changes (Eq. 4) coming from changes of intra (ΔH_{intra})- and intermolecular (ΔH_{inter}) interactions, configurational entropy change (ΔS_{conf}), and a free energy change coupled to (de)solvation processes (ΔG_{sol}), such as release of interface waters (Eq. 1) during binding. ΔG_s includes both enthalpic and entropic contributions of changes of solute-solvent interactions during complex formation. (Notably, there is an unclosed debate in the literature on the separability of the entropic terms for individual (molecular) contributions which may affect the above separation of ΔG_s from other terms of ΔG [85, 86]. In many ΔG calculators [84], ΔH_{inter} is estimated involving pair-additive potential terms such as the Coulomb ($E_{Coulomb}$) or the Lennard-Jones (E_{LJ}) formulas (Eq. 5) for electrostatic and van der Waals-interactions, respectively. Jacobson and Karlén [87] found that ΔG calculators built mostly on such enthalpic terms of ligand-target interactions (Eq. 5) produced high correlations with NHA hinting that ΔH_{inter} accounting for protein-ligand interactions is partly described by NHA. One possible explanation is that NHA can be related to surface area, and hence, to van der Waals interactions and, therefore, a high NHA can translate into a high ΔH_{inter}.

Besides ΔH_{inter}, some parts of the configurational entropy (S_{config}) can be also related to MW (Eq. 6)

$$S_{config} = S_{trans} + S_{rot} + S_{vib} \quad \text{Eq. 6}$$

$$S_{trans} + S_{rot} = R\ln(aMW) \quad \text{Eq. 7}$$

where trans, rot, and vib denote respectively, the translational, rotational, and vibrational ΔS contributions to the configurational entropy change, and 'a' is a constant. Several studies [88-94] calculate S_{config} using classical formulas relating S_{trans} and S_{rot} to the logarithms of MW and the principal moments of inertia, respectively. As known, the principal moments of inertia are also dependent on molecular size (and shape). Simplified formulas [95, 96] were also introduced (Eq. 7) showing the dependence of part of S_{config} on MW. However, this dependence was suggested to be very weak or zero for the change of S_{config}, i.e. for ΔS_{config} of the binding process [97, 98].

In summary, several studies have published relationships (Table **3**) at various correlation levels between experimental binding affinity and molecular property filters such as MW, NHA, etc. Since the article of Gilson *et al.* [97], which had also dealt with the ΔG-MW correlation, experimental collections have been published presenting new data. A collection of recent correlations was provided in Table **3** and the thermodynamic background was sketched to illustrate the problems of explaining these correlations. While the above considerations suggest that individual components of ΔG such as ΔH_{inter} are related to molecular size, and some of them, such as ΔS_{conf}, are probably not correlated with MW, the final explanation on the moderate, but significant correlations of ΔG with ligand size is still awaiting. Notably, these relationships are probably not linear as quantities obtained by simple normalization of ΔG with, e.g. MW, are still dependent on MW (see next Section for details).

4. THE CONCEPT OF LIGAND EFFICIENCY (EFFICIENCY INDEX, EI)

The dependence of binding affinity on ligand size (MW, NHA) discussed in the previous section raises the question whether it is possible to define a measure, the binding efficiency for comparison of 'intrinsic' binding affinities of ligands of any sizes *via* 'decoupling' ΔG from molecular size. In an early work, Andrews *et al.* [99] hinted at the possibility of definition of such intrinsic ΔGs for a limited number of functional groups of a molecule by using average values calculated from experimental ΔGs. Later, DeWitte and Shaknovich [100] calculated the intrinsic binding affinity per heavy atom and correlated these values with experimental K_i-s. Kuntz *et al.* [72] also used this intrinsic measure and showed that ΔG/NHA rapidly decreases up to ca. 15 NHA (see also previous Section).

Based on the above results, Hopkins *et al.* [101] recommended the introduction of ligand efficiency in the following explicit form (Eq. 8). The work of Wells and McClendon [78] provides information on the actual values of 'efficient' molecules. They collected several potent small molecules inhibiting protein–protein interactions and obtained $|EI_{NHA}|$ values of 0.2…0.4 for their data set. An alternative, idealized value of 0.5 has been recommended by others [63, 101, 102].

$$EI_{NHA} = \frac{\Delta G}{NHA} \quad \text{Eq. 8}$$

To note, throughout this review we use the name 'efficiency index (EI)' instead of 'ligand efficiency' to emphasize that this measure of intrinsic ΔG is a rational definition of the efficiency of a ligand, however, it is not the only possible definition.

Definition of other EIs was provided by Abad-Zapatero and Metz [24] using MW (EI_{MW}) and PSA (EI_{PSA}) in the denominator of Eq. 8 instead of NHA. A series of other EIs were introduced based on various size-dependent properties for normalization among which the Wiener-index (W) was found particularly useful in the form of EI_W [103]. Leeson and Springthorpe [45] proposed a ligand-lipophilicity-based efficiency index (EI_{lipo}, Eq. 9) to be used in "maximizing the minimally acceptable lipophilicity" per unit of binding affinity during drug design. They suggest that an average drug has an EI_{lipo} of 5-7 or greater.

$$EI_{lipo} = pIC_{50} \text{ (or } pK_i) - \text{clogP (or logD)} \quad \text{Eq. 9}$$

Although the definition of EIs involves normalization by ligand size (Eq. 8), Reynolds *et al.* [74] found that EI_{NHA} is still dependent on ligand size, as a very dramatic decline was observed in EI_{NHA} as size increases. Notably, Orita *et al.* [104], and Keserü and Makara [105] described a similar trend of EI_{NHA} vs. NHA. The drop in EI_{NHA} was large between ca. NHA=10…20, and flattened toward very large sizes (NHA>40). They found an interesting similarity between the maximal EI_{NHA} vs. NHA and the ASA vs. NHA curves suggesting that the primary driving forces behind the systematic decline in maximal EI_{NHA} with increasing molecular size is the reduced effective surface area for the larger compounds. In other words, large molecules possess relatively large buried surface area unavailable for binding. In a recent study, Reynolds and Holloway [83] concluded that the strong size dependence of EI_{NHA} (average or optimal) is mostly a consequence of the dependence of the enthalpic, and not the entropic part of EI. To eliminate the above size-dependency of EI_{NHA}, Reynolds *et al.* [74] introduced a new functional form called 'fit quality', and Nissink [106] derived a size-independent ligand efficiency measure of the form of binding affinity/$NHA^{0.3}$.

The concept of ligand efficiency is a simple way to merge binding and pharmacokinetic characteristics of a ligand into a single measure. EI has already been applied in many studies and it is suggested to become a useful tool of fragment-based drug discovery [102, 104, 107-109], lead optimization [46], and drug chemical (molecular) space localization for some diseases or organs [110]

5. SENSITIVITY AND SELECTIVITY OF PROPERTY FILTERS

Tables **1** and **2** list general or specific drug-likeness values of molecular properties. Most of these values are descriptive statistics (mean, median, percentile, etc.) of data sets including only drugs. That is, counter-examples of a set of non-drugs are generally not considered. Notably, the strict definition of such sets is not obvious (Section 2.2) due to possible change/evolution of the drug/non-drug status of any compounds. However, if sets of drugs were collected, then it is fairly plausible to expect a non-drug set for comparison. Introduction of a new statistical term on the selectivity of the property filter is also necessary showing the ability of the property to distinguish drugs from non-drugs. Since the investigated molecular properties (MW, NHA) are coupled to both pharmacokinetic drug-likeness (Section 2) and ΔG (Section 3), it would be also advantageous to 'switch off' the ΔG-coupling in an analysis to investigate the properties' selectivity only for drug-likeness. In the forthcoming Sections, selectivity and sensitivity measures of drug-likeness filters are introduced using a 631-compound database as an example.

5.1. Data Sets

Details of the collection of the data sets are provided in the Appendix and the sets are listed in the Supplementary Material. To decouple the ΔG-dependence (Section 3) of the property filters, the two sets (320 non-drugs and 311 drugs) were designed to have the same range of maximal experimental ΔG. To evaluate data sets and assess the similarity/dissimilarity of the distributions, a standard protocol of statistical analysis was followed (Appendix). The distribution of the data was checked, and it was found that the ΔG values in the sets and also in the entire database followed non-normal distributions ($p<0.001$). To check the distribution of an even larger sample of available experimental ΔG data, the same tests were performed for a set of more than 4,000 binding affinity values from the BindingDB [111] database and it showed a non-normal distribution as well. As the normality tests failed for the ΔG data sets, two non-parametric tests were applied and showed equal medians and distributions of ΔG between the drug and non-drug populations ($p>0.1$, $p>0.05$). In addition to the statistical tests, a high degree of overlap between the distributions of the two ΔG populations can be seen from the plot of their histograms (Fig. **1a**), and from the fitted mixed normal probability density functions (PDF, Fig. **1b**). The comparison of descriptive statistics also emphasizes the equality of drug and non-drug ΔG populations. The medians of the samples are in good agreement ($\Delta \approx 0.5$ kcal/mol) and the median difference between percentiles of the two samples (Appendix) is a marginal 3 % (Fig. **1c**). Details of the statistics are included as Supplementary Material.

In conclusion, a database of drug and non-drug compounds was collected wherein the two sets have ΔG distributions of significantly high similarity. Importantly, such criterion was not applied for the distribution of molecular properties and EIs of the two sets. Thus, it could be tested if the properties can describe general drug-likeness 'decoupling' effects common with ΔG. The outcome of this test is summarized in the next Section.

5.2. General Drug-Likeness Filters

Similarly to the previous section, the results of normality tests indicate that most of the investigated drug-likeness property filters (MW, NHA, W, logP) and the corresponding EIs (Section 4 and Appendix, EI_{MW}, EI_{NHA}, EI_W) are not normally distributed ($p<0.001$). In contrast with the previous section, the non-parametric tests of equivalence resulted in a highly significant difference ($p<0.001$) between the property/EI distribution of the drug and non-drug sets. There is a considerable increase in the medians of MW, NHA, and W with Δ≈150, 10, and 2000 units respectively, for non-drugs compared with drugs. Similarly, the corresponding median percentile differences are in the range of 15-230%, which is significantly larger than that of ΔG (Fig. **1c**). The histograms and Probability Density Functions (PDF's) (Fig. **2a**, **b**, **e**, & Supplementary Material) show a change in the shape of the distributions. Whereas the ΔG distributions (Fig. **1a**, **b**) are rather rounded, well-defined peaks appear in the case of MW, NHA and logW, reflected also by a change in the kurtosis value from negative to positive. For drugs, a sharp peak and a high kurtosis value appear, while the non-drug histogram is flat with a long tail.

A similar separation of the two sets can be observed using EIs (Fig. **2c**, **d** & Supplementary Material). The EI_{MW} histograms (Fig. **2c**) do not resemble the non-separable ΔG distributions of drugs and non-drugs (Fig. **1a**). The partial separation of EI values seems to be a plausible consequence of the differentiating power of the parent MW. By definition, in EIs the populations of ΔG and MW or NHA are connected and the distributions of EI_{MW} and EI_{NHA} reflect the shape of the one-peaked MW (Fig. **2a**) or NHA distributions, which are more suitable candidates for statistical evaluations than the flat ΔG distributions with dual maxima (Fig. **1a**, **b**).

Whereas significant separation power of the filters can be concluded from the above analysis, a considerable overlap of the drug and non-drug histograms can also be observed especially in the cases of W and EIw where the distributions have an exponential shape (Supplementary Material). Notably, taking the logarithm of W (logW) resulted in separate peaks (Fig. **2e**). For logP, (Fig. **2f**) the drug population is centered in a well-defined peak in the hydrophobic region (logP>0), as can be expected for drugs [12, 42], whereas the distribution of non-drugs is similar to the case of ΔG. The considerable overlap of drug and non-drug populations in the hydrophobic region, along with a separate non-drug sub-population

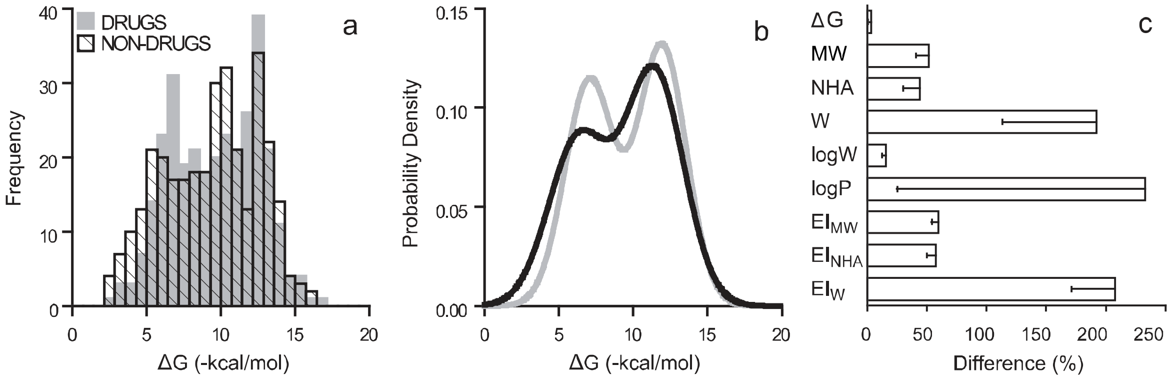


Fig. (1). Comparison of binding affinity distributions of sets of drugs and non-drugs. The two compound sets with N=311 and 320 members respectively, were designed to be non-separable by ΔG. Overlapping histograms of ΔG values in part (**a**) and two-component normal mixture probability density functions fitted to the histograms in part (**b**) reflect the similarity of the two datasets. In part (**c**), the median differences between the series of percentiles of the two sets are shown. Whereas the difference is marginal in the case of ΔG, it is significant for the filters. Error bars represent median absolute deviations.

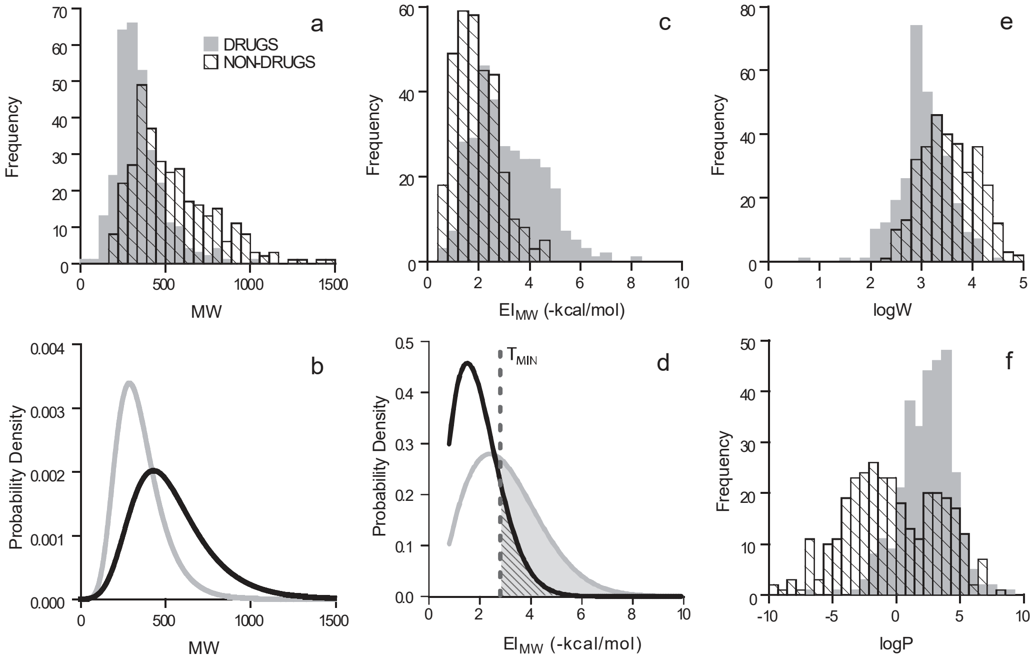


Fig. (2). Separation of sets of drugs and non-drugs by drug-likeness filters. Histograms in parts (**a**), (**c**), (**e**), (**f**), and fitted probability density functions in parts (**b**) and (**d**) reflect separation of the two compound sets by various filters. Part (**d**) also features key terms of this study with an example of EI$_{MW}$. The shaded area below the drugs group curve represents the sensitivity (σ) of EI$_{MW}$ above the threshold T$_{MIN}$=2.8 kcal/mol. The ratio of this shaded area and the striped area below the non-drug curve shows that drugs can be found with three fold higher probability than non-drugs above T$_{MIN}$ and by definition this equals the drug-likeness ratio (DR=3).

in the hydrophilic region (logP<0) explains the high spread of median differences at logP (Fig. **1c**).

5.3. Definition of Selectivity and Sensitivity of Drug-Likeness Filters

As it was shown in the previous section, sharp borders cannot be drawn between the partly overlapping drug and non-drug populations for the properties investigated. To achieve a coherent formulation of selectivity and sensitivity, fits of Probability Density Functions (PDF) of continuous distributions (Weibull, Gumbel, Exponential) were performed for histograms of the properties (Figs. **2b**, **d**). Using these explicit forms of PDFs, an analytical comparison of the distributions of drugs and non-drugs has become possible for the EI's and MW (Fig. **3** and Appendix). In the following discussion we will use the example of EI$_{MW}$ for the introduction of PDF-based sensitivity and selectivity of the filters.

The probability that a drug adopts an EI$_{MW}$ larger than a minimum threshold (T$_{MIN}$) is expressed as a percentage (Eqs. A4 and A6) and named sensitivity (σ) as it reveals whether a large enough section of the entire drug population is included in the region under question. A σ=51% is represented by an shaded area in Fig. (**2d**). In this case, 51% of the total drug population is located in the region above T$_{MIN}$. The larger the sensitivity of a filter, the fewer drugs are excluded erroneously above a minimum threshold T$_{MIN}$. Detailed definitions of probabilities are shown in the Appendix. Decidedly, σ is a necessary, but not a sufficient parameter of a property filter.

Further inspection of the fitted PDFs of EI$_{MW}$ (Fig. **2d**) reveals that in the region starting from T$_{MIN}$, the probability that a drug adopts an EI$_{MW}$ is three times higher than this probability for non-drugs. Thus, the ratio of the shaded area below the PDF curve of drugs and the striped area (Fig. **2d**) corresponding to non-drugs is three. Generalizing the previous observations, we introduce another measure of selectivity (Eqs. A5 and A7), the Drug-likeness Ratio (DR), relating the population of drugs with that of non-drugs by the ratio of their probabilities. In terms of the above-mentioned example, DR equals 3 as there is a three-fold higher chance for a compound to be a drug than a non-drug above T$_{MIN}$.

After fitting the PDFs, thresholds can be fine-tuned for a drug-likeness filter using the DR and σ functions as calibration curves (Fig. **4a**), i.e. the T$_{MIN}$ value can be read from the curve plot at a required level of DR or σ. According to the relative position of DR and σ functions, drug-likeness filters can be categorized into three types (see also Appendix for details): those with limits of T$_{MIN}$ (Fig. **4a**), both T$_{MIN}$ and a maximum threshold (T$_{MAX}$, Fig. **4b**), or only T$_{MAX}$ (Fig. **4c**). EI$_{MW}$ can be categorized under the first type (Fig. **4a**). In the above-mentioned example (Fig. **2d**), a T$_{MIN}$ of 2.8 kcal/mol is a realistic lower EI$_{MW}$ threshold at levels of DR=3, and σ=51%. As σ decreases with increasing DR (Fig. **4a**), thresholds with DR>10 may have no practical importance.

In Table **4**, general thresholds calculated for all filters at DRs from 2 to 3, and σ>50% are listed. Compared with values from the literature (Table **1**), it can be concluded that the calibrated range of 129-369 for MW (Fig. **4b**) correspond to drugs. Calibrated thresholds of NHA, EI$_{MW}$ and EI$_{NHA}$ at similar DR and σ values are also located at the drug/lead border (Table **1**). For logP there are various data published and our estimated range of 0.7-4.3 between T$_{MIN}$ and T$_{MAX}$ agrees well with the values from the literature (Table **1**). The above results allow experimenting with calibration, and fine-tuning of thresholds at different DR and σ levels depending on the nature of desired applications, i.e. if hits, leads or drugs are investigated requiring small/large selectivity and sensitivity criteria, etc. Example thresholds at various DR and σ levels and details of the calculations can be found in the Supplementary Material. Importantly, while the above description

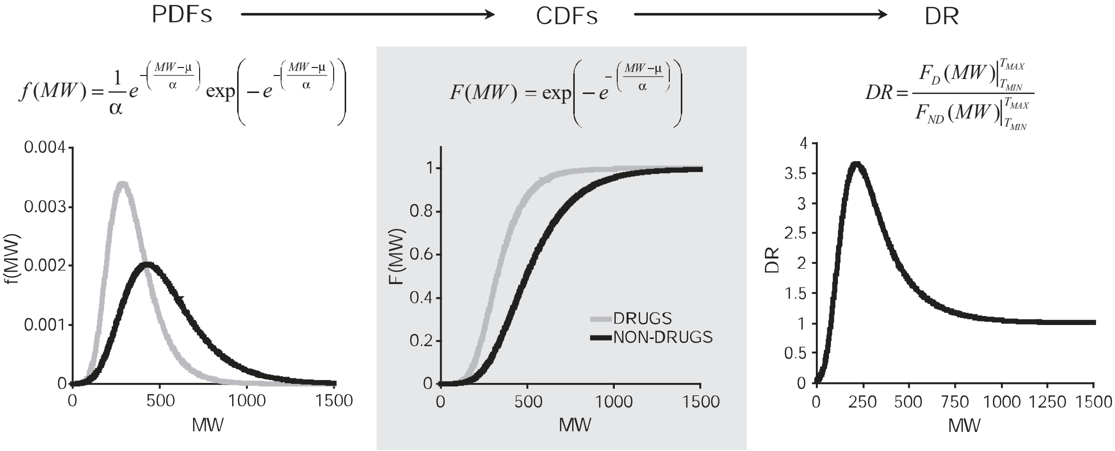


Fig. (3). An example of the use of fitted probability density function (PDF, f) and the corresponding cumulated density function (CDF, F) for the analytical calculation of selectivity and sensitivity measures DR and σ of MW. As F(MW)≈0 for small MWs, T_{MIN} was omitted from function DR. Gumbel distributions were fitted for both drugs (D) and non-drugs (ND) sets (see also Fig. **1**). Notably, the general functional formulae are provided in this figure and different scale (α) and location (∝) parameters were obtained for the two sets (see Supplementary Material for numerical values of the parameters and details of fit). The σ can be directly calculated (Eq. A6) from the CDFs according to $\sigma=100[F_D(T_{MAX})-F_D(T_{MIN})]$. Since the DR function has a maximum on the MW≤1500 domain investigated, T_{MIN} and T_{MAX} thresholds can be calculated (Fig. **4**) for DR values up to ca. DR=3.75. Plausibly, a DR≥ 1 is of interest.

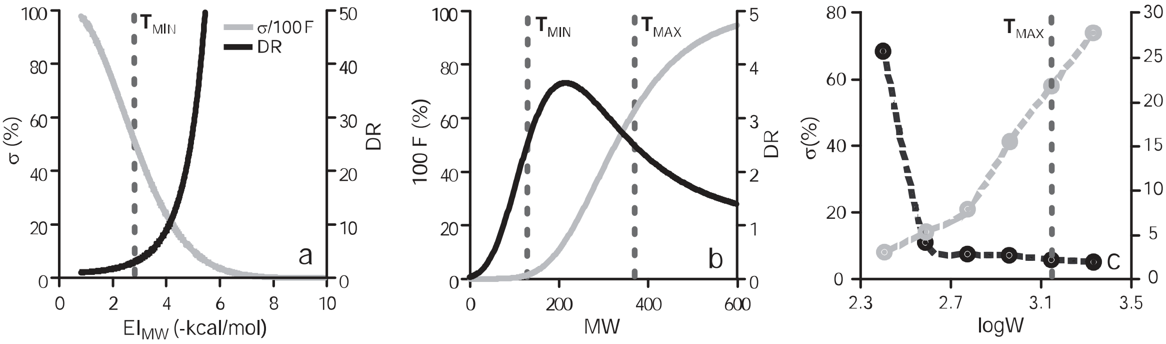


Fig. (4). Calibration curves of drug-likeness thresholds. The shape and relative location of sensitivity (σ) and Drug-likeness Ratio (DR) functions facilitate calibration of the three types of drug-likeness thresholds (T) of filters. (**a**) In the case of EI_{MW}, the DR function increases and σ decreases on the domain investigated. Thus, a minimum threshold (T_{MIN}=2.8 kcal/mol) can be calibrated (following the previous example of Fig. **2d**) with a DR=3, which is large enough that EI_{MW} can separate drugs from non-drugs. At the same time, the sensitivity of EI_{MW} is also acceptable (σ=51 %) for recognition of drugs above this threshold. (**b**) In case of MW, the DR function has a maximum, and therefore there are two thresholds (T_{MIN} and T_{MAX}) with the same DR value specifying a favorable MW interval with sufficiently high DR values. Here, $\sigma=100[F_D(T_{MAX})-F_D(T_{MIN})]$, where F_D is the cumulative distribution function of drugs. At higher DR values, i.e. narrower (T_{MIN}, T_{MAX}) intervals σ becomes smaller. (**c**) For a decreasing DR, the maximum of logW can be set (T_{MAX}), below which the separation of drugs from non-drugs is possible by logW. To note, the logW-related curves are not continuous functions, the points are derived from raw histogram data.

of filters with DR and *σ* were used for drug/non-drug (drug-likeness) separations, our present approach can easily be easily adopted to describe the filters in drug/lead or lead/hit relations (lead-likeness).

5.4. Disease-Specific Drug-Likeness

It is informative to characterize the discriminating power of the filters between drugs and non-drugs beyond general terms according to disease categories (Section 2.4). For this characterization, the set of drugs was divided into sub-sets by disease categories according to the classification of DrugBank [112]. Similarly to the case of general drug-likeness (Section 5.2), a non-drug companion with the closest ΔG was selected for each drug in each disease category. This method resulted in selected disease category sub-sets of non-drugs that are inseparable from the corresponding drugs by ΔG (Fig. **5**). In all cases, statistical comparisons of sub-sets of drugs and non-drugs were performed for ΔG and for all 8 filters. The overall results on separation of inter-quartile ranges are shown as a matrix (Fig. **5**), other details can be found in the Supplementary Material. (Notably, due to the relatively small number of drug/non-drug members of the sub-sets *σ* and DR were not calculated in this analysis by disease types. In forthcoming studies we plan to extend the selectivity and sensitivity calculation of the filters on large disease-specific data sets.)

In 70% of the cases, separation of the sub-sets at different levels can be observed (Fig. **5**), and in the remaining cases, the inter-quartile ranges of drugs and non-drugs are completely overlapping.

Table 4. Calibrated Thresholds, Selectivity, and Sensitivity of Property Filters

Property filter	General thresholds				Disease-specific thresholds
	T_{MIN}	T_{MAX}	DR	σ	T_{MIN}-T_{MAX}[c]
MW	129	369	2.5	61	206-322[d], 262-342[e], 258-342[f]
NHA[a]	9	27	2.0	67	
W[a]	-	2037	2.0	73	578-1180[f]
logW[a]	-	3.1	2.1	58	2.76-3.07[f]
logP[a]	0.7	4.3	2.6	67	1.02-3.27[g], 1.74-3.59[d]
EI_{MW}[b]	2.8	-	3.0	51	
EI_{NHA}[b]	4.2	-	3.0	52	
EI_{W}[b]	7.5	-	3.0	56	6.51-15.87[e], 6.01-17.78[h], 7.35-21.93[f]

Abbreviations. DR: drug-likeness ratio (selectivity); EI: efficiency index; logP: logarithm of octanol/water partition coefficient (small letters in brackets denote different logP definitions); MW: molecular weight; NHA: number of heavy atoms; σ: sensitivity; T: threshold; W: Wiener-index.

[a] Calibrated values of this filter were estimated from histograms and not from fitted distributions. [b] The dimension of EIs is –kcal/mol. [c] T_{MIN}-T_{MAX} is a (modified) inter-quartile range of the drugs set (no DR and σ values are given). [d] Musculo-skeletal system. [e] Nervous system. [f] Various. [g] Dermatologicals. [h] Respiratory system.

There is a minority (10%) of cases in which a high separation (>90%) was found. The "worst performance" occurred in the categories of Antineoplastic agents and Anti-infectives, with cancer drugs presented in the former category. This finding implies that the failure of new drug discovery in these areas [113, 114] may be partly due to the inefficacy of drug-likeness filters investigated in this study. There are also numerous cases where only a partial separation was achieved as, e.g. the cardiovascular system compounds at MW (Fig. **5b**).

Based on these observations, drug-likeness thresholds (Table **4**) were estimated for disease groups using the inter-quartile ranges of properties that provide the highest level of separation (>90%), which show agreement with Table **2**. These disease-specific thresholds provide in some cases (MW, logP) narrower drug-likeness ranges than the general thresholds found within reference values available in the literature.

Whereas the evaluation of the above mentioned negative or partly successful cases is not an easy task, certain positive results can be readily explained. For example, logP, well-known to describe skin permeability [115-117], performed well for the category of dermatological drugs which require absorption through the skin (Fig. **5c**). Similarly, MW, a good filter of nervous system drugs in this study (Fig. **5d**), describes blood-brain barrier penetration [66, 67], an important issue of drug design for CNS diseases. The MW-threshold calculated for nervous system diseases (Table **4**) is in good agreement with the MW<400 value recommended by other studies [34, 69].

Interestingly, the performance of EIs does not always correspond to their parent ligand-based properties (MW, NHA, W), emphasizing their different information contents. Besides the well-known drug-likeness filters such as MW and logP, the recently introduced EI_W [103] was one of the best separators according to the present analysis, emphasizing the benefits of using EIs.

6. SUMMARY AND FUTURE OUTLOOK

Molecular properties of drug candidates have been extensively used as drug-likeness filters of compound libraries. In the present

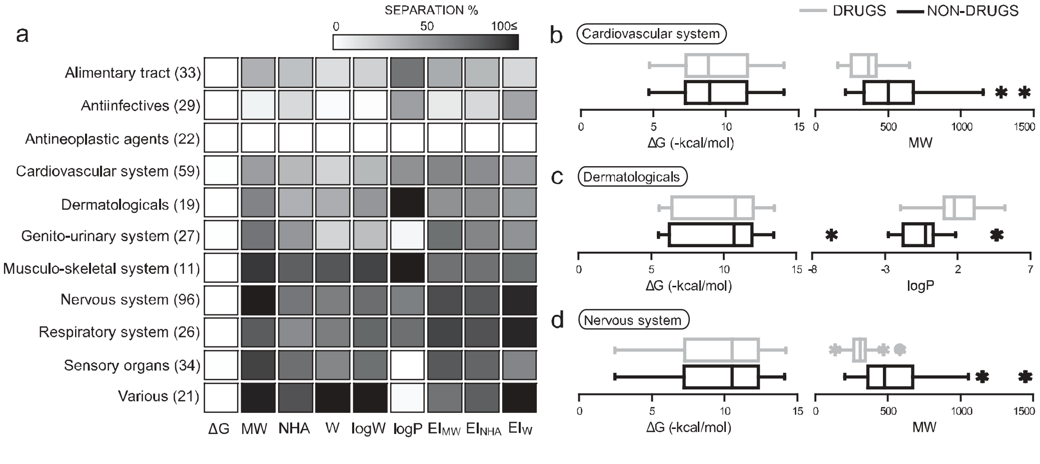


Fig. (5). Disease-specificity of drug-likeness filters. The set of drugs (N=311) used in this study was split into sub-sets according to various disease categories. The number (N) of members of these sub-sets is marked in brackets on the left side of the shaded matrix (**a**). Sub-sets of non-drug compounds possessing the same ΔG distribution as the sub-sets of drugs were formed. Each cell of the shaded matrix shows the level of separation of the inter-quartile range of a sub-set of drugs from that of non-drugs according to ΔG or a filter. The separation is 0 % if the two ranges are completely overlapping. This is the situation for all disease categories in the ΔG column due to the aforementioned selection of sub-sets of non-drugs. The separation is between 0 and 100 % if there is a partial overlap between the ranges as shown in the box plot for cardiovascular system drugs according to MW (**b**). If there is no overlap in the ranges, then the separation is 100 %. The latter situation is featured in examples of box plots of dermatologicals according to logP (**c**), and nervous system drugs according to MW (**d**).

review, a distinction was made between general and specific drug-likeness. While the investigated properties significantly differentiated between the sets of oral drugs and non-drugs in general, a considerable overlap remained between the two sets. Certain disease types or drug administration routes may require specific filter values instead of the broader, general ones. It was also discussed to which extent the molecular properties are coupled to ΔG. Statistical comparison of drug sets with non-drugs of similar binding affinity and use of selectivity and sensitivity measures were introduced as an improved description of the overlapping distributions of filter values. With the new measures filtering thresholds gain statistical meaning: namely, their selectivity against non-drugs (DR) and sensitivity for drugs (σ). In addition to the positive results of the general drug-likeness concept, relevant criticism and limits of its applicability were also surveyed.

Filtering thresholds can help in the future design of standardized, compound libraries assembled for binding assays, HTS, or other *in vivo* tests. However, precise statistical calibration of filtering thresholds − as shown in this work − may be required beyond simple descriptive statistics (mean, median) to assess full reliability of the thresholds. Disease-, target-, or administration-specific drug-likeness filters may help the design of focused libraries which may become a competitive alternative to general compound sets. Molecular property and EI-based filters have an increasing impact also in fragment-based design [101, 118], and in the chemical optimization of physico-chemical properties of natural products [113, 119] or other lead compounds.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGEMENT

This study was financed by the European Social Fund (grant agreement no. TAMOP-4.2.1/B-09/1/KMR-2010-0003), an Estonian Science Foundation grant JD80, an Estonian Ministry for Education and Research grant SF0140031Bs09, and an Innove Foundation grant 1.0101-0310. C.H.'s work was financed by a Bolyai Scholarship of the Hungarian Academy of Sciences. We are grateful to Professor David van der Spoel for critically reading of the manuscript and Professor David Meredith (Department of Mathematics, San Francisco State University) for providing the program Xplore.

APPENDIX

Collection and Verification of Compound Sets of Drugs and Non-Drugs

The structure of 311 drugs and 320 non-drugs and their experimental binding affinities (mostly as inhibition equilibrium constants, K_i) were collected from the following sources: PDBbind v2005 [120], KiBank [121, 122], SCORPIO [123], and from a previous study [124]. The BindingDB [110] database was also used for normality test comparisons. Where it was necessary, ΔG (precisely the standard Gibbs free energy change, ΔG^o − the standard sign is omitted in this study for simplicity) values were obtained from K_i by $\Delta G = RT\ln K_i$, using T=25 °C (298.15 K). The complete sets of drugs and non-drugs, as well as their raw and converted ΔG values are available as Appendices of the Supplementary Material. Similarly to other studies [72, 81], maximal ΔG values, i.e. ligand binding affinities corresponding to the complex with the relevant, strongest binding protein partner were collected. In the case of drugs, ΔG values with the pharmacologically relevant targets were considered. Whereas ΔG correspond to a multi-molecular interaction between the ligand compound, target, and solvent shell, it has been shown that ΔG is also related to molecular properties (MW, NHA, logP) [72, 78, 81] of the ligand only, as these properties hold information on both enthalpic (ΔH) and entropic (ΔS) constituents [103] of ΔG (through ΔG = ΔH - TΔS). Consequently, there is a ligand-based part of ΔG explained by the above properties (see also Section 3), which is constant regardless of the actual target. Since a compound can bind as a ligand to various targets, it can adopt different ΔG values due to target-specific interactions (ΔH) and, therefore, the maximal experimental ΔG, i.e. the maximum ΔG value of a compound with its relevant target(s), were collected for both sets in the present study. Using these maximum ΔG values helps decreasing target-specificity of the interaction (ΔH) part as they correspond to the ideal binding affinity of a compound.

The selection procedures of the two sets are following. (i) The list of all small-molecule approved drugs was downloaded from the DrugBank database, which also contain disease-specific data on drugs approved by the FDA (U.S. Food and Drug Administration agency). A standard, programmed procedure was applied to ensure purity the two sets. (ii) The ligand names were extracted from the PDB files, and queried in the DrugBank [112] database to identify those ligands that are FDA-approved drugs, and to avoid contamination of drug molecules in the non-drug collection. (iii) The set of non-drugs was designed to have overlapping binding affinity distribution with the set of drugs (Figs **1a**, **b**). While non-drugs were selected with a similar ΔG as drugs, but such criterion was not applied for the filtering properties of the compounds. Thus, there were no circumstances in the sampling which affected the composition of non-drugs set so as to determine/guarantee its similar/different property (MW, NHA, etc.) distribution compared with the drugs set (Fig. **1c**).

Filters

There are various properties applied in drug design as size, structural, or property filters. Whereas size filters such as molecular weight (MW) or number of heavy atoms (NHA) require solely the knowledge of a compound's atomic composition, structural descriptors also involve intra-molecular connectivity. The Wiener index (W) is a typical structural descriptor reflecting the branching and complexity of the molecule. The W is a robust measure as it does not depend on the molecular conformation. To be able to calculate the W of a compound, knowledge of its Lewis-structure is sufficient (Eq. A1). In this study, its logarithm (logW) is also used.

$$W = \frac{1}{2}\sum_{i,j}^{NHA} d_{ij} \qquad \text{Eq. A1}$$

where d_{ij} is the number of bonds in the shortest path connecting the pair of atoms i and j in the molecule. There are also other property filters, e.g. the logarithm of octanol/water partition coefficient (logP) which is generally applied as a measure of hydrophobicity for a non-ionized compound. The binding affinity and the aforementioned size or structural properties have been combined into hybrid filters called the efficiency indices (EI). The EIs are ΔGs normalized by these filters (Eq. A2). Exponents of ten were used as multipliers in the formulae to obtain human-readable EI values.

$$\begin{aligned} EI_{MW} &= 100\frac{\Delta G}{MW} \\ EI_{NHA} &= 10\frac{\Delta G}{NHA} \\ EI_{W} &= 1000\frac{\Delta G}{W} \end{aligned} \qquad \text{Eq. A2}$$

The program XLOGP v2.060 [125] was used to calculate the logarithm of octanol/water partition coefficient (logP) by an atom-additive method including correction factors. The calculations of molecular formula and number of heavy atoms (NHA), molecular weight (MW), and Wiener index were performed with Marvin Beans v4.1.861 [126]. The experimental ΔG's, calculated physicochemical properties and EI's are available as an Appendix in the Supplementary Material.

Descriptive Statistics

To check the similarity/dissimilarity of the distributions, a standard protocol of statistical analysis was followed for all ΔG and filter sets. A complete descriptive statistics including histogram (Figs. **2a**, **c**, **e**, **f**), minimum, maximum, range, median, median absolute deviation, arithmetic mean, standard error of arithmetic mean, 95.0% confidence interval, trimmed mean (10%, two sided), standard deviation, variance, coefficient of variation, skewness, kurtosis and a data vector of percentiles (1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 99%) was calculated for all data sets with program package Systat 12 [127]. The median of differences (%, Fig. **1c**) between vectors ($\vec{p}$) of tabulated percentiles of two sets (drugs and non-drugs) was calculated according to Eq. A3.

$$\text{Difference}(\%) = \text{median}\,(\vec{p}_{\text{DIFF}})\ ;\ \{p_{\text{DIFF}}\}_i = \frac{100\,|\,\{p_{\text{DRUGS}}\}_i - \{p_{\text{NON-DRUGS}}\}_i\,|}{\min(|\,\{p_{\text{DRUGS}}\}_i\,|;|\,\{p_{\text{NON-DRUGS}}\}_i\,|)} \qquad \text{Eq. A3}$$

Where $\vec{p}_{\text{DIFF}}$ is the difference vector and $\{p_{\dots}\}_i$ denotes the element of a vector. The spread of $\vec{p}_{\text{DIFF}}$ was given as median absolute deviation. Results of descriptive statistics are tabulated in the Supplementary Material**.**

Statistical Tests

The Shapiro-Wilk [128], Kolmogorov-Smirnov [129], and Anderson-Darling [130] tests were applied to check if the data sets came from a normally distributed population (α=0.05). The null hypothesis was that the population is normally distributed. If the p-value was smaller than significance level α, then the null hypothesis was rejected (the data are not from a normally distributed population). If the p-value was larger than α, then the null hypothesis that the data came from a normally distributed population was accepted. The statistics and p-values are tabulated in the Supplementary Material. As the data populations are not normally distributed, non-parametric tests are valuable, since they do not require assumptions on the distribution of the population and therefore are sometimes called distribution-free [131]. Thus, in the present study the non-parametric two-sided Kruskal-Wallis test (also called Wilcoxon rank sum test or Mann-Whitney [132] U test, α=0.1) and the two-sided Kolmogorov-Smirnov two sample test (α=0.05) were used to decide if two data sets came from the same population. The null hypothesis was that the two samples came from the same population and have the same distribution. If the p-value was less than the α level, then the null hypothesis was rejected (the data are not from the same distribution). If the p-value was greater than α, then the null hypothesis that the data came from the same population was accepted. The statistics and p-values are tabulated in the Supplementary Material. All tests were performed with Systat 12, many cases were counterchecked and p-values were calculated in parallel with the program R [133].

Fitting Distributions

In all cases where the data allowed, PDF's of the following 21 distributions were fitted to histograms of each data sets (and their parameters estimated by the respective methods) using Systat 12. Beta, Chi-square, Erlang, Gamma, Gumbel, Logistic, Loglogistic, Smallest extreme value (method of moments); Normal, Lognormal, Logit normal, Exponential, Double exponential (Laplace), Gompertz, Inverse Gaussian (Wald), Pareto, Rayleigh, Weibull, Uniform, (maximum likelihood method); Cauchy (method of quantiles or order statistics); Triangular (modified maximum likelihood and moments).

In all cases, 12 bin histograms were prepared for the fits. In the case of mixed normal distribution (Fig. **1b**), and for refinement of some fits (especially for calculation of location parameters of Weibull distributions), the software Dataplot [134] along with the probability plot correlation coefficient plot (PPCC) method was used. Quality of fits was confirmed by Kolmogorov-Smirnov and Anderson-Darling tests with Systat 12. Only highly significant PDF's (α=0.1) were selected for further use. The analytical form of PDFs (Fig. **2b**, **d**) facilitated the mathematically accurate calculation of calibration of thresholds (Fig. **4a**, **b**). Statistics of tests of fit, formulae of selected distributions and values of their location, shape and scale parameters of the PDF are listed in the Supplementary Material.

Calibration of Thresholds

The probability (P_D) that a filter χ adopts a value between thresholds T_{MIN} and T_{MAX} for drugs is expressed as a percentage and named sensitivity (σ) in this study (Eq. A4), as it reveals whether a large enough section of the entire drug population is included in the region under question. The random variable ξ_D^χ corresponds to the statistical event when a filter χ adopts a value for drugs (D).

$$\sigma = 100 P_D (T_{MIN} \le \xi_D^\chi \le T_{MAX}) \qquad \text{Eq. A4}$$

The drug-likeness ratio (DR) is expressed (Eq. A5) as the ratio of P_D and the corresponding probability for non-drugs (P_{ND}).

$$DR = \frac{P_D (T_{MIN} \le \xi_D^\chi \le T_{MAX})}{P_{ND} (T_{MIN} \le \xi_{ND}^\chi \le T_{MAX})} \qquad \text{Eq. A5}$$

In the cases where fitted continuous PDFs are available for drugs (f_D) and non-drugs (f_{ND}), the σ and DR of a filter χ can be expressed as Eqs. A6, and A7 respectively.

$$\sigma = 100 \int_{T_{MIN}}^{T_{MAX}} f_D(\chi) d\chi \qquad \text{Eq. A6}$$

$$DR = \frac{\int_{T_{MIN}}^{T_{MAX}} f_D(\chi) d\chi}{\int_{T_{MIN}}^{T_{MAX}} f_{ND}(\chi) d\chi} \qquad \text{Eq. A7}$$

Eqs. A6 and A7 and the cumulative distribution functions could be used in cases of $\chi = EI_{MW}$, EI_{NHA}, EI_W, and MW.

Depending on the types of the DR and σ functions, i.e. the relative location of the f functions, there are three cases to consider (Fig. **4**). (I) If DR is increasing on the investigated domain of filter χ, then T_{MAX}=+∞ and a T_{MIN} can be calculated. This situation was experienced at the EI's. (II) If DR has a maximum on the domain then both T_{MIN} and T_{MAX} can be calculated as in the case of MW. (III) Finally, if DR is decreasing then T_{MIN}=-∞ and T_{MAX} can be calculated as for logW.

The T_{MIN} and/or T_{MAX} thresholds were calculated by solution of Eqs. A6 and A7, for a set of different σ and DR values using the integral forms, i.e. the cumulative distribution functions of the respective PDF's at χ = MW, EI_{MW}, EI_{NHA}, EI_W. The equations were solved with the aid of Xplore, a program by Prof. David Meredith (Department of Mathematics, San Francisco State University). In the cases where continuous PDF's (χ = NHA, logP, W, logW) could not be fitted, the histograms were used to estimate thresholds applying the definitions of Eqs. A4, and A5. The details of the calculations can be found in the Supplementary Material.

Disease Specificity

A set of 309 drugs of this study (excluding the very small molecules ethanol and piperazine) was divided into sub-sets according to the 14 disease categories of the DrugBank database. These 14 disease categories were: Alimentary tract and metabolism, Blood and blood forming organs, Cardiovascular system, Dermatologicals, Genito-urinary system and sex hormones, Systemic Hormonal preparations (excluding sex hormones and insulins), Antiinfectives for systemic use, Antineoplastic and immunomodulating agents, Musculo-skeletal system, Nervous system, Antiparasitic products, including insecticides and repellents, Respiratory system, Sensory organs, and others which do not fit in the above categories (Various). Non-drug molecules having the closest ΔG were selected for each member of each drug sub-sets using an in-house program. The difference in ΔG was set not to exceed 1 kcal/mol for the drug-non-drug pairs, and indeed it was much less for all cases. One non-drug was used only once for each sub-set. Thus, two sub-sets of compounds (drugs and non-drugs) with overlapping ΔG distributions were available for a disease-specific analysis. Descriptive statistics (median, median absolute deviation, mean, standard deviation, 1st and 3rd quartiles, minimum and maximum) were calculated for all disease categories, and filters by the same program. Boxplots generated by the program R were used for visual comparison of distributions. Only sub-sets having N>10 members were used for final discussion (Fig. **5**). Details of the disease-specific analysis can be found in the Supplementary Material.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Hopkins, A.L. Network pharmacology. *Nat. Biotechnol.*, **2007**, *25*(10), 1110-1111.

[2] Yildirim, M.A.; Goh, K.-I.; Cusick, M.E.; Barabási, A.L.; Vidal, M. Drug-target network. *Nat. Biotechnol.*, **2007**, *25*(10), 1119-1126.

[3] Shoichet, B.K. Virtual screening of chemical libraries. *Nature*, **2004**, *432*, 862-865.

[4] Hopkins, A.L.; Witty, M.J.; Nwaka, S. Mission possible. *Nature*, **2007**, *449*, 166-169.

[5] Villoutreix, B.O.; Eudes, R.; Miteva, M.A.; Structure-based ligand screening: Recent success stories. *Comb. Chem. High T. Scr.*, **2009**, *12*, 1000-1016.

[6] Rajamani, R.; Good, A. C. Ranking poses in structure-based lead discovery and optimization: Current trends in scoring function development. *Curr. Opin. Drug Disc. Devel.*, **2007**, *110*, 308-315.

[7] Dobson, P. D.; Kell, D. B. Opinion - Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat. Rev. Drug Disc.*, **2008**, *7*, 205-220.

[8] Karelson, M. *Molecular descriptors in QSAR/QSPR*. Wiley- Interscience publication: New York, 2000.

[9] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, 2nd ed.*; WILEY-VCH: Weinheim, Germany, 2009

[10] Amidon, G.L.; Anik, S.T. Comparison of several molecular topological indexes with molecular surface area in aqueous solubility estimation. *J. Pharm. Sci.* **1976**, *65*, 801-806.

[11] Gutman, I.; Körtvélyesi, T. Wiener indices and molecular surfaces. *Z. Naturforsch.* **1995**, *50a*, 669–671.

[12] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature*, **2004**, *432*, 855-861.

[13] Katritzky, A.R.; Kuanar, M.; Slavov, S.; Dobchev, D.A.; Fara, D.C.; Karelson, M.; Acree, W.E. Correlation of blood-brain penetration using structural descriptors. *Bioorgan. Med. Chem.*, **2006**, *14*, 4888-4917.

[14] Hitchcock, S.A. Blood-brain barrier permeability considerations for CNS-targeted compound library design. *Curr. Opin. Chem. Biol.*, **2008**, *12*, 318-323.

[15] Charifson, P.S.; Walters, W.P. Filtering databases and chemical libraries. *Mol. Divers.*, **2000**, *5*, 185-197.

[16] Charifson, P.S.; Walters, W.P. Filtering databases and chemical libraries. *J. Comput. Aid. Mol. Des.*, **2002**, *16*, 311-323.

[17] Tounge, B.A.; Pfahler, L.B.; Reynolds, C.H. Chemical information based scaling of molecular descriptors: A universal chemical scale for library design and analysis. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 879-884.

[18] Walters, W.P.; Murcko, M.A. Prediction of 'drug-likeness'. *Adv. Drug Deliver. Rev.*, **2002**, *54,* 255-271.

[19] Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.*, **2003**, *23*(3), 302-321.

[20] Zheng, S.; Luo, X.; Chen, G.; Zhu, W.; Shen, J.; Chen, K.; Jiang, H. A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model.*, **2005**, *45*, 856-862.

[21] Jorgensen, W.L. The many roles of computation in drug discovery. *Science*, **2004**, *303*, 1813-1818.

[22] Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS data mining approaches. *QSAR Comb. Sci.*, **2004**, *23*, 207-213.

[23] Blomberg, N.; Cosgrove, D.A.; Kenny, P.W.; Kolmodin, K. Design of compound libraries for fragment screening. *J. Comput. Aid. Mol. Des.*, **2009**, *23*, 513-525.

[24] Abad-Zapatero, C.; Metz, J.T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today*, **2005**, *10*(7), 464-469.

[25] Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.*, **2005**, *44*, 1504-1508.

[26] Blum, L.C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, **2009**, *131*, 8732-8733.

[27] Martin, E.J.; Critchlow, R.E. Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.*, **1999**, *1*, 32-45.

[28] Walters, W.P.; Ajay; Murcko, M.A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.*, **1999**, *3*, 384-387.

[29] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **1997**, *23*, 3-25.

[30] Fecik, R.A.; Frank, K.E.; Gentry, E.J.; Menon, S.R.; Mitscher, L.A.; Telikepalli, H. The search for orally active medications through combinatorial chemistry. *Med. Res. Rev.*, **1998**, *18*(3), 149-185.

[31] Clark, D.E.; Pickett, S.D. Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today*, **2000**, *5*(2), 49-58.

[32] Ajay; Walters, W.P.; Murcko, M.A. Can we learn to distinguish between "drug-like" and "Nondrug-like" molecules? *J. Med. Chem.*, **1998**, *41*, 3314-3324.

[33] Ghose, A.K.; Viswanadhan, V.N.; Wendolowski, J.J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.*, **1999**, *1*, 55-68.

[34] Sakaeda, T.; Okamura, N.; Nagata, S.; Yagami, T.; Horinouchi, M.; Okumura, K.; Yamashita, F.; Hashida, M. Molecular and pharmacokinetic properties of 222 commercially available oral drugs in humans. *Biol. Pharm. Bull.*, **2001**, *24*(8), 835-940.

[35] Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, **2002**, *45*, 2615-2623.

[36] Lu, J.J.; Crimin, K.; Goodwin, J.T.; Crivori, P.; Orrenius, C.; Xing, L.; Tandler, P.J.; Vidmar, T.J.; Amore, B.M.; Wilson, A.G.E.; Stouten, P.F.W.; Burton, P.S. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J. Med. Chem.*, **2004**, *47*, 6104-6107.

[37] Hann, M.M.; Leach, A.R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 856-864.

[38] Proudfoot, J.R. Drugs, leads, and drug-likeness: An analysis of some recently launched drugs. *Bioorgan. Med. Chem. Lett.*, **2002**, *12*, 1647-1650.

[39] Feher, M.; Schmidt, J.M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 218-227.

[40] Wenlock, M.C.; Austin, R.P.; Barton, P.; Davis, A.M.; Leeson, P.D. A comparison of physicochemical profiles of development and marketed drugs. *J. Med. Chem.*, **2003**, *46*, 1250-1256.

[41] Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol.*, **2000**, *44*, 235-249.

[42] Leeson, P.D.; Davis, A.M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.,* **2004**, *47*, 6338-6348.

[43] Vieth, M.; Siegel, M.G.; Higgs, R.E.; Watson, I.A.; Robertson, D.H.; Savin, K.A.; Durst, G.L.; Hipskind, P.A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.*, **2004**, *47*, 224-232.

[44] Proudfoot, J.R. The evolution of synthetic oral drug properties. *Bioorgan. Med. Chem. Lett.*, **2005**, *15*, 1087-1090.

[45] Leeson, P.D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.*, **2007**, *6*, 881-890.

[46] Perola E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.*, **2010**, *53*, 2986–2997.

[47] Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M.C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model.*, **2008**, *48*, 613-628.

[48] Vieth, M.; Sutherland, J.J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem*., **2006**, *49*, 3451-3453.

[49] Tyrchan, C.; Blomberg, N.; Engkvist, O.; Kogej, T.; Muresan, S. Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds. *Bioorgan. Med. Chem. Lett*., **2009**, *19*, 6943-6947.

[50] Krämer, S.D.; Lombardi, D.; Primorac, A.; Thomae, A.V.; Wunderli-Allenspach, H. Lipid-bilayer permeation of drug-like compounds. *Chem. Biodivers.*, **2009**, *6*, 1900-1916.

[51] Vistoli, G.; Pedretti, A.; Testa, B. Assessing drug-likeness - what are we missing? *Drug Discov. Today*, **2008**, *13*(7/8), 285-294.

[52] Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discov.*, **2003**, *2*, 665-668.

[53] Ganesan, A. The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.*, **2008**, *12*, 306-317.

[54] Lajiness, M.S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behaviour. *Curr. Opin. Drug Disc.*, **2004**, *7*(4), 470-477.

[55] Bhal, S.K.; Kassam, K.; Peirson, I.G.; Pearl, G.M. The rule of five revisited: Applying logD in place of logP in drug-likeness filters. *Mol. Pharm.*, **2007**, *4*(4), 556-560.

[56] Good, A.C.; Hermsmeier, M.A. Measuring CAMD technique performance. 2. How ”druglike” are drugs? Implications of random test set selection exemplified using druglikeness classification models. *J. Chem. Inf. Model*., **2007**, *47*, 110-114.

[57] Gimenez, B.G.; Santos, M.S.; Ferrarini, M.; Fernandes, J.P.S. Evaluation of blockbuster drugs under the rule-of-five. *Pharmazie*, **2010**, *65*(2), 148-152.

[58] Zhang, M.-Q.; Wilkinson, B. Drug discovery beyond the ’rule-of-five’. *Curr. Opin. Biotech*., **2007**, *18*, 478-488.

[59] Dobson, P.D.; Patel, Y.; Kell, D.B. ’Metabolite-likeness’ as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today*, **2009**, *14*(1/2), 31-40.

[60] Tronde, A.; Nordén, B.; Marchner, H.; Wendel, A.-K.; Lennernäs, H.; Hultkvist Bengtsson, U.; Pulmonary absorption rate and bioavailability of drugs *in vivo* in rats: Structure-absorption relationships and physicochemical profiling of inhaled drugs. *J. Pharm. Sci.*, **2003**, *92*(6), 1216-1233.

[61] Ritchie, T.J.; Luscombe, C.N.; Macdonald, S.J.F. Analysis of the calculated physicochemical properties of respiratory drugs: Can we design for inhaled drugs yet? *J. Chem. Inf. Model.*, **2009**, *49*, 1025-1032.

[62] Morphy, R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *J. Med. Chem*., **2006**, *49*, 2969-2978.

[63] Paolini, G.V.; Shapland, R.H.B.; van Hoorn, W.P.; Mason, J.S.; Hopkins, A.L. Global mapping of pharmacological space. *Nat. Biotechnol.*, **2006**, *24*(7), 805-815.

[64] Ajay; Bemis, G.W.; Murcko, M.A. Designing libraries with CNS activity. *J. Med. Chem.*, **1999**, *42*, 4942-4951.

[65] Kelder, J.; Grootenhuis, P.D.J.; Bayada, D.M.; Delbressine, L.P.C.; Ploemen, J.-P. *Pharm. Res.*, **1999**, *16*(10), 1514-1519.

[66] Levin, V.A. Relationship of octanol-water partition coefficient and molecular-weight to rat-brain capillary-permeability. *J. Med. Chem.*, **1980**, *23*, 682-684.

[67] Kaliszan, R.; Markuszewski, M. Brain/blood distribution described by a combination of partition coefficient and molecular mass. *Int. J. Pharm.*, **1996**, *145*, 9-16.

[68] Barn, D.; Caulfield, W.; Cowley, P.; Dickins, R.; Bakker, W.I.; McGuire, R.; Morphy, J.R.; Rankovic, Z.; Thorn, M. Design and synthesis of a maximally diverse and druglike screening library using REM resin methodology. *J. Comb. Chem.*, **2001**, *3*, 534-541.

[69] Chico, L.K.; Van Eldik, L.J.; Watterson, D.M. Targeting protein kinases in central nervous system disorders. *Nat. Rev. Drug Discov*., **2009**, *8*, 892-909.

[70] Garcia-Sosa, A.T.; Mancera, R.L. Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex. *Mol. Inf.*, **2010**, *29*, 589-600.

[71] Lie, M.A.; Thomsen, R.; Pedersen, C.N.S; Schiøtt, B.; Christensen, M.H. Molecular docking with ligand attached water molecules. *J. Chem. Inf. Model.*, **2011**, *51*(4), 909-917.

[72] Kuntz, I.D.; Chen, K.; Sharp, K.A.; Kollman, P.A. The maximal affinity of ligands. *Proc. Natl. Acad. USA*, **1999**, *96*, 9997-10002.

[73] Reynolds, C.H.; Bembenek, S.D.; Tounge, B.A. The role of molecular size in ligand efficiency. *Bioorgan. Med. Chem. Lett.*, **2007**, *17*, 4258-4261.

[74] Reynolds, C.H.; Bembenek, S.D.; Tounge, B.A. Ligand binding efficiency: trends, physical basis, and implications. *J. Med. Chem.*, **2008**, *51*, 2432–2438.

[75] Ferenczy, G.G.; Keserü, G.M. Enthalpic efficiency of ligand binding, *J. Chem. Inf. Model.*, **2010**, *50*, 1536-1541.

[76] Ferrara, P.; Gohlke, H.; Price, D.J.; Klebe, G.; Brooks, C.L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem*., **2004**, *47*, 3032-3047.

[77] Velec, H.F.G.; Gohlke, H.; Klebe, G.; DrugScore^{CSD}-Knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem*., **2005**, *48*, 6296-6303.

[78] Wells, J.A.; McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **2007**, *450*, 1001-1009.

[79] Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem*., **2008**, *29*, 1316–1331.

[80] Olsson, T.S.G.; Williams, M.A.; Pitt, W.R.; Ladbury, J.E. The Thermodynamics Of Protein-Ligand Interaction And Solvation: Insights For Ligand Design. *J. Mol. Biol*., **2008**, *384*, 1002-1017.

[81] Cheng, A.C.; Coleman, R.G.; Smyth, K.T.; Cao, Q.; Soulard, P.; Caffrey, D.R.; Salzberg, A.C.; Huang, E.S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol*., **2007**, *25*(1), 71-75.

[82] Perozzo, R.; Folkers, G.; Scapozzo, L. Thermodynamics of Protein-Ligand Interactions: History, Present, and Future Aspects. *J. Recept. Sig. Transd.,* **2004**, *24*, 1–52.

[83] Reynolds, C.H.; Holloway, M.K. Thermodynamics of ligand binding and efficiency. *ACS Med. Chem. Lett.*, **2011**, *2*(6), 433–437.

[84] Brooijmans, N.; Kuntz, I.D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct*., **2003**, *32*, 335–373.

[85] Brady, G.P.; Sharp, K.A. Entropy in protein folding and in protein-protein interactions. *Curr. Opin. Struct. Biol.*, **1997**, *7*, 215-221.

[86] Zhou, H.-X.; Gilson, M.K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev*., **2009**, *109*, 4092-4107.

[87] Jacobsson, M.; Karlén, A. Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model*., **2006**, *46*, 1334-1343.

[88] Steinberg, I.Z.; Scheraga, H.A. Entropy changes accompanying association reactions of proteins. *J. Biol. Chem.*, **1963**, *238*(1), 172-181.

[89] Karplus, M.; Kushick, J.N. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **1981**, *14*, 325-332.

[90] Tidor, B.; Karplus, M. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J. Mol. Biol*., **1994**, 238, 405-414.

[91] Yu, Y.B.; Privalov, P.L.; Hodges, R.S. Contribution of translational and rotational motions to molecular association in aqueous solution. *Biophys. J*., **2001**, *81*, 1632-1642.

[92] Schwarzl, S.M.; Tschopp, T.B.; Smith, J.C.; Fischer, S. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas correction? *J. Comput. Chem*., **2002**, *23*, 1143-1149.

[93] Carlsson, J.; Åqvist, J. Absolute and relative entropies from computer simulation with applications to ligand binding. *J. Phys. Chem. B,* **2005,** *109,* 6448-6456.

[94] Irudayam, S.J.; Henchman, R.H. Entropic cost of protein-ligand binding and its dependence on the entropy in solution. *J. Phys. Chem. B*, **2009,** *113,* 5871–5884.

[95] Searle, M.S.; Williams, D.H. The cost of conformational order: Entropy changes in molecular associations. *J. Am. Chem. Soc*., **1992**, *114*, 10690-10697.

[96] Murray, C.W.; Verdonk, M.L. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aid. Mol. Des*., **2002**, *16*, 741-753.

[97] Gilson, M.K.; Given, J.A.; Bush, B.L.; McCammon, J.A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J*.; **1997**, *72*, 1047-1069.

[98] Finkelstein, A.V.; Janin, J. The price of lost freedom: entropy of biomolecular complex formation. *Prot. Engin*., **1989**, *3*, 1-3.

[99] Andrews, P.R.; Craik, D.J.; Martin, J.L. Functional group contributions to drug-receptor interactions. *J. Med. Chem*., **1984**, *27*, 1648-1657.

[100] DeWitte, R.S.; Shakhnovich, E.I.; SMoG: de Novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.,* **1996,** *118,* 11733-11744.

[101] Hopkins, A.L.; Groom, C.R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today*, **2004**, *9*(10), 430-431.

[102] Abad-Zapatero, C. Ligand efficiency indices for effective drug discovery. *Expert Opin. Drug Discov*.; **2007**, *2*(4), 469-488.

[103] Hetényi, C.; Maran, U.; García-Sosa, A.T.; Karelson, M. Structure-based calculation of drug efficiency indices. *Bioinformatics*, **2007**, *23*(20), 2678-2685.

[104] Orita, M.; Ohno, K.; Niimi, T. Two ‘golden ratio’ indices in fragment-based drug discovery. *Drug Discov. Today*, **2009**, *14*(5/6), 321-328.

[105] Keserü, G.M.; Makara, G.M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov*., **2009**, *8*, 203-212.

[106] Nissink, J.W.M. Simple size-independent measure of ligand efficiency. *J. Chem. Inf. Model.,* **2009**, *49,* 1617–1622.

[107] Bembenek, S.D.; Tounge, B.A.; Reynolds, C.H. Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today*, **2009**, *14*(5/6), 278-283.

[108] Reitz, A.B.; Smith, G.R.; Tounge, B.A.; Reynolds, C.H. Hit triage using efficiency indices after screening of compound libraries in drug discovery. *Curr. Top. Med. Chem.,* **2009**, *9,* 1718-1724.

[109] Murray, C.W.; Rees, D.C. The rise of fragment-based drug discovery. *Nat. Chem*., **2009**, *1*, 187-192.

[110] Garcia-Sosa, A.T.; Oja, M; Hetényi, C.; Maran, U. Disease-specific differentiation between drugs and non-drugs using principal component analysis of their molecular descriptor space. *Molecular Informatics*, **2012**, *In the press.*.

[111] Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **2007**, *35*, D198-D201.

[112] Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico

drug discovery and exploration. *Nucleic Acids Res*., **2006**, *34*, D668-D672 Sp. Iss. SI.

[113] Neidle, S.; Thurston, D. E. Chemical approaches to the discovery and development of cancer therapies. *Nat. Rev. Cancer* **2005**, *5*, 285-296.

[114] von Nussbaum, F.; Brands, M.; Hinzen, B.; Weigand, S.; Habich, D. Antibacterial natural products in medicinal chemistry - Exodus or revival? *Angew. Chem.-Int. Edit.* **2006**, *45*, 5072-5129.

[115] Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **2007**, *6*, 29-40.

[116] Barratt, M.D. Quantitative structure-activity relationships for skin permeability. *Toxicol. Vitro,* **1995**, *9*, 27-37.

[117] Cross, S.E.; Magnusson, B.M.; Winckle, G.; Anissimov, Y.; Roberts, M.S. Determination of the effect of lipophilicity on the *in vitro* permeability and tissue reservoir characteristics of topically applied solutes in human skin layers. *J. Invest. Dermatol.,* **2003***, 120*, 759-764.

[118] Rees, D.C.; Congreve, M.; Murray, C.W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discov.*, **2004**, *3*, 660-672.

[119] Koehn, F.E.; Carter, G.T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.*, **2005**, *4*, 206-220.

[120] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem*., **2004**, *47*, 2977-2980.

[121] Aizawa, M.; Onodera, K.; Zhang, J.-W.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata. K. KiBank: A database for computer-aided drug design based on protein-chemical interaction analysis. *Yakugaku Zasshi,* **2004**, *124*, 613-619.

[122] Zhang, J.–W; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem*., **2004**, *28*, 401-407.

[123] Ababou, A.; Ladbury, J.E. Survey of the year 2005: literature on applications of isothermal titration calorimetry. *J. Mol. Recognit*., **2007**, *20*, 4-14.

[124] Hetényi, C.; Paragi, G.; Maran, U.; Timár, Z.; Karelson, M.; Penke, B. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc*., **2006**, *128*, 1233-1239.

[125] Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design,* **2000**, *19*, 47-66.

[126] *Marvin 4.8.1*, ChemAxon **2008**, www.chemaxon.com.

[127] *SYSTAT 12*, SYSTAT Software, Inc., 1735 Technology Dr., Ste. 430, San Jose, CA 95110.

[128] Shapiro, S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611.

[129] Kolmogorov, A. *Foundations of the Theory of Probability*, 2nd ed.; New York: Chelsea, 1956.

[130] Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, **1952**, *23*, 193–212.

[131] Otto, M. *Chemometrics*, Wiley-VCH: Weinheim, Germany, 1999.

[132] Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **1947**, *18*, 50–60.

[133] *Program package R*. http://www.r-project.org (accessed 24 February **2009**).

[134] Heckert, N.A.; Filliben, J.J. *NIST Handbook 148: DATAPLOT Reference Manual, Volume I: Commands*, National Institute of Standards and Technology Handbook Series, 2003.

Received: October 11, 2011 **Revised: January 26, 2012** **Accepted: January 27, 2012**