

toth.gergely_304_24

**Adatok értelmezése szimulációkkal és a többváltozós
statisztika eszközeivel a kémiában**

Tóth Gergely

MTA doktori értekezés

Budapest

2025.

Szakmai köszönetnyilvánítás

Szakmai tevékenységemet négy részre tagolhatom, ahol mindegyik résznél kiemelt köszönetet kell mondanom: a szakdolgozói és doktori időben nyújtott segítségért Pusztai Lászlónak, a mainzi kutatói munkáért Karl Heinzingernek, a 90-es évek második felétől Baranyai Andrásnak az igényes elméleti munkába való bevezetésért és 2010-től Szepesváry Pálnak a kemometria irányába való eltérítésért. Köszönet illeti az összes hallgatót, akik engedték, hogy néha partikuláris problémák kiszámolásával bízzam meg őket. Külön köszönet illeti azokat, akiknek társ- vagy egyedüli témavezetője lehettem PhD tanulmányaik során. Itt két nevet emelnék ki, akiktől a számítástechnika területén inkább én tanultam: Vrabcz Attilát és Király Pétert. Köszönet illeti minden további társszerzőmet is.

Köszönettel tartozom azoknak is, akik nem közvetlen szakmai támogatást nyújtottak, hanem lehetőséget adtak a munkához: Elsősorban a feleségemnek, aki gyerekeink nevelésében nálam nagyobb részt vállalt és így megszakítás nélkül teljes állásban dolgozhattam. Másodsorban is neki, de már a gyerekeinkkel közösen, hogy elfogadták a nagyobb szellemi szabadságot, de kisebb anyagi megbecsülést jelentő akadémiai karrierválasztásomat.

Tartalomjegyzék

1. Bevezető	3
2. Célkitűzések	7
3. Kérdések a fordított Monte Carlo módszerrel kapcsolatban	11
<u>3.1 A fordított Monte Carlo módszer egyértelműsége és sokasága</u>	11
<i>A fordított Monte Carlo módszer és a fluktuációk [s18]</i>	12
<i>A párkorrelációs és a szerkezeti függvény fluktuációja [s15] [s18]</i>	13
<i>A fordított Monte Carlo módszer és a termodinamika [s23]</i>	14
<i>A fordított molekuláris dinamika [s28]</i>	16
<i>Ökölszabályok a fordított Monte Carlo szimulációhoz [s21] [s24]</i>	17
<u>3.2 A részecskék belső és a részecskék közötti szerkezet párhuzamos meghatározása a fordított Monte Carlo módszerrel</u>	18
<i>A polidiszperzitás szerepe a fényszóródási mérések fordított Monte Carlo kiértékelésénél [s16] [s22] [s25]</i>	19
<i>Ni katalizátor szerkezetének vizsgálata fordított Monte Carlo szimulációval [s30]</i>	22
<i>Egy nemionos felületaktív anyag szerkezetének vizsgálata fordított Monte Carlo szimulációval [s33]</i>	25
4. Kölcsönhatások, a statisztikus mechanikai inverz probléma, „coarse graining”	29
<u>4.1. Potenciális energia felület meghatározása kvantumkémiaiával, klasszikus mechanikai kölcsönhatások paraméterezése</u>	29
<i>Hg-ion kölcsönhatási potenciálok paraméterezése [s10]</i>	29
<i>Ion-víz kölcsönhatások paraméterezése [s14]</i>	30
<u>4.2. Potenciál meghatározás „on fly” a 90-es években</u>	33
<i>Az FSCF-MNDO Monte Carlo módszer és alkalmazásai [s5] [s6] [s7] [s19]</i>	33
<i>A Car-Parrinello módszer alkalmazása egy legkörkémiai kérdésnél [s17]</i>	35
<u>4.3. Az inverz probléma alapjai saját tapasztalatokkal</u>	38
<i>Henderson tétele a mi szemünkkel [s24]</i>	38
<i>A fordított Monte Carlo módszer fázistérbeli mintavételezése [s32]</i>	39
<i>A fordított molekuláris dinamika és a valós dinamika [s32]</i>	40
<u>4.4. Saját módszerek az inverz probléma megoldására</u>	42
<i>A Born-Green-Yvon hierarchia és a fordított Monte Carlo módszer kombinálása [s26]</i>	42
<i>A módosított inverz Monte Carlo eljárás [s27] [s29]</i>	44
<i>Az inverz probléma megoldása mesterséges ideghálával [s31]</i>	48
<i>Párpotenciál meghatározása potenciális energia illesztésével [s36]</i>	49

<u>4.5 Az inverz probléma későbbi vonatkozásai</u>	55
<i>A 2007-es áttekintő cikkem utáni kutatási irányok [37]</i>	55
<i>Egy alkalmazás: párkölcsönhatás meghatározása fázismező kristály modellezés eredményeire [43] [45]</i>	56
<u>4.6 Lépések a kinetikai Monte Carlo irányában</u>	58
<i>Karbamid kristályosodása [s39]</i>	58
<i>Argon kristályosodás sebességi állandói [s46]</i>	61
<i>Több minimumú rendszerek kommittor függvényei [s54]</i>	63
5. Kemometria	68
<u>5.1 Szeriálás</u>	68
<i>Saját szeriáló módszerek és alkalmazásuk [s41] [s53] [s63]</i>	69
<u>5.2 In memorian Szepesváry Pál: trendvizsgálat, normáeloszlás kérdései</u>	77
<i>Trendvizsgálat varianciaanalízissel [s42] [s47]</i>	77
<i>Kérdések a normál eloszlással kapcsolatban [s44] [s51]</i>	80
<u>5.3 Az allokáció és a modellvalidálás kérdései</u>	83
<i>Periodikus adatok allokációja kemometriai számolásokban [s52]</i>	84
<i>Az allokáció hatása validációs és kísérlettervezési paraméterekre [s58]</i>	86
<i>Validációs stratégiák az OECD 4. elv és a mintaméretfüggés figyelembevételével [s59] [s61] [s65]</i>	91
<i>Nagy hatású pontok detektálása keresztellenőrzési és illeszkedési validációs paraméterek alapján [s48]</i>	100
<i>A robusztusság vizsgálatának két alapmódszere [s64]</i>	103
6. Összefoglalás	108
7. Tézispontok	111
8. Irodalomjegyzék	115
<i>Az értekezésben szereplő saját munkák</i>	115
<i>További saját publikációk</i>	118
<i>Külső irodalomjegyzék</i>	120
9. Függelék	127
<u>9.1. Rövidítések, angol szakkifejezések és a magyar szakzsargon feloldása/értelmezése</u>	127
<u>9.2. Validációs és optimálási paraméterek táblázata</u>	135

1 Bevezető

A természettudomány fejlődése szorosan összekapcsolódott az adatok gyűjtésével, értelmezésével és feldolgozásával. Kezdetben egyszerű megfigyelések adatait próbálták meg a nagy elődök értelmezni, majd egyre több lett a szisztematikus és célzott adatgyűjtés. A műszeres vizsgálatok tömegessé tették az adatokat, a számítógépek elterjedésével megjelentek az elvi modellekre generált adatok is. Ezek egy részét a valóságot imitáló/követő szimulációk segítségével generáljuk, egy másik része csak egyszerű statisztikai alapon létrehozott adatsor, sokszor egy-egy hatás, elvi modell vagy jelenség vizsgálatára. Eddigi tudományos tevékenységem adatokhoz kötődött, felölelve a teljes spektrumot az egyszerű észleléstől (kristályok makroszkopikus alakja), az összetett méréseken (diffrakciós mérések), számítógéppel generált kvázi mérési adatsorokon (célzott adatstruktúrák), klasszikus mechanikai szimulációkon át (szerkezeti, termodinamikai és dinamikai adatok) az összetett adatokig (molekuláris deszkriptorok, kvantummechanikai számolások, ipari adatsorok).

Kutatásaim során többnyire olyan módszereket alkalmaztam és fejlesztettem, amelyek az adatokból a klasszikus statisztikai módszereken túlmutató információk kinyerését tették lehetővé. Olyan rejtett információk keresésével próbálkoztam, amelyekhez néha szimulációkat és számolásokat kellett végrehajtanom, gyakran megfordítva a természettudomány egy-egy bevett irányát, pl. szerkezetből kölcsönhatási potenciált határoztam meg, vagy kinetikai állandókat vezettem le makroszkopikus morfológiából. Sokszor csak a keresett információ töredékét és azt is csak feltételesen tudtam megkapni. A számomra a legérdekesebb kutatások azok voltak, ahol egy szokatlan kérdésre szokatlan megoldással találtam választ.

Az első témakör a folyadékok és hasonló rendszerek diffrakciós méréseihez kapcsolódik, itt a fordított Monte Carlo szimulációt [Pusztai&McGreevy1988] vizsgáltam statisztikus mechanikai szemmel, illetve összetett anyagok szerkezetének felderítésével kapcsolatban. Ezzel párhuzamosan kerültem a többi klasszikus mechanikai szimuláció területére, egyre nagyobb súllyal a szimulációkhoz szükséges kölcsönhatások meghatározásának irányába, ami a második szélesebb kutatási témám lett. A hagyományos kutatási megítélés szerint ezen a területen végeztem a legtöbb elméletet igénylő „komoly” kutatást, amikor a statisztikus mechanika potenciálmeghatározáshoz kapcsolódó inverz problémájának a megoldására dolgoztam ki eljárásokat. A témából egy kiágazásnak tekinthető a kinetikai adatok meghatározása, ahol mezoszkopikus szimulációs módszereket használtam. A harmadik témakör a kemometria. Itt két területre jutott nagyobb súly, az egyik a szeriálás, vagyis az adattömbök átrendezése, a másik az adatok értelmezésére és predikciójára készített modellek validálásának kérdése. További kemometriai publikációim vannak a trendvizsgálatokkal és a normál eloszlás néhány részletével kapcsolatban is.

A látszólag eltérő három fő terület (vagy 10-15 részterület) azonban ugyanolyan módon mind számítógépes kutatást jelentett, adatok felkutatását, beolvasását, generálását, esetleges szimulációk elvégzését, a hozzá kapcsolódó programozást, programok módosítását és alkalmazását. Kutatói pályámat a témák és a szükséges ismeretek folyamatos eltolódása jellemezte. A kezdeti szimulációs témához nagyon hamar hozzákerült valamennyi statisztikus mechanikai ismeret, numerikus matematikai módszerek használata, majd a többváltozós statisztika kémiai ága, a kemometria. Ezekre a területekre nem csak a kutatás miatt kerültem, hanem ugyanerre haladtam az oktatási feladataimmal is az ELTE-n. Hogy melyik-melyiket segítette, az sokszor utólag azonosítható. Például a numerikus matematika oktatásakor sok olyan módszert kellett megértenem, amelyeket később alkalmazhattam a kutatásban, esetleg egy adott cikk csak arról szólt, hogy egy problémát milyen a területen új numerikus módszerrel oldottam meg. Szakmai utammal kapcsolatban két kollégát mindenképpen ki kell emelnem. Szimulációs területen, illetve azok statisztikus mechanikai megközelítésével kapcsolatban Baranyai András, a kemometria felé fordulásomnál pedig Szepesváry Pált. Nélkülük nem kutattam volna se az inverz probléma, se a kemometria területén, még ha egyikőjükkel se kerültem hagyományos főnök-beosztott kutatói kapcsolatba. A kutatási területeken való váltásokhoz hasonlóan számos számítástechnikai módszert is meg kellett ismernem, bár itt inkább az eredményezi a szélesebb skálát, hogy ez az értekezés több mint 30 év munkájából készült. Különböző programnyelveket használtam: Fortran, C, majd Python, volt időszakom a Mathematica, a Matlab és az R alkalmazására, közben a standard Microsoft szoftverek mellett szövegszerkesztőnek Latex-t, gnuplot-t és számos egyedi ábrázoló-szerkesztő programot használtam. De szinte sose profi szinten! Ez alól talán a Fortran volt a kivétel, ahol a kezdeti években tiszta munkaidőben is több évnnyit programoztam azon a nyelven. Mindeközben oktatási feladataimba is rövidebb-hosszabb időszakokra bevettem ezeket a szoftvereket.

A dolgozat gerincét azok a cikkek adják, amelyek belefértek a választott címbe (45 cikk). Azokat a munkákat hagytam ki, amelyek témája az adatfeldolgozási címtől távoli volt. Az 1993-ban szerzett egyetemi doktori oklevelemet az ELTE két évvel később nyilvánította egyenértékűvé a közben bevezetett PhD fokozattal. A korábbi fokozatszerzésre vonatkozó elvi kizárása a publikációimnak az egyetemi doktori fokozathoz kapcsolódik, mivel az ELTE nem PhD címet adományozott nekem 1995-ben, hanem az 1993-as eljárást nyilvánította egyenértékűvé a PhD-val. Nem vettem be a dolgozatba a mainzi posztdoktori időm felületi molekuláris dinamikai szimulációit, illetve azokat az alkalmazásokat, ahol egy már kész potenciállal molekuláris szimulációk történtek. Kihagytam több egyedi témát, pl. izotóphelyettesítéssel kapcsolatos szimulációimat, a coulomb kölcsönhatás periodikus rendszerekben levő illesztésével kapcsolatos munkát, egy-két alkalmazott kvantumkémiai közreműködésemet, egy-egy konkrét adatsorra való kiértékelést. Kihagytam azokat a dolgozat címétől eltérő kutatásokat is, ahol úgy lettem társszerző, hogy se a ráfordított idő, se a megvalósítással kapcsolatos ötleteknél nem

volt jelentős a hozzájárulásom, pl. az elmúlt évek pseudouridinnal foglalkozó cikkeit. Az így kihagyott közleményeim száma 20. Történt egy további szűkítés is, nem minden a dolgozatban tárgyalt publikációhoz kapcsolódik tézispont. Két esetben ez olyan többszerzős publikációt jelent, ahol se első, se levelező szerző nem voltam. Az adott kutatásban egy-egy jól lehatárolt feladatom volt, ami szorosan kapcsolódik a dolgozat témájához és csak ez szerepel a dolgozatban. A tézispontban való szerepléssel formailag megnehezítettem volna, hogy pl. az első szerzők az adott publikációkban saját minősítésükben felhasználják. A harmadik kihagyott publikáció egy „comment”, ez formai szempontból nem alkalmas. Ez utóbbinál ehelyett egy összefoglaló mű szerepel a tézispontnál hivatkozásnak. A dolgozattal együtt a kérelemben leadott lista 61 publikációt tartalmaz. Nem szerepel a hivatalos listán a már említett „comment”, két oktatással kapcsolatos cikk és egy a leadás előtti MTMT zárolás alatt megjelent publikáció. A dolgozatban szereplő saját munkák számozása a teljes 65 tételes listára vonatkozik az időrendi sorrendnek megfelelően. „s” előjellel, ha tárgyalom az értekezésben, „e” előjellel, ha legfeljebb megemlítem.

A dolgozat három fejezete időrendi sorrendet mutat, a tudományos munkám három szakaszát. A 2. és 3. szakasz mellett, részben között, volt egy időszak, amikor időm nagyrésztét egyetemi oktatásszervezés, oktatás, oktatási adminisztráció és hallgatókkal való egyedi, nem szakmai foglalkozás tette ki (2004-2017). Ez kutatói eredményeimben egy inaktív időszakot eredményezett, de az ebben a közegben való mozgásomnak mind a már pár éve lezárt hazai szakaszát, mind a még élő nemzetközi részét ma is fontosnak tartom. Ebben az időszakban részt vettem néhány olyan közlemény elkészítésében, amelyek valamilyen oktatási relevanciával bírtak [e50, e56, e57]. Ugyanígy örülök, hogy szervezőjeként, sőt három alkalommal ötletadójaként is részt vehettem egy őszi iskola (Gyógyszeripar A-tól Z-ig, Egis-ELTE őszi iskola, 2017.) és három workshop létrehozásában (Szabadenergia workshop, 2011., Szimulációk több skálán, időben és térben, 2022., és Occam borotvája vagy gépi tanulás? Természettudományos modellezés kevés vagy sok paraméterrel, 2024.). 2019 óta szervező házigazdája lehetek az Alkímia Ma előadássorozatnak, amely évi 12-13 ismeretterjesztő előadásból és azt követő látványos kísérleti bemutatóból áll.

Egy hagyományos értekezésnél részletesen szokott szerepelni elsőként a bemutatott kutatások irodalmi előzménye, majd az eredmények értékelése a tudomány jelenlegi állása és az elvégzett kutatás tudományos visszhangja alapján. A klasszikus értekezésben a használt eszközöket és a módszereket is illik bemutatni a saját kutatások előtti részben. Az esetemben ez az út nem járható, mivel sokszor egy-egy cikkhez teljesen önálló irodalmi összefoglalás és a módszerek egyedi leírása kellene. A tudományos utóélet számítógépes kutatásoknál sokszor irrelevánssá válhat, pl. egy 20-30 évvel ezelőtt Cray szuperszámítógépen, ma egy telefonon kiszámolható eredményre. A kutatásaimhoz kapcsolható eszközpark alapvetően egy asztali számítógépet, monitort és nyomtatót jelentett az

toth.gergely_304_24

elmúlt közel 35 évben. Ilyen környezetben dolgoztam az ELTE kémiai épületének számos szobájában, mainzi posztdoktori időszakomban, vagy pár hetes erdélyi, szlovén és lengyel oktatói és kutatói útjaimon. Az idők során annyi változott, hogy az asztal szélére felhalmozott könyv és újságkupacokat ma már többnyire ezek digitális verzióira cseréltem. A 90-es évek végéig ezen felül a komolyabb számításigényű feladatoknál igénybe vettem a mainzi, majd az ELTE Elméleti Kémiai Tanszék munkaállomás jellegű nagyobb teljesítményű számítógépeit, jellemzően IBM RISC-eket. A 90-es évek végétől tudatosan már csak olyan kutatásokba kezdtem, amelyek egy személyi számítógépen elvégezhetőek voltak. Kerülni próbáltam azt a frusztrációt a publikálásnál, ahol az eredményeket a számítási teljesítmény korlátozta és a bírálók kérdéseinek a felénél csak erre tudtam hivatkozni.

A függelék első fejezete egy összevont rövidítésjegyzék és pár mondatos leírása egy-egy lényegi módszernek, illetve leírása a dolgozatban szereplő olyan szakkifejezéseknek, amelyeket nem fordítottam le széleskörűen elterjedt magyar szakkifejezés hiányában. A függelék ilyen formája az elővédéskor derült ki, mivel elvétele található olyan kolléga, aki mind a szimulációs, mind a kemometria területen járatos. Ha valaki számára nem világos egy szakmai terminus technikus az értekezésben, érdemes először ebben a függelékben rákeresni.

2. Célkitűzések

Ebben a fejezetben a dolgozat célkitűzéseit nem a hagyományos szerkezetben írom le, hiszen az elvégzett kutatásokhoz nem kapcsolható egy közös anyagcsoport vagy egyetlen közös módszer. Helyette tématerületenként megemlítem, hogy mi volt az elsődleges motiváció, illetve mik voltak azok a menet közben kialakuló új kérdések, amelyekre választ kerestem. A publikációk nagy száma miatt csak a fontosabb kérdéseket és célokat sorolom fel. Egyes publikációkhoz egyáltalán nem írom le a célokat, ha az adott esetben egy tervezett kutatási téma mélyebb megértésekor részeredményként alakult ki az adott önálló publikáció; egy az eredeti témából elágazó kutatás és nem egy tervezett stratégia vezetett oda.

A 3.1-es fejezetben a fordított Monte Carlo módszer (RMC) [Pusztai&McGreevy 1988] algoritmusával és a szimuláció által generált konfigurációk statisztikus mechanikai szemmel való vizsgálatával foglalkozom. Az RMC eredetileg egy sztochasztikus minimalizáló eljárás, ahol rendezetlen kondenzált fázisokról készült diffrakciós mérésekkel konzisztens részecske elrendeződések generálása a cél. A névhez ragadó szimuláció kifejezés téves képzetet keltett már a kezdetekben is. A módszer egyértelműségének a kérdése szinte a módszer 1988-as bevezetésével egyidőben alakult ki, hogy a kapott konfigurációk (pontosabban az ezeket leíró szerkezettel kapcsolatos különböző függvények) mennyire tükrözik a vizsgált rendszerek valós viszonyait. Első körben az általános egyértelműségi kérdések jelentek meg, hogy az egyváltozós függvények segítségével generált konfigurációk mennyire egyértelmű és stabil szerkezeteket generálnak, valamint milyen információk szükségesek pl. többkomponensű, molekuláris és kovalens rendszerekben. Ehhez képest az általam 1995-től végzett kutatásokban a fő hangsúly az RMC statisztikus mechanikai értelemben vett sokaságán volt, milyen termodinamikai és szerkezeti jellemzőkkel vannak gondok, lehet-e az RMC-t közelebb vinni a referencia kanonikus sokasághoz. A kutatás során több új részkérdés merült fel, pl. a párkorrelációs függvény fluktuációjának alakját ábrákon vizsgáltam, a kialakult sejtéshez elméleti magyarázatot kellett keresni. Hasonlóan menetközben alakult ki a dinamikai változat kérdése, reálisan számolható-e egy szimuláció minden pillanatában a χ^2 RMC potenciál gradiense alapján.

A 3.2 fejezet az egyetemi doktorimban szereplő egyik téma újrakezdése volt. Az első kolloidokra vonatkozó RMC-s számolásaimban észlelt kontakt párok és néha furcsa alakú párkorrelációs függvények valós létezésében nem voltam biztos, ezért az akkor megjelenő Monte Carlo alapú Fourier transzformációs eljárásban, az MCGR [Pusztai&McGreevy 1997] algoritmusban egy gyors szkennelési lehetőséget láttam annak a vizsgálatára, hogy ezt nem a helytelen részecskeátmérő-választás okozza-e. Ahogy jobban megismertem a kisszögű mérések problémáit, úgy kezdtem el gyanakodni, hogy hátha nem is az átmérő értéke, hanem a szimulációkban feltételezett monodiszperzitás okozza a szokatlan

párkorrelációs függvényeket. A két további rendszer RMC szimulációjára (Ni katalizátor és BRIJ-35) konferenciák alkalmával jött az ötlet, ahol ezeknek a rendszereknek hagyományos kiértékeléséről hallottam előadásokat. Itt a feladat a megfelelő modellek és az RMC szimulációs algoritmus módosítása volt a célzott kihívás.

A 4.1 fejezet potenciálmeghatározásainak feladatát részben németországi főnökömtől, Karl Heinzinger-től kaptam. A célkitűzés megfelelt az akkor tipikus alkalmazott kvantumkémiai eseteknek: mekkora a legnagyobb rendszer, amelyen a kvantumkémiai számolásokat el tudom végezni úgy, hogy a meghatározandó klasszikus kölcsönhatások valamennyire reálisan legyenek.

A 4.2 fejezet szemiempirikus kvantumkémiai Monte Carlo szimuláció témáját Náray-Szabó Gábor vetette fel. Ennél a témánál a célkitűzés egy az adott kvantumkémiai szinten minden Monte Carlo lépésben energiát számoló algoritmus megvalósítása volt. Számomra a cél maga a megvalósíthatóság volt, hiszen egy addig szinte alig létező kvantumkémia+szimuláció kombinációt kellett számítási algoritmusában kidolgoznom és működően beprogramoznom úgy, hogy reálisan alkalmazható is legyen a kor számítógépes kapacitása mellett. Az alfejezet második témájánál fontos volt számomra, hogy megismerjem az időszak egyik legújabb szimulációs módszerét, a Car-Parrinello molekuláris dinamikai szimulációt [Car&Parrinello 1986]. Valószínűleg az első magyar kutató voltam ezen a témán, sőt a választott rendszerem, a szilárd határfelületen levő katalízis ilyen megközelítése is sokat ígért, bár a célt, a feltételezett katalízist nem tudtam igazolni és értelmezni ezzel a szimulációs technikával.

A 4.3-as fejezetben szereplő eredmények a 4.1-es fejezet tapasztalataiból indultak ki, de kicsit bolyongás jelleggel. Kerestem a megoldást, hogy a klasszikus Monte Carlo és molekuláris dinamika keretein belül az RMC közelítéséből indulva lehet-e jó szerkezetet és dinamikát elérni egyszerre. A 4.4-es fejezetben a továbblépést az inverz probléma irányába tettem meg. A Born-Green-Yvon (BGY) hierarchia [Hill 1956] és az RMC módszer kombinálása egy feldobott labda volt, az RMC módszer születése óta néha utaltak rá a szakirodalomban, de a megvalósításhoz kellett Baranyai András ismerete a triplétt korrelációs függvények számolásával kapcsolatban. Itt a fő célkitűzés a BGY hierarchia numerikus igazolása és egy megvalósítható potenciál meghatározás bemutatása volt a BGY hierarchia és az RMC kombinálásával, beleértve az ehhez kapcsolódó elméleti és numerikus matematikai kérdések megoldását is. A következő esetben már célzottan az inverz problémára koncentráltam, teljesen kilépve az RMC témakörből: az inverz Monte Carlo (IMC) módszer [Ljubartsev&Laaksonen 1995] gyengeségeit próbáltam meg módosításokkal kiküszöbölni, pl. kikerülni a párkorrelációs függvény, mint szekunder kísérleti adat használatát, illetve a numerikus instabilitását az IMC-nek. Az energia alapú inverz módszer alapötlete a lineáris egyenletrendszerekre vonatkozó numerikus matematikai ismereteimből alakult ki, kombinálva a szimulációs szakirodalom néhány alapegyenletével. A célt sikerült elérni, energia alapon is meghatároztam párkölcsönhatási

potenciálokat. A harmadik reprodukálható tényező, a dinamika illesztése is elkezdődött, de ebből publikálható szintig nem jutottunk el. Az inverz probléma mesterséges ideghálós megoldása egy korai tudományos játék volt az akkor még szinte ismeretlen fekete doboz modellezésnek, általánosabban a gépi tanulásnak. A cél a Henderson-féle funkcionál [Henderson 1974] fekete doboz modellel való közelítése volt.

A 4.6. fejezet első témája a kinetikai Monte Carlo szimuláció a kristályosodáshoz kapcsolódik. Piana és munkatársai [pl. Piana&Gale 2005] több cikkben alkalmazták a hagyományos irányból induló megközelítést az idő és méretskálák váltásának, vagyis a kvantumkémia irányából (klasszikus kölcsönhatási potenciálok meghatározására) a klasszikus mechanikai szimulációkon át (elemi lépések sebességi állandójának meghatározása) jutottak el a mezoszkopikus kinetikai Monte Carlo szimulációkban elért már „látható” mérettartományig. A nikkeles és BRIJ35-ös RMC szimulációk és a neurális hálós inverz probléma megközelítésben már olyan rejtett információt kerestem az adatokban, amelyekkel kapcsolatban előtte nem voltam biztos, hogy tényleg tartalmazzák-e a bemenő adatok. Itt a karbamidnál bevállaltam annak a kérdésnek a vizsgálatát, hogy egy kristály makroszkopikus formájából vissza lehet-e bármilyen mikroszkopikus folyamatra következtetni, akár a mikroszkopikus sebességi állandók szintjéig. Az argonos kristálynövekedési közlemény a szabadenergia számoló klasszikus szimulációs lehetőségeket és a kinetikai Monte Carlo szimulációt kötötte össze. Kicsit kilóg a több minimumú rendszerek kommittor függvényeinek vizsgálata (kommittor definícióját lásd a függelékben). Meglepő módon, nem találtuk meg az irodalomban, hogyan kell definiálni több minimum esetén a kommittort és a kapcsolódó mennyiségeket, amelyek pl. a kinetikai Monte Carlo szimulációhoz szükséges információkat adja. A célkitűzés itt ezeknek függvényeknek a levezetése és egy egyszerű modellrendszeren való bemutatása volt. Menetközben alakult ki másodlagos célként, hogy milyen grafikai megoldásokkal lehet ezeket a függvényeket szemléltetni.

A kemometria témánál az értekezésben először a szeriálást mutatom be az 5.1 fejezetben. Itt a célkitűzést az 5. fejezetben is részletezett előadás adta az Elba hosszában mért szennyezőkről. Lehet-e úgy rendezni egy adatmátrixot, hogy a sorok sorrendje valami hasonlóságát mutassa az objektumoknak. Ugyanakkor azt, hogy melyik két objektum miatt, melyik változók miatt lett hasonló, szintén mutassa a mátrix. Az Elbára vonatkozó példánál, lehet-e úgy rendezni az adatokat, hogy pl. az azonos jellegű ipari szennyezőket tartalmazó mintavételi helyeket egymás mellé tegye és azt is lássuk, melyek ezek a szennyezők. Az első szeriálási megoldásunknál kiemelt szerepet kapott az adatmátrix diagonálisa (átlója), de ez csak egyfajta lineáris rendezést tett lehetővé. Nem volt alkalmas pl. annak a mutatójára, ha több különböző csoport is részben ugyanazokat a jellegzetes változókat tartalmazta. A második módszer, a folszeriálás ennek a megoldására született.

Az 5.2 fejezetben a trendvizsgálattal kapcsolatos kutatást egy külső oktatási feladat hozta, ahol a Neumann-féle trendtesztet kellett tanítani, annak minden nehezen értelmezhető részével, pl. a kritikus értékek mögött bújó elméleti ismeretekkel. Abban az időben sokat foglalkoztam varianciaanalízissel (ANOVA) oktatási feladatok miatt. A kíváncsiság hozta a célkitűzést, megoldható-e a trendvizsgálat az ANOVA feltételezéseivel analóg módon, beleértve az ehhez a teszthez tartozó F-eloszlást is. Az időfüggő trendvizsgálat ötletét a mainzi szimulációs múltam adta, ahol sokat számoltam autokorrelációs függvényeket. A keresztkorreláció ötletet Salma Imre kollégám légköri/légszennyezési adatai adták. A cél egyszerű volt: legyen egy igen/nem statisztikai teszt jellegű választ adó módszer az időfüggő adatok időeltolódásos korrelációinak vizsgálatára.

A normál eloszlással kapcsolatos kutatások célja annak a szisztematikus megértése volt, hogy miért kaptam egy oktatáshoz szimulált adatsorra elsőre érthetetlen eredményeket, a részletek leírása az adott fejezetnél szerepel.

Az 5.3 fejezetben a periodikus adatok allokációjának ötlete onnan eredt, hogy Mahalanobis transzformációt [Mahalanobis 1936] szerettem volna végezni egy Monte Carlo szimulációval kapott konfiguráción. Az eredeti kérdés időközben lényegtelenné vált, de közben megértettük a periodikus adatok kemometriai feldolgozásának hiányosságait. A publikáció így kialakult célja egy olyan adatablak választási elv kidolgozása lett, ami illeszkedik a hagyományos intervallum és nem tórusz alapú kemometriai adatkiértékelésekhez.

A kísérlettervezési optimálisok és a validációs paraméterek korrelációi vizsgálatának eredeti célja az volt, hogy szerettem volna megérteni gyakorló módon is a kísérlettervezést és hogy miért van ennyi kísérlettervezési optimális. A végső cél fokozatosan alakult ki, vagyis a modellezés előtti kísérlettervezés és a modell teljesítménye tényleg olyan jól korrelál-e, mint ahogy a kísérlettervezési könyvek állítják.

A mintaméretfüggéses vizsgálatok alapötlete az volt, hogy a belső és külső validálási eljárások adáz híveinek vitáit nem értettem. Úgy éreztem, hogy léteznie kell valami kritikus mintaméretnek, ahol a belső validációról át kell térni a külső validációra. A kritikus mintaméretet nem találtuk meg, de a közben megismert OECD irányelvek [OECD 2004] megadtak egy kutatási irányt fokozatosan újabb és újabb kérdéseket felvetve. Persze itt is nagy volt a miértek szerepe, főleg, hogy a saját ismereteimet rendezzem a hatalmas és sokszor ötletszerűnek tűnő validációs paraméter rengetegben.

A dolgozat végén tárgyalt két cikknél is csak egy-egy kérdésre kerestem a választ. Igazak-e az irodalmi állítások, hogy pl. „Beware of Q2!” [Golbraikh&Tropsha 2002]? Vagy a mai oly divatos „bagging” kérdése a gépi tanulásban. A bagging módszer alapján egy bootstrap alapú ismételt mintavételezés. De miért bootstrapet és nem keresztellenőrzést használunk? Bár ez utóbbira egyelőre még nem végeztem kutatást, de egy fordított kérdésről szól az utolsó fejezet, hogy a robusztussághoz miért keresztellenőrzést használunk és nem bootstrapet.

3. Kérdések a fordított Monte Carlo módszerrel kapcsolatban

3.1. A fordított Monte Carlo módszer egyértelműsége és sokasága

A fordított Monte Carlo módszer [McGreevy&Pusztai 1988] megjelenése után Pusztai László révén hamar a magyarországi folyadékszerkezeti kutatások egyik leggyakoribb témája lett. Vegyész szakdolgozatomban is ezzel a módszerrel foglalkoztam, egyetemi doktori értekezésemnek is egy részét ez adta. A módszer alapjai röviden a Függelékben találhatóak, ahol érdemes az ott szereplő hagyományos Monte Carlo szimulációval összehasonlítva értelmezni a fordított Monte Carlo módszert. Az RMC-vel kapcsolatos kutatások elég hamar két irányba ágaztak, kísérleti diffrakciós adatoknak az adott anyag tulajdonságaira fókuszáló kiértékelésére, illetve a módszer határainak elméleti/gyakorlati vizsgálatára. Több kérdés is felmerült:

- 1) A módszer egyváltozós függvényekre való illesztés során háromdimenziós szerkezeteket hoz létre, vagyis kevés információ alapján rengeteg koordinátát tartalmazó konfigurációkat ad. Ezek a konfigurációk a kétrészecske korreláción túlmutató szerkezeti függvényekkel is jellemezhetőek, pl. szögeloszlásokkal és szerkezeti ujjlenyomatokként értelmezhető invariánsaival a gömbi harmonikus függvényeknek. Kérdés volt, hogy milyen eltérések észlelhetőek, amikor a célzott „kísérleti” függvény klasszikus mechanikai szimulációkból származott, ahol a párkorreláción túlmutató függvények is rendelkezésre álltak referenciaként [pl. e1].
- 2) Többkomponensű rendszerek esetében hány és milyen koefficiensekkel képzett totál vagy parciális függvény alapján lehet érdemben modellezni?
- 3) Molekuláris és kovalens kötések tartalmaz rendszerekre mit jelent az egyértelműség?
- 4) Inverz (q) vagy valós (r) térbeli információra kell illeszteni és milyen intervallumban szükséges a függvények ismerete?
- 5) Mik a reális és beépíthető kényszerek?

A gyakorlati oldalról elindult a nem standard diffrakciós módszerek adatainak kiértékelése is (pl. elektrondiffrakcióra [McKenzie et al. 1992], EXAFS-ra [Bras et al. 1994] különböző összetett vegyületeken, akár nem is egyensúlyi, hanem üveg halmazállapotú rendszerekre [pl. Duine et al. 1994]. Az RMC módszer azóta rutineljárás, máig is aktívan használják, fejlesztik, rendszeres konferenciákat tartanak a témában. Az RMC akkori, a 90-es évek második felének főbb kérdéseivel kapcsolatban Robert McGreevy összefoglaló cikke jó kiindulási pont lehet [McGreevy 2001].

Új megközelítést jelentett 1993-tól a volt Ruff Imre csoportnál Baranyai András hazatérte, aki a több részecske eloszlásoknál a szög és a szférikus harmonikusokon felül a tripllett korrelációs függvények ($g_3, g(r,s,t)$) számolásában, illetve azok szerepében az entrópia sorfejtésben ($S_2, S_3, \dots$) is járatos volt. Ebben az időben, 1993 szeptemberétől – 1995. szeptemberéig, Németországban más

szimulációs témákon dolgoztam, de az ELTE-n Gereben Orsolya részvételével az RMC módszerrel foglalkoztak, beleértve S_2 és S_3 entrópia hozzájárulások, $g(r)$ (párkorrelációs függvény) és g_3 (triplett korrelációs függvény) számolásokat is, ez utóbbinál a fluktuációkat is nézték. RMC-vel különböző kényszerekkel készített amorf szilícium modelleket próbáltak minősíteni és jellemezni ezekkel az értékekkel [Gereben et al. 1994a, 1994b].

A fordított Monte Carlo módszer és a fluktuációk [s18]

1996-ban Baranyai Andrással újrakezdtük az egyértelműségi vizsgálatokat, de más megközelítésből [s18]. A korábbi szimulációs görbék (átlag $g(r)$, $S(q)$, szögeloszlások, szférikus harmonikusok rotációs invariánsai) összehasonlításán alapuló módszer helyett inkább a statisztikus mechanika elméleti szempontjait kezdtük el nézni, különösen a fluktuációkra koncentrálni. A vizsgált alapkérdés az volt, hogy az RMC által generált sokaság mennyire felel meg a kanonikus sokaságnak.

3.1. táblázat. A belső energia és szórásának összehasonlítása MC és különböző RMC szimulációkban. χ^2 a célfüggvény típusát jelöli az RMC szimulációkban. A 2-es hatványkitevő a mért és számított függvények eltérésének négyzetére utal az illeszkedési célfüggvényben. i index jelöli, ha r - vagy q -függő mérési hibát alkalmaztunk. [s18]

Módszer	χ^2	1/elf. arány	$\langle U \rangle / N$ J·10 ⁻²⁰	σ_U / N J·10 ⁻²⁰
MC		2,23	21,0233	0,1299
RMC	ΔS^2	2,23	20,9365	0,2259
	Δg^2	2,23	20,8532	0,2643
	ΔS_i^2	2,23	20,8713	0,2996
	Δg_i^2	2,23	20,9362	0,1054
	ΔS^2	5,00	21,0082	0,0748
	ΔS_i^2	5,00	20,9859	0,1318
	Δg_i^2	5,00	21,0010	0,0504
	ΔS^2	10,00	21,0148	0,0504
	Δg_i^2	10,00	21,0098	0,0416

Többnyire egykomponensű Lennard-Jones rendszereket vizsgáltunk referenciaként klasszikus Monte Carlo szimulációkkal, majd az annak eredményeképpen kapott $g(r)$ és $S(q)$ függvényekből kiindulva RMC modellezéseket végeztünk. Meghatároztuk az RMC szimulációk belső energiáját és annak fluktuációját is úgy, hogy RMC szimulációk konfigurációira az eredeti Lennard-Jones párpotenciállal

számoltunk. A hagyományos MC szimulációkban a maximális mozdítási hossz alapján beállított elfogadási aránya a generált konfigurációknak 0,5 körül szokott lenni, elvileg ez effektív bejárását jelenti a fázistérnek átlagos atomos rendszerekben. Az RMC szimulációkban ennél jóval kisebb elfogadási aránnyal szoktak dolgozni (0,05-0,1). Az eredeti MC szimuláció átlagos belső energiájánál az RMC szimulációk még a 0,1-es elfogadási arány esetén is nagyobb értéket adtak, de a belső energia fluktuációja, ami a rendszer hőkapacitásával hozható kapcsolatba, viszont ilyen elfogadási arány mellett rendkívül kicsi lett (3.1. táblázat).

A párkorrelációs és a szerkezeti függvény fluktuációja [s15] [s18]

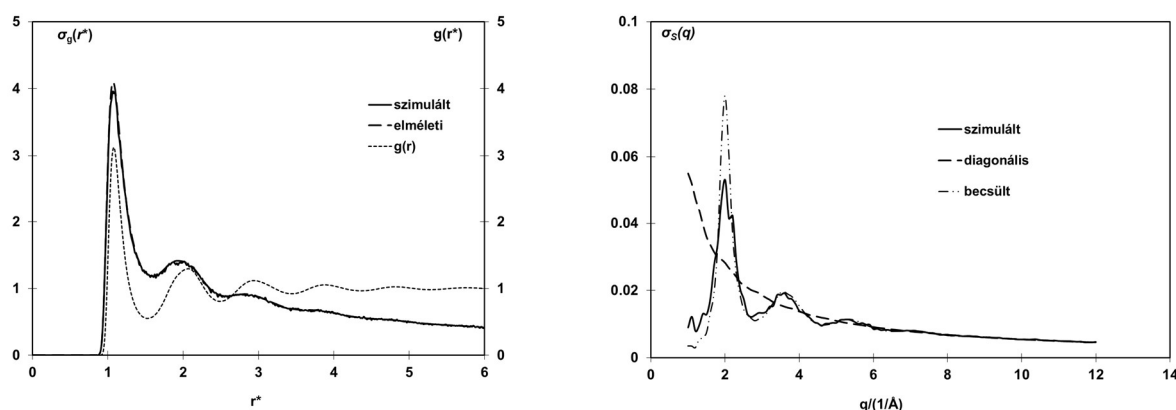
Az előző kutatásnak egy váratlan elágazása volt, hogy numerikusan és elméleti levezetéssel is be tudtuk mutatni, hogy kanonikus sokaságban a $g(r)$ függvény meghatározza a saját fluktuációját [s15]:

$$\sigma_g^2(r) = \frac{\langle g(r) \rangle}{4\pi r^2 dr} \quad 3.1. \text{ egyenlet}$$

Az analitikus forma létezésének sejtése úgy alakult ki, hogy EXCEL-ben közös ábrán nézegettem a $g(r)$ -t és a varianciáját, majd játékképpen megpróbáltam az egyikből előállítani a másikat, végül r^2 -tel való osztással és skálázó faktor EXCEL SOLVER-rel való meghatározásával egyértelmű kapcsolatot találtam (3.1. ábra). Az elméletet ezek után Baranyai Andrással közösen vezettük le, már tudva, hogy honnan hová kell eljutni. A képletből jól látható, hogy adott r távolságnál a fluktuáció nem függ más r távolságnál levő $g(r)$ értéktől. A 3.1. egyenlet egy részecskére adja meg a fluktuációt, az N részecskére és diszkrét felosztásra vonatkozó egyenlet [s18]-ban található. Az $S(q)$ esetére nem tudtunk zárt alakú képletet adni, de sikerült meghatároznom egy oszcilláció mentes diagonális tagot, ami $S(q)$ -val szorozva elég jól közelíti annak varianciáját (3.1. ábra). $v(r_j)$ az adott térfogatelemet jelenti.

$$\sigma_{S, \text{becslés}}^2(q_i) = \sigma_{S, \text{diag}}^2(q_i) S(q_i) = \rho \frac{2}{N} \sum_j v(r_j) g(r_j) \frac{\sin^2(q_i r_j)}{q_i^2 r_j^2} S(q_i) \quad 3.2. \text{ egyenlet}$$

Foglalkoztunk azzal is, hogy az $S(q)$ fluktuációjának a kiszámításához, vagy akár a belső energia varianciájának a kiszámításához közelebb visz-e minket a $g(r)$ varianciájának ismerete, de az ezekhez szükséges diszkrét $g(r)$ kovariancia mátrixára nem találtunk jó közelítő formulát. Egzakt megoldás feltehetően nem is létezik, mert abban az esetben a standard szimulációs eredmények nagy része megszerezhető lenne akár csak a $g(r)$ ismeretében, esetleg szimuláció nélkül is.



3.1. ábra. Jobb oldal: $g(r)$ és számolt, illetve elméletileg levezetett fluktuációja [s15]. Bal oldal: $S(q)$ számolt és közelítően becsült fluktuációja, valamint a diagonális rész [s18].

A fordított Monte Carlo módszer és a termodinamika [s23]

Visszatérve az RMC szimulációk minősítéséhez [s18], azt kaptuk, hogy ezeknél mind a $g(r)$, mind az $S(q)$ fluktuációja szignifikánsan kisebb, mint a kanonikus MC szimulációkban. Ugyanígy, a belső energia fluktuációja is kisebb, ha az eredeti Lennard-Jones kölcsönhatás paramétereivel számoljuk egy adott RMC konfiguráció potenciális energiáját. Vizsgáltuk, hogy a klasszikus $g(r)$ vagy $S(q)$ négyzetes eltéréssel definiált χ^2 -e helyett más definíció esetén (pl. negyedik hatvány, abszolútérték) hogyan teljesít a módszer, illetve kombináltuk a valós és a reciproktérbeli információt, valamint kiegészítettük a kétrészecske entrópiára vonatkozóval is (a részletek az s18-as publikáció 2-4. táblázataiban láthatóak).

Később beemeltük az összehasonlításba a háromrészecske eloszlások információit is a háromrészecske korrelációs függvények eltérésnégyzet-összegeinek illetve abszolút eltéréseinek formájában. Kiszámoltuk az RMC konfigurációkra a konfigurációs hőmérsékletet is. Az ekvipartíciós tételek közül a kinetikus energiára vonatkozót mindenki ismeri, vagyis a részecskék hőmérséklete az ideális gáz hőmérőnek megfelelően számolható a kinetikus energiából, ami szabadsági fokként $1/2 kT$. Kevésbé ismert a Clausius-féle ekvipartíció, ami a részecskére ható erőkire is hasonlóan állít. Ebből levezethető egy olyan szimulációkra alkalmazható összefüggés, ami pusztán a részecskék helyzete és a kölcsönhatási potenciál alapján definiál egy pillanatnyi konfigurációs hőmérsékletet. Szemléletesen úgy fogalmazhatunk, hogy hőmérsékletet számolunk annak alapján, hogy éppen milyen magasan és milyen meredek pozícióban helyezkednek el a részecskék a lokális potenciálfelületeiken. A 3.2. táblázatban jól látszik, hogy az RMC szimulációk különböző variánsai a kanonikus MC-nél magasabb konfigurációs hőmérsékleteket adnak. A következő táblázat az [s23] cikkből származik, és mutatja, milyen módon különböznek a MC és az RMC szimulációk termodinamikai jellegű adatai (3.2. táblázat):

3.2. táblázat. MC és RMC szimulációk termodinamikai adatainak összehasonlítása [s23]

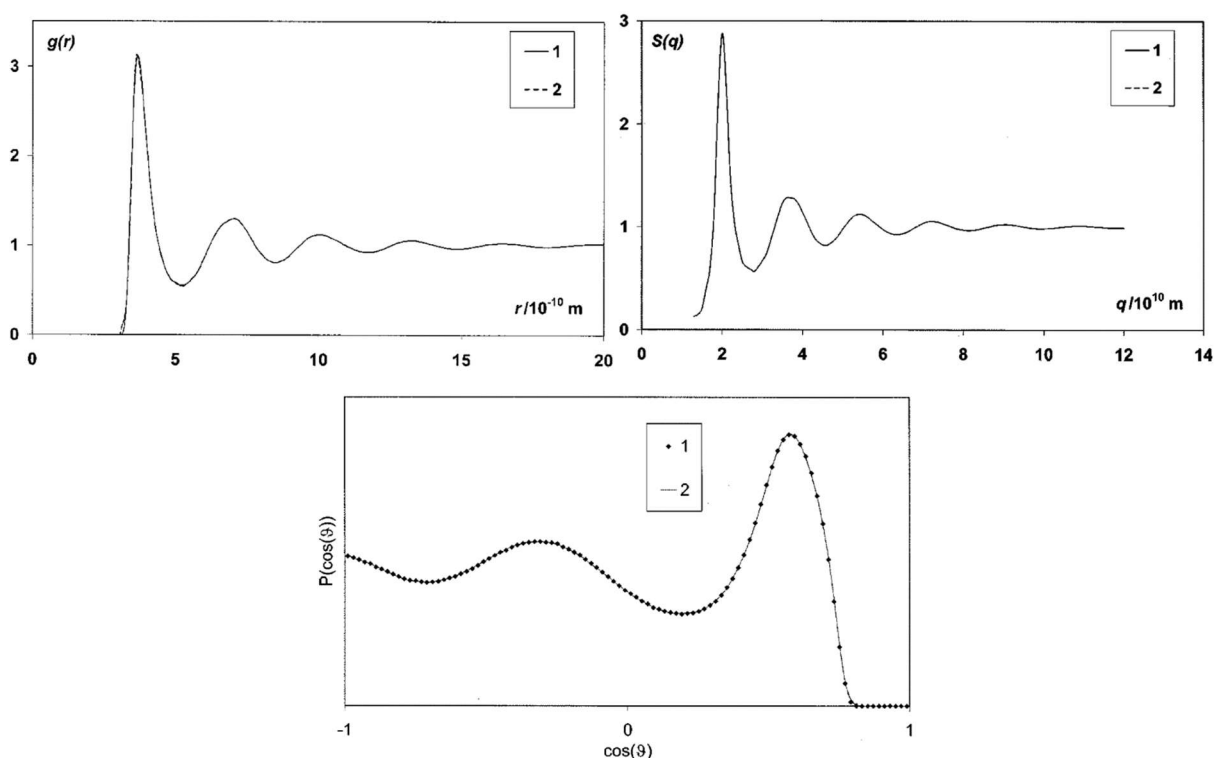
No.	szim.	cél-függvény ^a	teljes és részleges elfogadási arány	$E/$ atom kJ/mol	S_E kJ/mol	$D_1(r)$ $r=8.5 \times 10^{-10}$ m	$S_2(r)$ (k) $r=17.0 \times 10^{-10}$ m	$S_3(r)$ (k) $r=8.5 \times 10^{-10}$ m	$S_2(r)+S_3(r)$ (k) $r=8.5 \times 10^{-10}$ m	Konf. hőm. K
1	MC	-	0.45	-5.305	0.017	-	-3.79	-1.23	-4.76	86.3±3.0
2		-	0.45	-5.306	0.021	2.39X10 ³	-3.77	-1.22	-4.74	
3	RMC	ΔS^2	0.45	-4.708	0.041	1.09X10 ⁴	-3.55	-1.15	-4.44	601.8±55
4		ΔS^2	0.20	-5.194	0.017	3.72X10 ³	-3.77	-1.25	-4.73	228.7±33
5		ΔS^2	0.10	-5.245	0.012	3.26X10 ³	-3.79	-1.27	-4.78	159.8±24
6		Δg_i^2	0.45	-4.832	0.016	1.45X10 ⁵	-3.22	-0.82	-3.82	127.9±2.8
7		Δg_i^2	0.20	-5.172	0.008	1.35X10 ⁴	-3.65	-1.03	-4.41	100.0±1.9
8		Δg_i^4	0.20	-5.222	0.005	7.10X10 ³	-3.71	-1.09	-4.52	94.7±1.7
9		Δg_i^4	0.10	-5.250	0.004	4.12X10 ³	-3.75	-1.14	-4.61	93.1±1.7
10		$\Delta S^2, \Delta g_i^2$	0.20 (0.41, 0.41)	-5.253	0.005	3.33X10 ³	-3.75	-1.19	-4.67	97.2±1.9
11		$\Delta S^2, \Delta g_i^4$	0.20 (0.41, 0.41)	-5.271	0.004	3.29X10 ³	-3.79	-1.23	-4.73	95.2±1.7
12		$\Delta S^2, \Delta g_i^4$	0.10 (0.29, 0.29)	-5.286	0.003	2.75X10 ³	-3.80	-1.23	-4.75	92.3±1.9
13		$\Delta g_i^4, S_2$	0.20 (0.21, 0.85)	-5.224	0.005	8.82X10 ³	-3.75	-1.10	-4.56	95.0±1.7
14		$\Delta g_i^4, S_2$	0.10 (0.12, 0.75)	-5.255	0.004	4.83X10 ³	-3.81	-1.16	-4.67	93.1±1.6
15		$\Delta S^2, \Delta g_i^4, S_2$	0.20(0.46,0.46,0.74)	-5.272	0.004	3.20X10 ³	-3.79	-1.22	-4.73	95.0±1.8

a) S , g , és S_2 a szerkezeti függvényt, a párkorrelációs függvényt és az entrópia jellegű függvényt jelöli. A kitevők a célfüggvényben a hatványkitevőt jelentik. Az i index az r - és q -függő mérési hibahasználatát jelenti. Az entrópia járulékoknál a távolság az összegzés felső határát jelöli.

A táblázat mutatja, hogy az RMC konfigurációi távolról sem tekinthetők azonosan válogatottnak a kanonikus sokaságéval. A hagyományos statisztikai mechanikai képen a kanonikus sokaság energiaeioszlása úgy jön ki, hogy ha az x -tengelyen energiát szemléltetünk és a pl. bal oldali tengelyen a megfelelő Boltzmann faktort az adott konfigurációhoz, akkor erre egy $\exp(-x)$ -szel arányos lecsengő görbét, míg egy pl. ugyanezen az ábrán jobb oldali y -tengelyen az adott energiához tartozó lehetséges konfigurációk számát ábrázoljuk, akkor egy hasonlóan emelkedő görbét kapunk (lásd később 4.6. ábra). Az eredő energia eloszlása a konfigurációknak egy adott hőmérsékleten sematikus szintén berajzolható. Az 3.2. táblázatból látszik, hogy az MC konfigurációk energia eloszlása nem igaz az RMC esetében, hiszen jelentősen más átlagértékkel és szórással jellemezhetők az RMC eredmények. A táblázat azt is mutatja, hogy ha az RMC-t vezérlő $\Delta\chi^2$ értékét több és a konfigurációs tér elemeit egymástól eltérően súlyozó parciális célfüggvényeket alkalmazunk, pl. $g(r)$, $S(q)$ és kétrészecske entrópia, akkor az energetikai tagok valamint a konfigurációs hőmérséklete a rendszernek valamennyivel közelebb vihető az eredeti MC rendszeréhez.

A fordított molekuláris dinamika [s28]

Az RMC szimulációknak statisztikus mechanikai megközelítése generálta azt a kérdést, hogy az ergodikus hipotézisnek megfelelő klasszikus Monte Carlo és molekuláris dinamikai szimulációs párnak megvan-e a dinamikai változata fordított Monte Carlo módszer esetében. [s28]-as cikkünkben Baranyai Andrással levezettük a fordított molekuláris dinamikai (RMD) szimuláció képleteit és bemutattuk az RMD dinamikai szimuláció egyenértékűségét a sztochasztikus RMC-vel. A következő ábrákon látható, hogy az RMD esetében is tökéletes egyezés érhető el hagyományos MD szimuláció eredményeiből kapott célfüggvény esetében az RMD és az MD szerkezetre vonatkozó függvényeinél, egyszerű atomi rendszerekre (3.2. ábra).



3.2. ábra. MD és RMD szerkezetre vonatkozó függvények összehasonlítása. Az RMD szimulációk célfüggvénye az MD szimulációkból számolt $S(q)$ volt. Felül balra: párkorrelációs függvények Felül jobbra: szerkezeti függvények Alul: szögeloszlások - A szimulációk további részletei az eredeti [s28] cikkben találhatóak meg. [s28]

Egykomponensű rendszerek mellett klasszikusan modellezett víz RMD szimulációira is mutattunk példát. Olyan eseteket is bemutattunk, ahol kétkomponensű rendszerben kevesebb „mérési” görbével dolgoztunk, mint ahány parciális függvénye volt a rendszernek. Ezekben az esetekben a parciális $g(r)$ -ek reprodukálása ugyanúgy nem sikerült, mintha azt RMC szimulációval próbáltuk volna.

Ökölszabályok a fordított Monte Carlo szimulációhoz [s21] [s24]

A 3.1. alfejezet bevezetőjében említettem, hogy az RMC módszer egyértelműségére és alkalmazási hatáira többkomponensű vagy irányított (pl. kovalens vagy hidrogén-) kötésekkel is rendelkező rendszerekre számos kutatás folyt. Ezek egy részében „általánosnak” nevezett, sokszor hangzatos konklúziókra jutottak, kifejezetten sokszor figyelmen kívül hagyva más csoportok publikált eredményeit. Az egyetemi doktori dolgozatomban szereplő [s1-s3] cikkeim az első lépések közé tartoztak ebben a témában, így megszólítva éreztem magam minden alkalommal, amikor a témában kutató Pusztai László / Gereben Orsolya / Baranyai András és jómagam eredményeit nem vették figyelembe, vagy azokkal ellentétes következtetésekre jutottak. Egyik alkalommal olyan szinten, hogy az újságban „Comment” formájában vontunk kétségbe egy cikket, amelyben szerintünk elnagyolt állításokat tettek. A kommentünk szakmai újdonsága volt, hogy az RMC-vel kapcsolatos 1998-ig kialakult elméleti és gyakorlati véleményünket tézispontszerűen szedtük össze. [s21]

Nagyjából ugyanezeket a kérdéseket részletesebben tárgyaltuk Baranyai Andrással a Trends in Statistical Physics-ben megjelent rövid összefoglaló cikkünkben [s24]. Az összefoglalóban újdonságként szerepelt a szabadenergia és a $g(r)$ közti kapcsolat létezésének, mint egyértelmű funkcionálnak egy új elméleti levezetése (részletesen lásd később), valamint az RMD-vel kapcsolatban bemutatott egyértelműségi probléma, ha a parciálisok számánál kevesebb mérési adat áll rendelkezésre.

Ha mai szemmel nézek rá ezekre az állításainkra, úgy gondolom, hogy időtállóan mértük fel a módszer alkalmazhatóságát, előnyeit és korlátait. Általános szabályként a következőket állapítottuk meg:

- Ha lehetséges, minden esetben az elsődleges diffrakciós adatot (pl. $S(q)$) szükséges használni a származtatott függvénnyel (pl. $g(r)$ -rel) szemben.
- Fontos a merevgömbi paraméter korrekt használata, érdemes egy tartományban részletesen vizsgálni a hatását az eredményekre.
- Jelentős az eltérés a hagyományos MC és RMC szimulációk között a kapott szerkezeteket leíró két és teljes háromrészescke entrópius tagokban, különösen a teljes $g(r,s,t)$ esetében (pontos értelmezését lásd később a 4.4 alfejezetben a BGY elméletnél).
- Nagy rendszerméretet kell alkalmazni a Fourier transzformáció hibáinak csökkentésére.
- Az RMC jó szerkezeti eredményeket csak akkor szolgáltat, ha pl. a párkorrelációs függvény fluktuációját nagyon lecsökkentjük kis σ paraméter alkalmazásával. Ugyanez igaz pl. a becsülhető konfigurációs hőmérsékletre is. Ugyanakkor pl. a párkorrelációs függvény fluktuációjának az általunk levezetett elméleti r -függő értéket kellene adnia.

- A szabadenergia és a $g(r)$ közti funkcionál jellegű kapcsolat nem igazolása az RMC módszernek, hanem csak egy támogatása annak, hogy egyáltalán működhet a módszer. Nemegyensúlyi és amorf rendszerek modellezése ezért különösen fenntartásokkal értelmezhető.
- Hatékony kizárási módszer az RMC, ha a kérdéses tulajdonságra kényszereket alkalmazva nem illeszthető a diffrakciós adat.

A rövid összefoglaló cikkben az RMD-vel kapcsolatban bemutattunk néhány egyértelműségi szempontot, pl. ha a parciálisok számánál kevesebb mérési adat állt rendelkezésre, illetve érintettük a parciálisok súlyát a teljes $S(q)$ -kban. Ezekkel a kérdésekkel többször is szembesültem a 90-es években, de mivel kevés esetben dolgoztam valós diffrakciós adatokkal vagy molekuláris rendszerekkel, lehetőleg minden alkalommal legalább annyi független mérést vagy szimulált mérést bevontam az RMC jellegű számításokba, amennyi a parciálisok száma volt. Ettől az elvtől csak később tértem el, amikor a coarse graining kutatási irányba indultam el.

Ha összegeznem kell az ide kapcsolható cikkek [s15,s18,s21,s23,s24] lényegét, akkor a) megállapítottuk, hogy az RMC szimulációk által generált részecske elrendeződések nem tekinthetők egy kanonikus sokaság reprezentatív mintájának b) Egyes mutatók, mint pl. az energetikai és szerkezeti fluktuációk, a konfigurációs hőmérséklet és a 2-3 részecske entrópia hozzájárulások jobban hasonlónak tehetők az eredeti Monte Carlo szimulációk értékeihez, ha az RMC szimulációk során több célfüggvényt alkalmazunk, pl. párhuzamosan valós és inverz térbeli célfüggvényeket és S_2 tagot. c) Bemutattuk, hogy az RMC módszernek is létezik a dinamikai párja, az RMD. d) A kutatásokhoz kapcsolódóan lett néhány elméleti eredményünk is: sikerült bemutatni és elméletileg levezetni, hogy a párkorrelációs függvény fluktuációját egyértelműen meghatározza a $g(r)$ maga, míg az $S(q)$ esetében egy közelítő képletet tudtunk javasolni, valamint Baranyai Andrással közös cikkünkben szerepel az inverz elmélet egyértelműségének egy új bizonyítása is.

3.2. A részecskék belső és a részecskék közötti szerkezet párhuzamos meghatározása a fordított Monte Carlo módszerrel

Az RMC módszert a kezdeti időszakban atomi felbontású röntgen és neutron diffrakciós mérések kiértékelésére használták. Az ELTE-s vegyészképzés során egy sokunknak emlékezetes kolloidikai laboratóriumi mérés volt, amikor milliméterpapírra rajzolt Zimm-diagramm segítségével fényszóródási mérést értékeltünk ki. Doktori időszakom elején jött az ötletem, hogy a fényszórás mérés két fő célja, a kolloid részecskék közötti és részecskén belüli szerkezet felderítéséből az elsőre alkalmas lehet az RMC módszer. Néhány egyszerűbb gömbszimmetrikus kolloid rendszerre rendelkezésre állt az

irodalomban $S(q)$ függvény. 1992-ben Pusztai Lászlóval publikáltuk az első idevágó cikket [e3], ahol a kolloid részecskék konfigurációit állítottuk elő RMC szimuláció segítségével.

A polidiszperzitás szerepe a fényszóródási mérések fordított Monte Carlo kiértékelésénél [s16] [s22]

Az irodalomban számos intenzitás függvényt, $I(q)$ -t, publikáltak monodiszperz homogén gömbként modellezhető kolloid rendszerre fényszóródás mérés alapján. Mint az

$$\langle I(q) \rangle = n^2 f^2 N P(q) S(q) \quad 3.3. \text{ egyenlet}$$

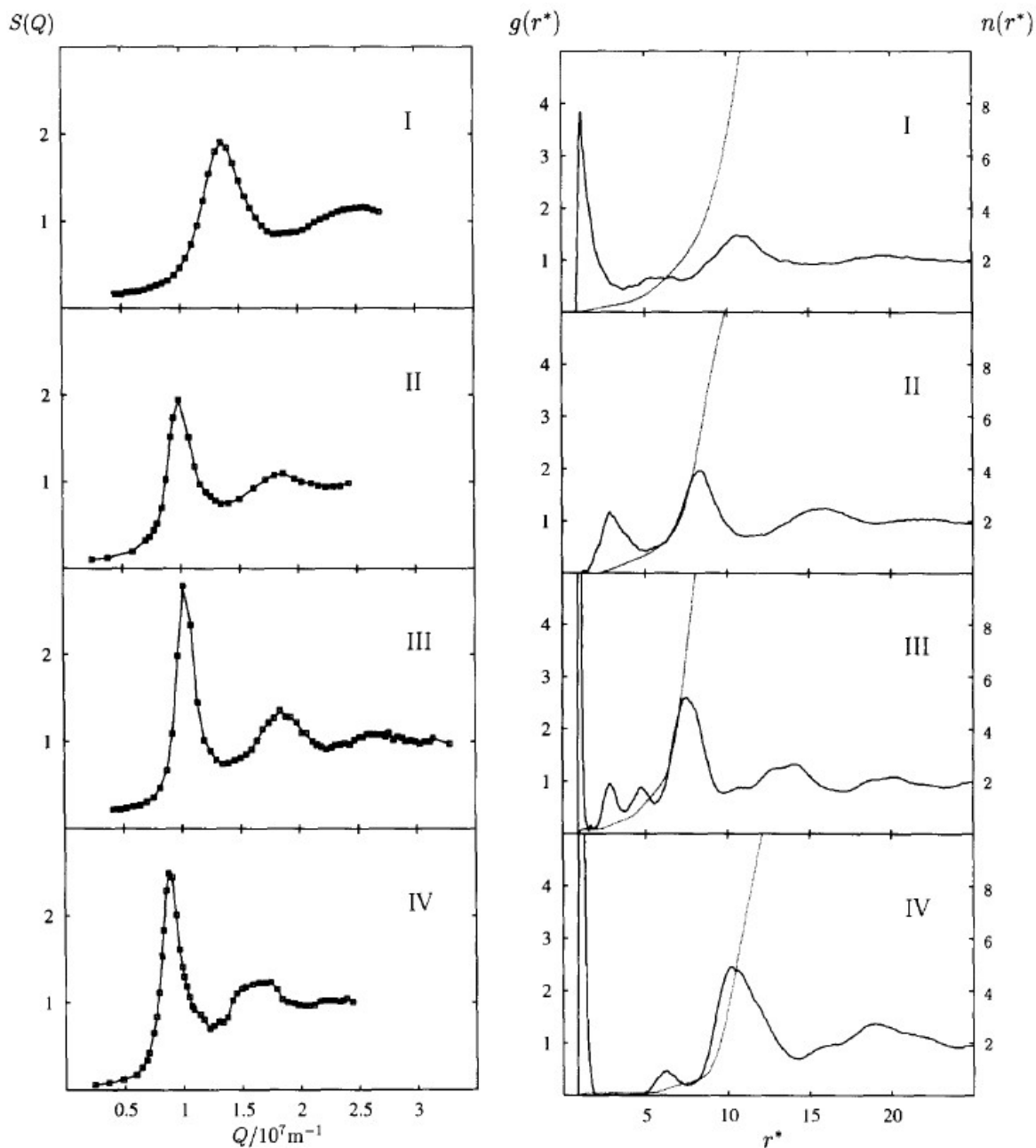
egyenletből látszik, a mért $I(q)$ intenzitás tartalmazza a $P(q)$ részecskeszóró függvényt, amely monodiszperz homogén gömbökre számítható. $\langle I(q) \rangle$ jelölés a függvény sokaságátlagát mutatja. n az f szórási amplitúdójú szórócentrumok száma egy részecskében, amelyből N darab σ átmérőjűt tartalmaz a rendszer.

$$P(q) = \left\{ 3 \left[\sin\left(\frac{q\sigma}{2}\right) - \left(\frac{q\sigma}{2}\right) \cos\left(\frac{q\sigma}{2}\right) \right] / \left(\frac{q\sigma}{2}\right)^3 \right\}^2 \quad 3.4. \text{ egyenlet}$$

A RMC háromdimenziós módszernek 1997-ben jelent meg az egydimenziós kistestvére, az MCGR módszer (Monte Carlo determination of $g(r)$) [Pusztai&McGreevy 1997]. Ez egy Monte Carlo alapú Fourier transzformációs eljárás, ahol a bemenő $S(q)$ függvényt irányítottan, de véletlen módosításokkal generált $g(r)$ függvényből lehet előállítani, elérve, hogy a Fourier transzformáció az RMC-hez hasonlóan a valós tér $\rightarrow$ inverz tér irányába történjen olyan függvényen, amely valós térbeli hossza miatt közel lecsengett, vagyis nem jelentkezik levágási hiba. Pl. Gereben et al. ezt a módszert használta korábban fémüvegek sűrűségének becslésére az RMC szimulációt megelőzően, mivel az adott rendszerek diffrakció szempontjából releváns sűrűségének mérése bizonytalan volt [Gereben&Pusztai 1996].

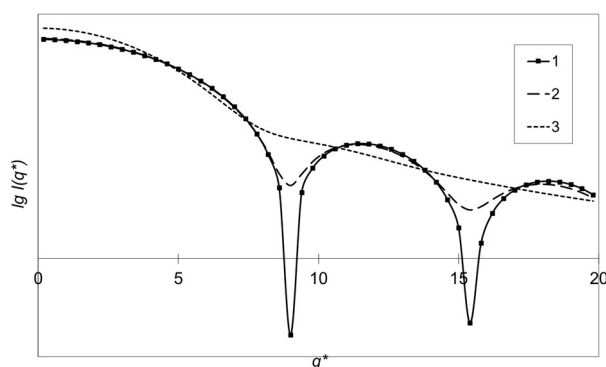
Az [s16]-os tanulmányban négy irodalmi $I(q)$ mérési görbét értékeltem ki, kipróbálva az eredeti szerzők által más mérések alapján javasolt részecske átmérőket és az MCGR-es legjobb illesztésekkel kapott saját átmérőket is, ahol az MCGR-ben végigszkenneltem egy-egy átmérő tartományt mindegyik méréshez. Az eljárás során a 3.3-3.4. egyenletek segítségével a mért $I(q)$ -ból adott részecskeátmérő mellett meghatároztam az $S(q)$ -t, majd arra alkalmaztam az MCGR eljárást. Az RMC szimulációkból kapott $g(r)$ és szögeloszlás görbék segítségével mutattam be a jellemző szerkezeteket (3.3.ábra). Viszonylag jelentős kontakt csúcsokat észleltem, illetve többször kisebb előcsúcsokat is. A kontakt értékek a szakirodalomban „deplation force” néven ismert jelenséggel értelmezhetők, ez gyakran a kolloid koaguláció egyik hajtóereje. A deplation force-nak nevezett jelenség az, hogy a kolloid részecskék közötti térből a statisztikus fluktuáció miatt alkalmanként kiszoruló oldószer és ellenion részecskék által keltett nyomáskomponens a két kolloid részecske közül eltűnik és a két részecske összetapad. Kinetikai szempontból elég ritkán követi ezt olyan fluktuáció, ahol valamiért szétválna a két részecske és megint bekerülne közéjük oldószer. Valós kolloid rendszereknél gyakran ez az összetapadás ad időbeli lehetőséget a két részecske végleges egyesülésére. Nem emeltem be a doktori

értekezésem cikkei közé, de volt egy kutatásunk [e35], ahol aszimmetrikus kétkomponensű rendszerek parciális $g(r)$ függvényeit határoztuk meg, olyanokat, ahol ezeknek a kontakt csúcsoknak az értéke akár 40-es volt a $g(r)$ -ben. Összehasonlításképp, az ideális gáz $g(r)$ -je konstans 1.



3.3. ábra. Négy gömbszimmetrikusnak feltételezett kolloid rendszerről készült fényszóródási mérés RMC szimulációs eredményei. Baloldal: mért és illesztett szerkezeti függvények. Jobb oldal: párkorrelációs függvények és azok integrálja (szomszédszám). r^* a távolság redukált egységben. Részletek az eredeti publikációban találhatóak. [s16]

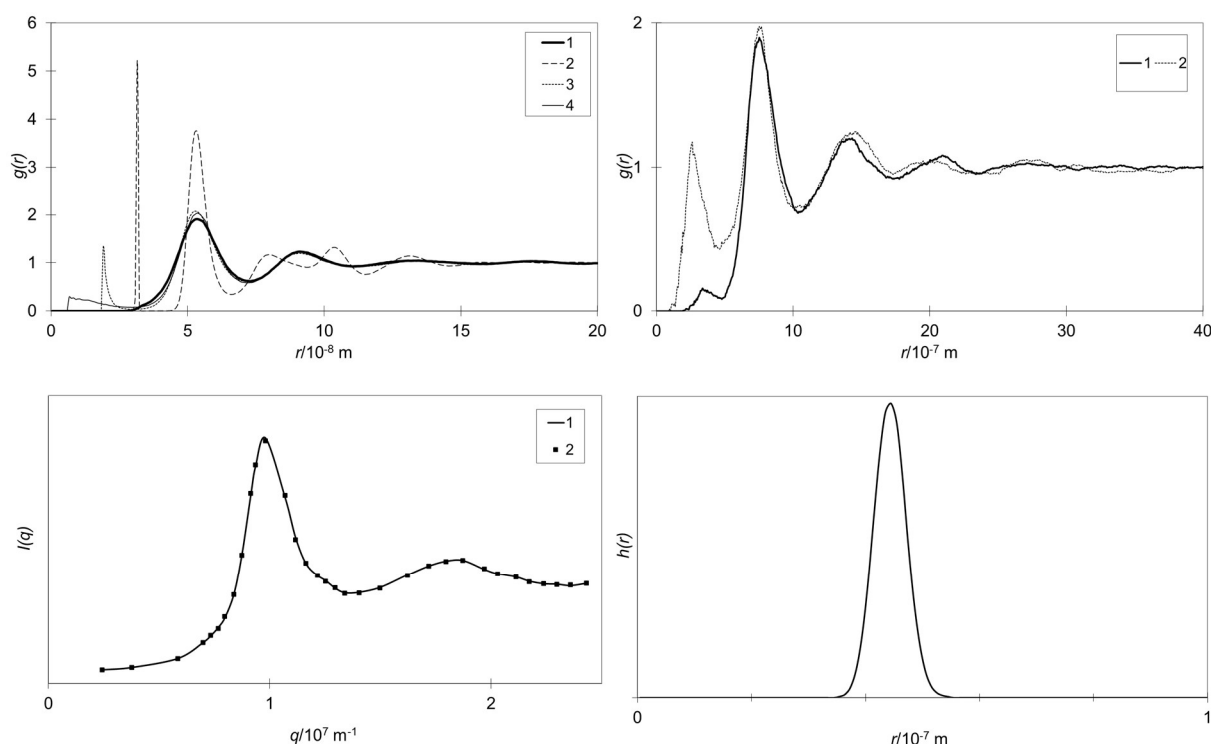
Az 3.3. ábrán szerepel a szomszédok r értékig integrált száma is, az $n(r)$, ami a párkorrelációs függvényből a gömbhéjakra vonatkozó térfogattal és az átlagos sűrűséggel való szorzással kapható. Látható, hogy a kontakt elrendezés kevés részecskét érint, az integrált szomszédszám legfeljebb egy tizednyi szomszédot jelent a III-IV. részában. Ugyanakkor a nagy csúcs előtti további csúcsok integrálja kb. egy (II-IV részában). Ez egy közeli szomszédot jelent, ami eléggé szokatlannak tűnt. Az eredmények ismeretében értettem meg, hogy a monodiszperz közelítés rendkívül karakterisztikus minimumokkal rendelkező $P(q)$ részecskeszóró függvényt ad. Azonban kis polidiszperzitást megengedve $P(q)$ kisimul (3.4. ábra).



3.4. ábra. Ideálisan híg merevgömbi rendszerek szórási intenzitása redukált q egységben. 1: monodiszperz gömbök 2-3: polidiszperz gömbök, ahol a koloidok polidiszperzitására jellemző Schultze-eloszlás félértékszélessége 2: 5%, illetve 3: 20%. [s22]

Az [s22] és [s25] publikációkban egy olyan RMC variáns kifejlesztéséről számoltam be, ahol a hagyományos részecske elmozdítással párhuzamosan az adott részecske átmérőjét is változtattam. Az így számolt $I(q)$ értéket vettem össze a megszokott χ^2 kritériummal a fordított Monte Carlo eljárás során. Mivel előfordulhatott volna, hogy egy-egy részecske sugara nagyon eltérhet a többitől, akár olyan mértékben, hogy a részecskeszóró függvénye karakterisztikus értékei az adott mérési q tartományon is kívülre essen, ezért egy további feltételt is alkalmaztam. Közel monodiszperz koloid részecskék méreteloszlása jól közelíthető a Schulz elméleti eloszlással. A részecskepozícióktól függő χ^2 -t kiegészítettem egy olyan χ^2 értékkel, ami azt mutatta, hogy a szimulációban levő méreteloszlás mennyire írható le Schulz eloszlással. Az elméleti Schulz eloszlás várható értéke és szórása megegyezett az adott konfiguráció részecske sugarainak várható értékével és szórásával. A kombinált RMC módszerrel kapott párkorrelációs függvényekről eltűntek az előző kiértékeléseknél kapott kontakt párok, illetve az előcsúcsok is vagy eltűntek, vagy jelentősen kisebbek lettek. A 3.5. ábra mutatja a párkorrelációs függvényt két rendszerre, illetve az egyikre a modellezett $I(q)$ függvényt és a kapott

részecskesugár-eloszlást is az ábra alsó sorában. A rendszerek térkitöltésének definíciója: a szimulációs cellában levő gömbök térfogatának összege osztva a szimulációs cella térfogatával.



3.5. ábra. Polisztrén latex gömböket tartalmazó rendszerekről készült fényszóródási mérések kiértékelése RMC módszerrel. Felső ábrák: párkorrelációs függvények. Bal felső ábra: 0,13-as térkitöltésű rendszer. 1: polidiszperz közelítés a részecske méreteloszlás párhuzamos optimálásával 2-4: monodiszperz közelítés $1,56 \cdot 10^{-8} \text{ m}$, $0,94 \cdot 10^{-8} \text{ m}$, $0,312 \cdot 10^{-8} \text{ m}$ merevgömbi részecskesugárral. Jobb felső ábra: 0,0211-as térkitöltésű rendszer. 1: polidiszperz közelítés a részecske méreteloszlás párhuzamos optimálásával 2: monodiszperz közelítés $4,55 \cdot 10^{-8} \text{ m}$ merevgömbi részecskesugárral. Bal alsó ábra: a 0,0211-as térkitöltésű rendszer mért és illesztett $I(q)$ -ja. Jobb alsó ábra: részecskeeloszlás 0,0211-as térkitöltésű rendszerben. További részletek az eredeti publikációkban találhatóak. [s22, s25]

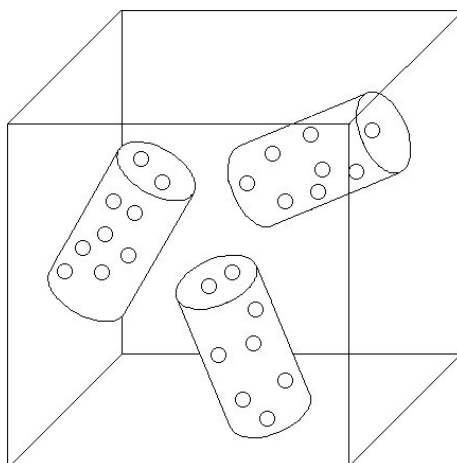
Ni katalizátor szerkezetének vizsgálata fordított Monte Carlo szimulációval [s30]

A kolloidok fényszóródási mérései körülbelül ugyanarról a távolságtartományról szolgáltatnak információt, mint a kiszögű röntgen- és neutroindiffrakciós mérések. Bóta Attila kollégám jelezte, hogy vannak ilyen mérési eredményei pár érdekes rendszerről. Végül egy nem pirofóros Raney-nikkel típusú katalizátor szerkezetéről próbáltunk meg információt kapni módosított RMC algoritmus segítségével [Bóta et al. 2002, s30].

A porózus alumínium tartalmú hordozóanyag gibbsit és bayerit alapú volt, ebben kb. 25-30 tömeg %-ban Ni volt eloszlva, a különböző mérések alapján kis hengerek formájában. A nikkel részecskékről

való kisszögű diffrakciós kép anomális röntgenszórás segítségével lett meghatározva. Egy a nikkel abszorpciós csúcsához közeli és egy távolabbi energiájú röntgen sugárzással készített spektrum különbsége közelíthető úgy, mint ha csak a nikkel részecskékről készítettünk volna felvételt.

A fő kérdés az volt, hogyan számítsuk a hengerek részecszeszóró függvényét, illetve mekkora és milyen sugár/magasság arányú hengereket használjunk. De voltak további kérdések is, pl. nem ismertük a pontos sűrűséget, így az intenzitás skálázása és alapvonal eltolása is szükséges volt. Ez a modellezés egy olyan tudományos ötletelés volt, ahol a hengereknek az RMC szimulációban a megszokott translálása és forgatása mellett a többi feladatot is meg kellett oldani. A végső modell a 3.6. ábrán látszik, a hengerek atomi méretű nikkel részecskékként lettek feltöltve, de nem a valós sűrűséggel, mert akkor milliós nagyságrendű nikkel atomról lett volna szó.



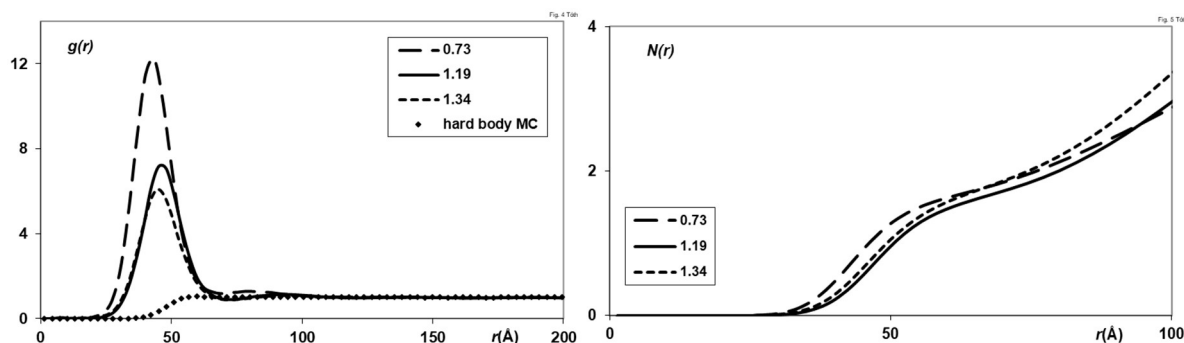
3.6. ábra A szimulációs cellába helyezett nikkel hengerek, és az azokba véletlenszerűen elhelyezett Ni szórócentrumok sematikus képe. [s30]

Kihasználva, hogy a kisszögű tartományban az atomi szórási tényező megközelítőleg konstans, valamint bevezetve a c_1 és c_2 skálázó tényezőket, az intenzitás számolható a hengerekbe rakott próba részecskék pozíciójából, illetve a minden egyes számoláskor egy-egy dimenziós minimumkereséssel meghatározott c_1 - c_2 paraméterekből.

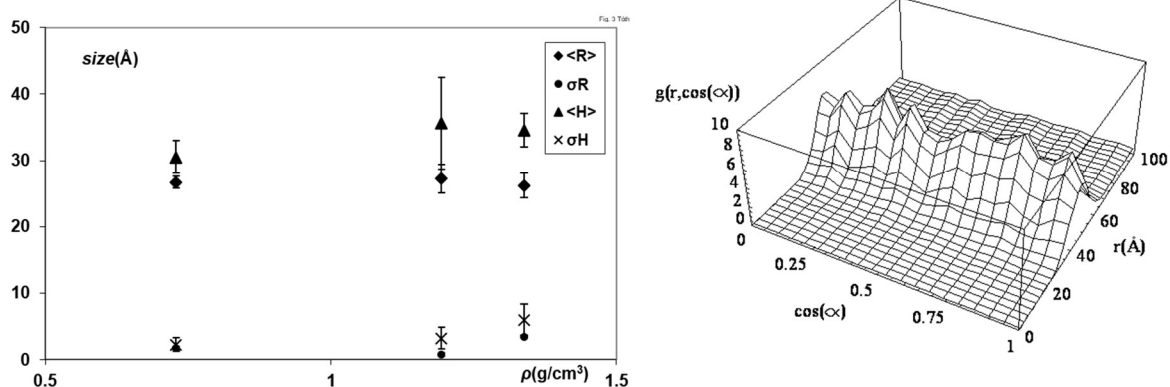
$$I(q) = c_1 \cdot S(q) + c_2 \quad \text{3.5. egyenlet}$$

A szimuláció során 80 hengert helyeztünk el a periodikus szimulációs cellában és ezekbe összesen 2000-4000 Ni szórócentrumot helyeztünk el véletlenszerűen. A hengerek átlagos hossza, annak szórása a 80 hengerre számítva, valamint a sugaruk és annak szórása négy extra változót jelentett, amelyeket szintén RMC eljárással változtattunk. Vagyis a hagyományos translálási és forgatási lépések mellett néha próbálkoztunk olyan Monte Carlo lépésekkel is, ahol ezeket a paramétereket változtattuk. A hengerek méretének változásával a rendszer sűrűsége is változott. Ez szándékosan nem lett korigálva,

mert nem volt pontos mérési eredmény arról, hogy mekkora a helyi nikkel sűrűség a katalizátorban. Több sűrűség értékből több szimulációt indítottunk abban bízva, hogy lesz olyan sűrűség, ahol kimagaslóan jól sikerül reprodukálni a mért kísérleti görbét. Végül nem találtunk ilyen értéket, de meglepő módon a hengerek átlagos paraméterei, a párkorrelációs, integrált párkorrelációs és orientációs függvények nagyon hasonlóak lettek. Az eredményeket a következő ábrákon mutatom be (3.7. és 3.8.ábrák):



3.7. ábra. Bal oldal: Ni részecskék centrumának párkorrelációs függvénye három sűrűsége (g/cm^3 egységben) és összehasonlítása merev test (hard body) szimulációval. Jobb oldal: az előbbi görbék szomszédszám integrálja annak szemléltetésére, hogy mindegyik szimuláció hasonló részecskeszámmra vonatkozó sűrűséget jelentett. [s30]

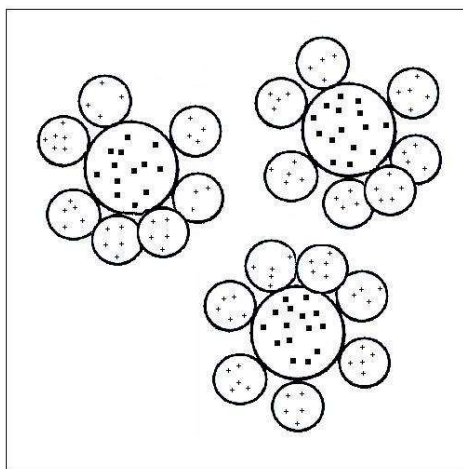


3.8. ábra. Bal oldal: a Ni hengerek sugarai és magassága (átlaguk, szórásuk és ingadozásuk a szimulációk alatt). Jobb oldal: A hengerek egymással bezárt szögének az eloszlása a távolságuk függvényében az $1,19 \text{ g/dm}^3$ -as sűrűségű szimulációban. Az ekvidisztáns felosztás a szögek koszinusza szerint történt, mivel ahhoz tartozik egyforma térfogatelem a térben. [s30]

Egy nemionos felületaktív anyag szerkezetének vizsgálata fordított Monte Carlo szimulációval [s33]

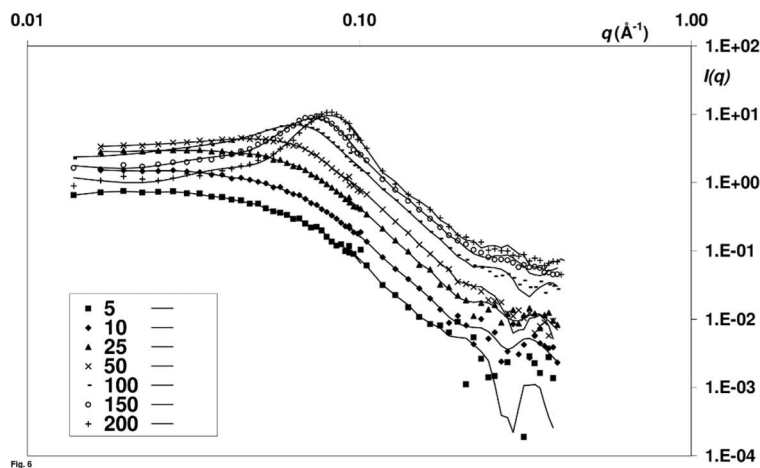
A nikkeles RMC szimuláció példájából kiindulva 2004-ben egy Borbély Sándor által mért adatsor kiértékelésébe is belekezdünk. A $\text{CH}_3(\text{CH}_2)_{11}(\text{OCH}_2\text{CH}_2)_{23}\text{OH}$ nemionos felületaktív anyag kommerciális neve BRIJ-35. Az anyag micelláris szerkezetével kapcsolatban több feltevés volt abban az időben, az egyik érdekes közelítés a „hajas golyó”, vagy angolul hairy sphere elnevezésű volt. Szemléletesen úgy ábrázolták ezeket a micellákat, hogy egy viszonylag nagy, kb. 40 molekulából képződő aggregátumokból álló hidrofób belső mag és az azt körülvevő hidrofil lelógó molekula részekből áll a micella. Borbély Sándor kisszögű neutroddiffrakcióval vizsgálta ezt a rendszert 5-200 g/dm³ koncentráció tartományban 20, 40 és 60 °C hőmérsékleteken [Borbély 2000].

Az RMC modellünk az 3.9. ábrán látható, ahol a hidrofil magot egy nagy gömb, míg minden egyes molekulához az összegubancolódott hidrofób láncot egy-egy kis gömb tartalmazta. A belső és a külső gömböket szórócentrumokkal töltöttük fel, két különböző fajtaival. Ennek megfelelően kétkomponensű rendszerként modelleztük a rendszert, ami három parciális párkorrelációs függvény és egy konstans tag hozzájárulását adta a számolt $I(q)$ intenzitáshoz (lásd a_1 - a_4 paramétereket az eredeti közleményben, [s33]). Ezek a paraméterek a parciálisok súlyozása mellett a normálás feladatát is elvégezték, most is egyváltozós minimumkereséssel optimáltuk ezeket minden RMC próbálkozásnál. A hagyományos egy-egy micellára vonatkozó translációs és rotációs lépésen felül alkalmanként RMC analóg módon (vagyis $\Delta\chi^2$ alapon) változtattuk az átlagos aggregációs számot (N_{agg}), annak szórását ($\sigma_{N_{agg}}$) és a külső gömbök nedvesedési faktorát (W), ami azt mutatta, hogy hányszorosára dagad a gömb elvi molekuláris mérete az oldószer felvétele miatt. A belső gömb méretét az aggregációs szám határozta meg a megfelelő hidrofób BRIJ-35 molekularészletek térigényével arányosan. A szimulációs cellák 200 micellát tartalmaztak és összesen 35000-50000 szórócentrumot.

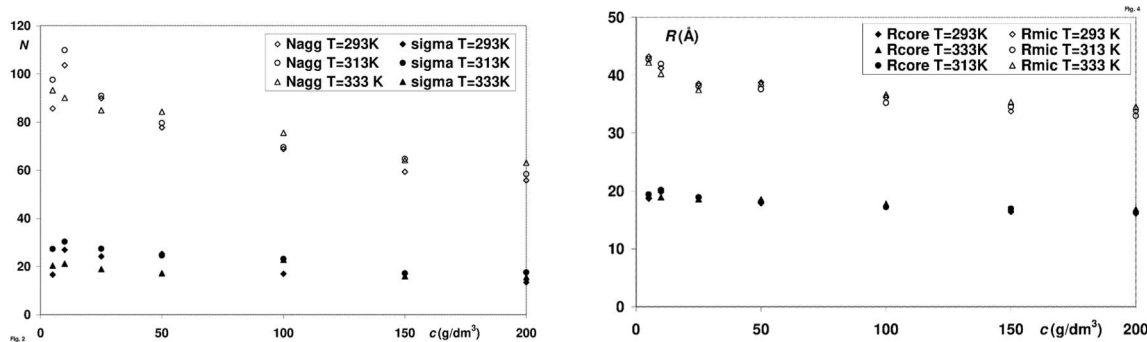


3.9. ábra. A BRIJ-35 anyag szimulációjakor használt modell. [s33]

A szimulációkkal jól reprodukáltuk a kísérleti eredményeket, valamint eredményeket kaptunk N_{agg} értékére és szórására, valamint a micellák méretére (3.10. és 3.11. ábrák).

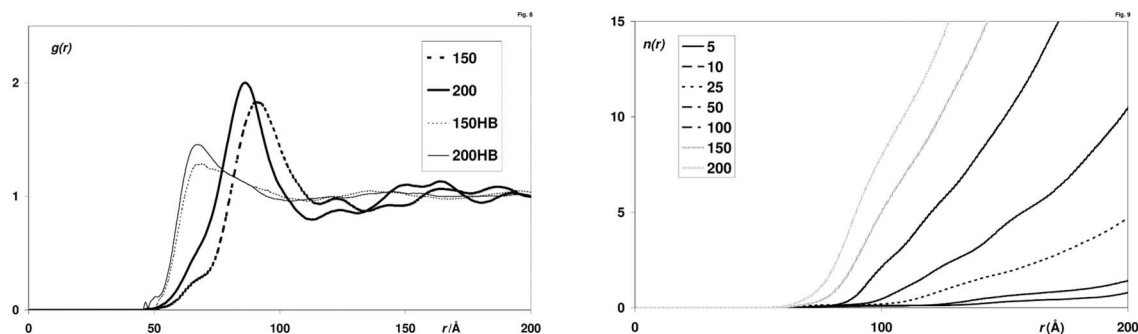


3.10. ábra. BRJ-35 szórási intenzitások reprodukciója a szimulációkban különböző koncentrációknál (g/dm^3 egységben). [s33]



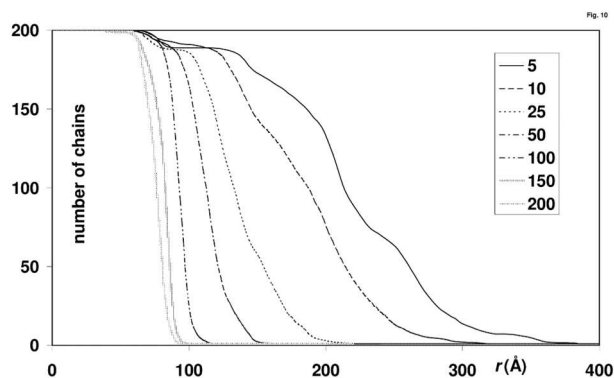
3.11. ábra. BRJ-35 mérés RMC kiértékeléskor kapott értékei különböző koncentrációknál: Bal oldal: aggregációs számok és szórásuk Jobb oldal: A mag és a teljes micellák becsült sugara. [s33].

A W nedvesedési faktorra reális értékeket kaptunk, 1.7-5-ös tartományban a koncentrációtól és a hőmérséklettől függően. A szimulációkból a részecskék közötti szerkezetről is információkat szereztünk (3.12. ábra).



3.12. ábra. A BRIJ-35 mérés RMC kiértékeléskor kapott párkorrelációs függvényei (bal oldal) és a részecskeszámra kiintegrált értékek (jobb oldal). A bal oldali ábrán a HB-s görbék merev testekre végzett szimulációkat jelentenek diffrakciós célfüggvény illesztése nélkül. [s33].

A legutolsó ábra a rendszerek perkolációját mutatja a különböző koncentrációkra. Esetünkben a perkoláció azt mutatja, hogy amennyiben két szomszédos BRIJ-35 micellát összekapcsolnak tekintünk egy adott távolságon belül, akkor hány klaszter található a rendszerben. Az ábra ezeket az értékeket az összekapcsoló határtávolság függvényében ábrázolja (3.13. ábra).



3.13. ábra A BRIJ-35 mérési eredmények kiértékeléskor kapott perkolációs görbék. [s33].

Az [s34]-s áttekintő publikációm alapján új eredményeket nem tartalmazott, de jól mutatta az [s3]-[s16]-[s22]-[s25]-[s30]-[s33] munkák folyamán nyert tapasztalatot. Például a polidiszperzitás figyelembevételét elengedhetetlennek tartom ma már. A két különleges alakú rendszer, a nikkel katalizátor és a BRIJ-35 pedig egy olyan megoldást mutatott a korábban kezelhetetlen térbeli alakzatok részecske szóró függvényének közelítésére, ami könnyen beilleszthető volt az RMC kiértékelési sémába. Ezek munkák minden esetben lehetővé tették olyan információk kinyerését diffrakciós

toth.gergely_304_24

adatokból, különösen a részecskék közötti szerkezettel kapcsolatban, amelyek más módszerekkel addig nem voltak lehetségesek. A munkák utóéletét a hivatkozások mutatják, ezen a területen főleg japán kisszögű diffrakciós mérések kiértékelésével foglalkozó kollégák hivatkoznak rám, illetve a BRIJ-35 esetében annak az anyagcsoportnak a kutatói említenek meg.

4. Kölcsönhatások, a statisztikus mechanikai inverz probléma, „coarse graining”

4.1 Potenciális energia felület meghatározása kvantumkémiai, klasszikus mechanikai

kölcsönhatások paraméterezése

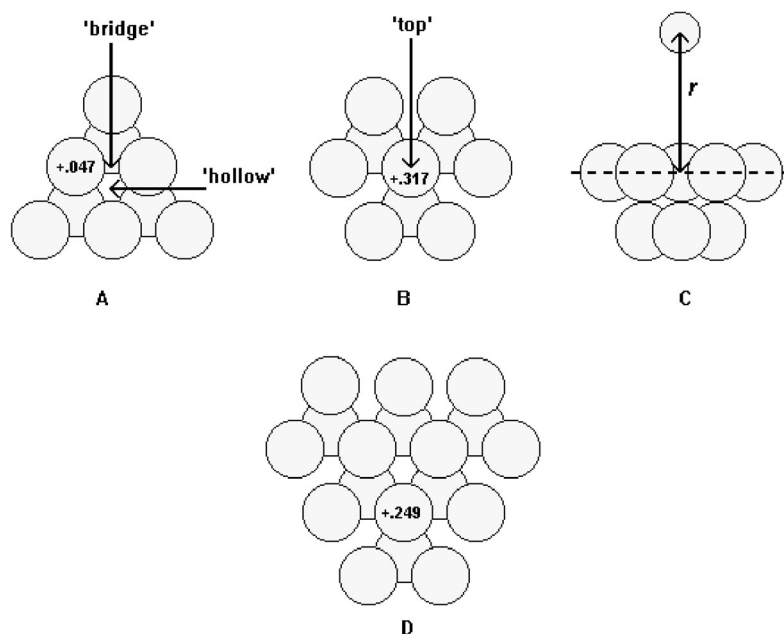
Ebben az alfejezetben azok a 90-es években folytatott kutatásaim szerepelnek, ahol kvantumkémiai módszereket használtam alapvetően két megközelítéssel. Két esetben a klasszikus mechanikai szimulációkhoz szükséges kölcsönhatási potenciálok kvantumkémiai meghatározására mutatok példát, a másik két eset pedig olyan módszerek fejlesztéséről vagy alkalmazásáról szól, ahol a szimulált rendszer energiájának vagy a részecskékre ható erőknek a számolása menet közben („on fly”) kvantumkémiai módszerekkel történt.

Hg-ion kölcsönhatási potenciálok paraméterezése [s10]

Karl Heinzinger mainzi kutatócsoportjában a 90-es években a határfelületi szimulációk jelentették a fő kutatási irányt. A szimulációkhoz különböző klasszikus mechanikai kölcsönhatási modelleket használtak. Az első modellek az úgynevezett tükörtöltés (image charge) alapján lettek meghatározva. Ez az elmélet jól leírja az ion-fémes vezető közti mikroszkopikus párkölcsönhatást közepes és nagy távolságoknál, de nem tudja figyelembe venni a felületi inhomogenitást és az ott levő eltérő kötési helyek különbségeit. Az elmélet kiegészíthető pl. Lennard-Jones kölcsönhatásokkal a kötőhelyek megkülönböztetésére, de ez csak egy a térbeli hatást leíró korrekció, nem veszi figyelembe pl. az ionok és a felület közti hatás következtében az elektronfelhőben bekövetkező változásokat. A 90-es évek elején kezdett elterjedni, hogy kvantumkémiai sorozatszámolások alapján határozzanak meg klasszikus mechanikai kölcsönhatási potenciálokat, mi is ezt az utat választottuk. Az adott potenciál kifejlesztése során az egyik első kérdés akkor is az volt, hogy milyen kvantumkémiai szinten, mekkora rendszermérettel és milyen korrekciókat végezve határozzuk meg a potenciált. A 90-es évek elejének számítástechnikai kapacitása mai szemmel nagyon korlátozott volt. A potenciálisenergia-felület meghatározása után a megfelelő, többnyire csak párkölcsönhatási tagokat tartalmazó egyenlet paraméterezése már jóval egyszerűbb feladat volt.

A kvantumkémiai sorozatszámítási lehetőségek 9 és 10 higany atomot tartalmazó klaszterekből és egy ionból álló rendszereket tettek lehetővé Hartree-Fock szinten közepes, diffúz és polarizációs bázis függvényekre részben általam optimalt bázisok mellett [s10, e11]. A belső elektronhéjakat pszeudopotenciálok segítségével írtuk le, részben a relativisztikus hatások, részben a kisebb számításigény miatt. A véges méret miatt egyértelműen nem beszélhettünk fémes Hg klaszterekről. Többfajta korrekciót is alkalmaztunk vagy vizsgáltunk, pl. a báziskészlet szuperpozíciós hibát, a véges méret miatt a higany klasztereken kialakuló egyenletlen töltéseloszlás korrigálását, illetve ellenőrző

számolásokat végeztünk 18 atomos klasztereken is. Az 4.1. ábrán a vizsgált szerkezetek és kötési pozíciók láthatóak.



4.1. ábra: A kvantumkémiai számolások során használt Hg klaszterek sematikus ábrája. Az ábrán az ionkötő pozícióknak az angol szakirodalomban ismert nevei szerepelnek. Magyarul inkább az egy-, kettő, három- atommal való kötés szerinti megnevezés a pontosabb, mint a pontatlan tető, híd és üreg fordítások. [s10]

A kapott potenciálisenergia-felületekre a klasszikus két részecske távolságától ($r_{\text{Hg-ion}}$) függő potenciál tartalmazott még egy tagot (ρ), amely az ionnak a felületre projektált helyzete alapján módosította a kölcsönhatást (4.1. egyenlet). A potenciálparaméterek optimálásához Gauss-Newton-Marquardt-Levenberg algoritmus alapján C programot írtam és minden esetben több száz különböző kezdeti becslésből indítottam a számolást.

$$U(r, \rho) = p_2 \exp(-p_3 r) - p_4 \exp(-p_5 r) + f(\rho) p_6 \exp(-p_7 r)$$

$$f(\rho) = \exp(-p_1 \rho^a)$$

$$a = (1,52718 - \log p_1) / 0,633954 \quad 4.1. \text{ egyenlet}$$

Az optimált paraméterek az eredeti publikációban találhatóak mind a négy anionra (F-, Cl-, Br-, I-) és az 5 kationra (Li+, Na+, K+, Rb+ és Cs+).

Ion-víz kölcsönhatások paraméterezése [s14]

A fém/elektrolit határfelületi molekuláris dinamikai szimulációkhoz fém-fém, fém-víz, víz-víz, ion-víz és ion-ion kölcsönhatásokra van szükség. Ha olyan kis elektrolitkoncentrációban szimulálunk, ahol egy ion

található a rendszerünkben ellenion nélkül, akkor az ion-ion kölcsönhatásra nincs szükség. Ebben az esetben a rendszer neutralitását egy szétkent háttértöltéssel lehet korrigálni, ami a klasszikus dinamikai képben nem jelenik meg a mozgásegyenletekben. Az általunk szimulált teljes alkáli-halogenid sorhoz a csoport által használt BJH vízmodellnél nem volt szisztematikusan fejlesztett ion-víz kölcsönhatási potenciál, ezért a mainzi időm végén ennek a fejlesztésébe is belekezdtem [s14]. Iononként 420 különböző ion - egy víz molekula szerkezetekre számoltam kvantumkémiával és 'counterpoise' korrekcióval kötési energiákat, MP2-MP4 szinten, közepes bázisokat használva. A számolásokat elvégeztem mind a csoport által használt BJH, mind az akkor egyik legismertebbnek számító TIP4P vízmodellre. Rendkívül flexibilis potenciál formákat illesztettem párkölcsönhatást feltételezve külön-külön a víz atomjaira. Az anion-víz kölcsönhatásra:

$$U_{io}(r_{io}, r_{im}) = \frac{A_1}{r_{im}} + \frac{A_2}{r_{io}^4} + \frac{A_3}{r_{io}^6} + A_4 \exp(-A_5 r_{io})$$

$$U_{ih}(r_{ih}) = \frac{B_1}{r_{ih}} + \frac{B_2}{r_{ih}^2} + \frac{B_3}{r_{ih}^4} + B_4 \exp(-B_5 r_{ih}) - B_6 \exp(-B_7 r_{ih})$$

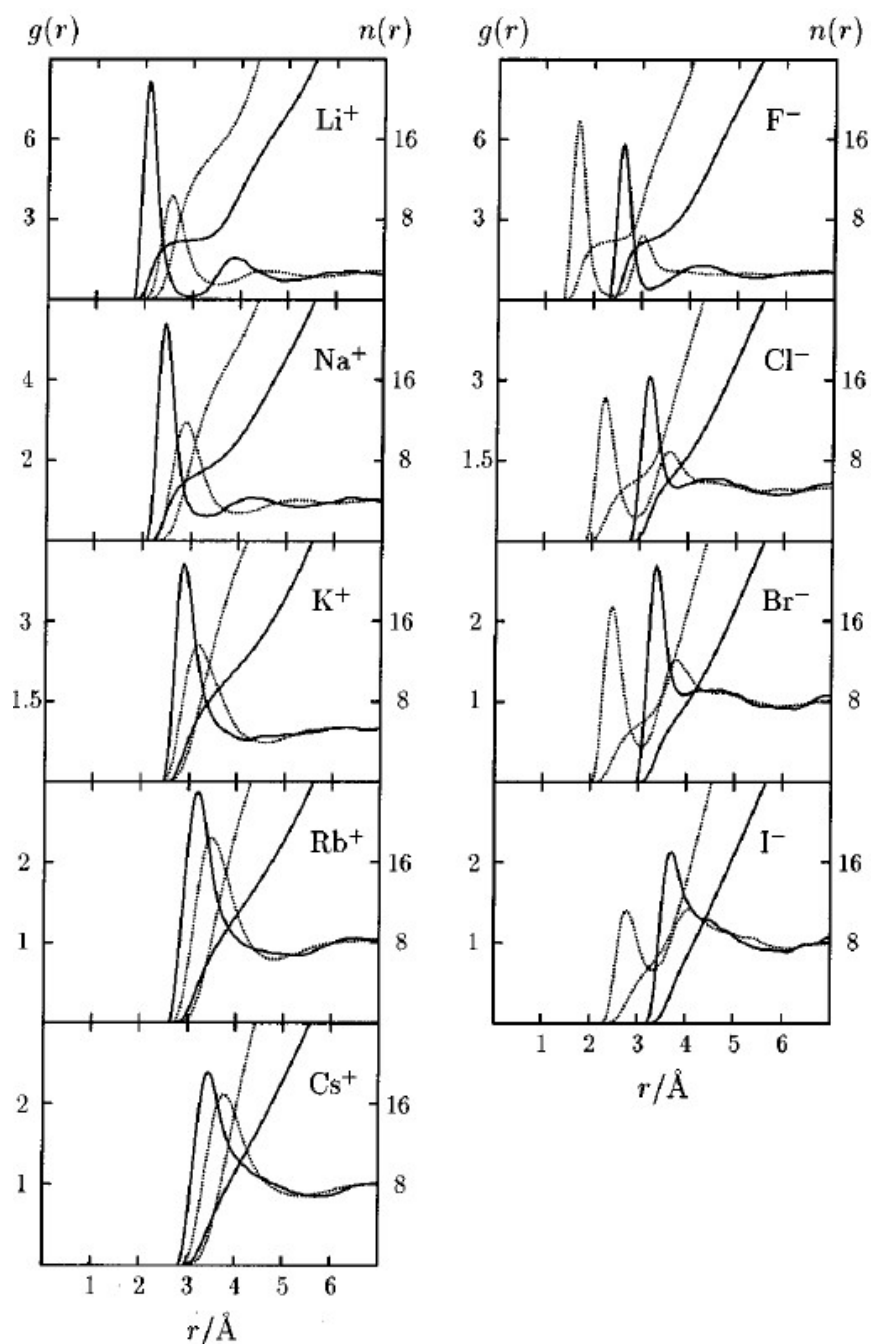
, illetve a kation-víz kölcsönhatásra:

$$U_{io}(r_{io}, r_{im}) = \frac{A_1}{r_{im}} + \frac{A_2}{r_{io}^4} + \frac{A_3}{r_{io}^6} + A_4 \exp(-A_5 r_{io}) - A_6 \exp(-A_7 r_{io})$$

$$U_{ih}(r_{ih}) = \frac{B_1}{r_{ih}} + \frac{B_2}{r_{ih}^2} + \frac{B_3}{r_{ih}^4} + B_4 \exp(-B_5 r_{ih}) \quad 4.2. \text{ egyenlet}$$

, ahol r_{io} és r_{ih} az ion oxigén és ion hidrogén távolságot jelöli, r_{im} az ion és a vízmodell negatív töltésű pozíciójának a távolságát.

Az optimált paraméterek az eredeti publikációban találhatóak. A potenciálokkal molekuláris dinamikai (MD) szimulációkat végeztem. Az eredményeket összehasonlítottam a kor legismertebb más modelljeivel és a rendelkezésekre álló kísérleti adatokkal. Máig se állítanám, hogy a legjobb kölcsönhatási készletet határoztam meg, de a szerkezeti és energetikai értékek beleillettek a kor más modelljeinek a teljesítőképeségébe. Viszont az egységes meghatározási módszer miatt olyan homológ potenciálsorozatot kaptam, ami alkalmassá tette, hogy szimulációkban ezekkel az alkáli kation illetve a halogenid sorozatokban a változásokat konzisztensen tudjuk vizsgálni. A tömbfázisbeli szimulációs eredményeket az 4.2. ábra mutatja:



4.2. ábra. Ion-oxigén (folytonos vonal) és ion-hidrogén (szaggatott vonal) párkorrelációs függvények és azok $n(r)$ integráljai. [s14]

Nem vettem be az értekezésbe azokat a cikkeket, ahol molekuláris dinamikai szimulációkat végeztünk ezekkel az ion-Hg és ion-víz kölcsönhatásokkal, mivel azok nehezen illeszthetők lennének ennek a dolgozatnak az „Adatok értelmezése szimulációkkal és a többváltozós statisztika eszközeivel a kémiában” címéhez. A kihagyott cikkeken olyan trendekre voltunk kíváncsiak, mint pl. ionok elhelyezkedésének változása a homológ sorokban, spektroszkópai eltolódások [e9] [e11] [e13]. A

szimulációkban levő többi kölcsönhatás meghatározásával ottani posztdoktor kollégáim foglalkoztak [Nazmutdinov et al. 1994, Böcker et al. 1995].

4.2 Potenciál meghatározás „on fly” a 90-es években

Az FSCF-MNDO Monte Carlo módszer és alkalmazásai [s5] [s6] [s7] [s19]

ELTE TTK Elméleti Kémiai Laboratóriumban világszínvonalú kvantumkémiai kutatások folytak, többnyire ab initio módszerekkel kapcsolatosan. Kicsit eltérő irányban kutatótt Náray-Szabó Gábor, aki a biológiai szempontból érdekesebb, nagyobb rendszerek kvantumkémiai számolásaival foglalkozott. Abban az időben, ez akár ab initio akár sűrűség funkcionál elméleti (DFT) szinten elképzelhetetlen lett volna és emiatt az ún. szemiempirikus módszereket alkalmazták a kor két populáris parametrizációjával, az AM1-gyel és a PM3-mal. A „tight binding” módszerek teljesítőképessége a molekuláris rendszerekre abban az időszakban messze rosszabb volt, mint ezeké a szemiempirikus parametrizációké. További egyszerűsítéseket is be lehet vezetni a szemiempirikus módszereknél, ilyen volt a fragment self consistent field módszer (FSCF), amellyel növelni lehetett a rendszerméretet a pontosság csökkentése mellett [Náray&Surján 1983]. 1992-ben Ferenczy György egy olyan szoftvert írt [Ferenczy et al. 1992], amelyben az akkor ismert QM/MM (kvantummechanika/klasszikus mechanika) felosztás helyett a külső részben szigorúan lokalizált molekulapályákkal szerepeltek MNDO (modified neglect of diatomic overlap) szemiempirikus kvantumkémiai szinten, míg a rendszer középső részében hagyományos MNDO közelítéssel számolták a rendszer energiáját, és így az egész rendszer akár több száz pl. szénatomot is tartalmazhatott. Felmerült, hogy használható-e a kombinált energiaszámolás menetközben (‘on fly’) Monte Carlo szimulációban.

A módszer elméleti részletei az eredeti publikációban olvashatóak [s7]. A Fortran nyelvű kód megírása, pontosabban Ferenczy György kódjának az átírása és felgyorsítása kb. háromnegyed évig tartott számomra. Ebből az utolsó három hónap a különböző gyorsítási algoritmusok kidolgozásával telt el. Első cikkünkben amorf szilíciumra alkalmaztuk a módszert [s7]. A módszer nem volt képes kötések felbontására vagy új kötések létrehozására, ezért egy fix kötések tartalmazó modellt, az ún. Wooten-modellt [Wooten&Wearie 1987] használtuk kiindulási konfigurációnak. Az akkori számítási kapacitással elérhető méretű on fly szemiempirikus kvantumkémiai Monte Carlo szimulációs módszert alkottunk meg, a futtatási adatokat az [s7] publikációból másoltam ide (4.1. táblázat). Látható, hogy 216 Si atomot tartalmazó rendszerre percenként kb. 4 konfigurációra tudtuk kiszámolni az energiát. Mai szemmel még a „multiscale” és a „shell” kifejezéseket lehet itt megemlíteni (esetünkben térben több skálás héj szerkezetű a szimuláció), hiszen két szintű leírást használtunk egy szimuláción belül, a szigorúan lokalizált MNDO és a klasszikus az MNDO-nak megfelelő alrendszereket, vagyis egy QM/QM felosztást.

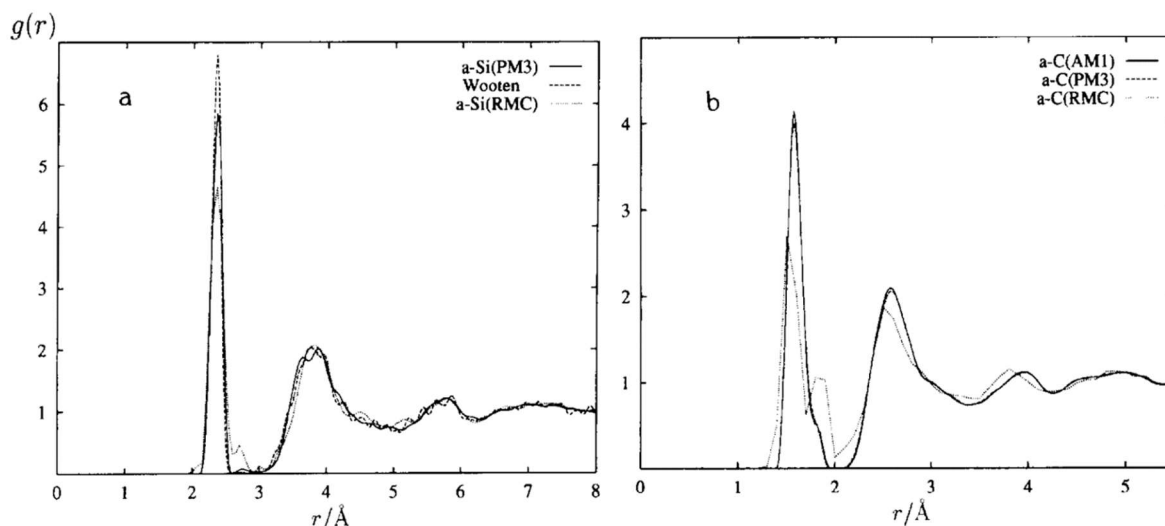
4.1. táblázat. A kifejlesztett FSCF-MNDO Monte Carlo szimuláció CPU időigénye, másodperc/konfiguráció (IBM RISC/6000 560) [S7]

Atomok száma	8	64	216
Alrendszer sugara (Å)	2,7	4,8	4,8
Nulladrendű SLMO			
első konfiguráció	0,1	22	230
további konfiguráció	0,1	1,2	7,9
Alrendszer	0,03	4,6	7,7

Az eredmények konzisztensek voltak más szimulációs eredményekkel (4.3. ábra). Ugyanebben az időben a rendelkezésünkre álló számítógépes kapacitás más módszerrel csak kisebb rendszerméretet tett volna lehetővé, pl. Car-Parrinello MD esetén alap sűrűségfukcionállal használata esetén se tudtunk volna 32 vízmolekulánál nagyobb rendszert szimulálni.

A kódot amorf szén szerkezetének modellezésére is felhasználtuk [s19]. Az eredmények értelmezése lényegében nem tért el az amorf Si-ra kapottól. A következő ábrán (4.3. ábra) az amorf Si-ra és az amorf C-re kapott párkorrelációs függvényt mutatom be. Érdeemes megjegyezni, hogy a szigorúan lokalizált és hagyományos MNDO határvonalnál (480 és 330 pm) nem látható törés, vagyis az rendszer két részre bontása szinte teljesen simán történt meg, ellentétben a kor QM/MM próbálkozásaival ilyen kovalensen kötött rendszerekre. Az ábrán diffrakciós kísérletekre számolt RMC eredmények is láthatóak.

Az FSCF-MNDO MC programot molekuláris folyadékokra is használtuk, vizsgáltuk kloroszilánok folyadékfázisú szerkezetét, illetve a gáz- és a szilárdfázisú különbségeket. Erősen limitált szimulációs rendszerünk volt, 50 molekulát tudtunk elhelyezni a periodikus szimulációs cellában. Ennek ellenére mindegyik kloroszilánra olyan elrendeződéseket kaptunk, ahol minél nagyobb számban próbált egyik molekula hidrogénje hidrogénkötést létesíteni egy másik molekula Cl atomjával. Ez valamennyire megváltoztatta a molekulák belső térszerkezetét (4.2. táblázat). Ezekről a molekulákról, illetve a folyadékfázisban levő változásaikról korábban semmilyen összehasonlítható adat nem állt rendelkezésre. Az MNDO FSCF módszerrel elért statikus és MC szimulációs eredményeket egy közös közleményben is összefoglaltuk, amelybe az MC szimulációk közül az amorf szilíciumos és kloroszilános számolásokat vettük be [s5]. Az [s5], [s6] és [s7] cikkek megjelenési sorrendje nem tükrözi a bennük leírt kutatások sorrendjét, valójában az [s7] volt az első munka, majd következett az [s6] és az [s5], és az [s19] pár évvel később ebből a témából.



4.3.ábra. Amorf szilícium (bal oldal) és amorf szén (jobb oldal) párkorrelációs függvényei különböző módszerekkel. A szilícium esetében az eredeti Wooten modell, a FSCF-MNDO MC (PM3 parametrizációval) és egy kísérlethez való RMC illesztés látható (ez utóbbi is a Wooten modell konstitúciójának megfelelően). Amorf szén esetében az RMC görbe mellett két fajta szemiempirikus paraméterezés szerepel az FSCF-MNDO MC számításnál. [s19]

4.2. táblázat. Kloroszilánok vákuumbéli (felső sorok) és folyadékfázisban (alsó sorok) kapott átlagos geometriai paraméterei (AM1 parametrizáció, MDDO és FSCF-MDDO MC módszerek). [s6]

Molekula	Paraméter		
	R(Si-Cl)	H-Si-Cl	Cl-Si-Cl
	kötéstáv (pm)	kötésszög	kötésszög
SiHCl ₃	205	109,5	109,4
	211	111,4	106,6
SiH ₂ Cl ₂	206	109,5	109,7
	214	109,2	105,5
SiH ₃ Cl	207	110,0	
	223	107,0	

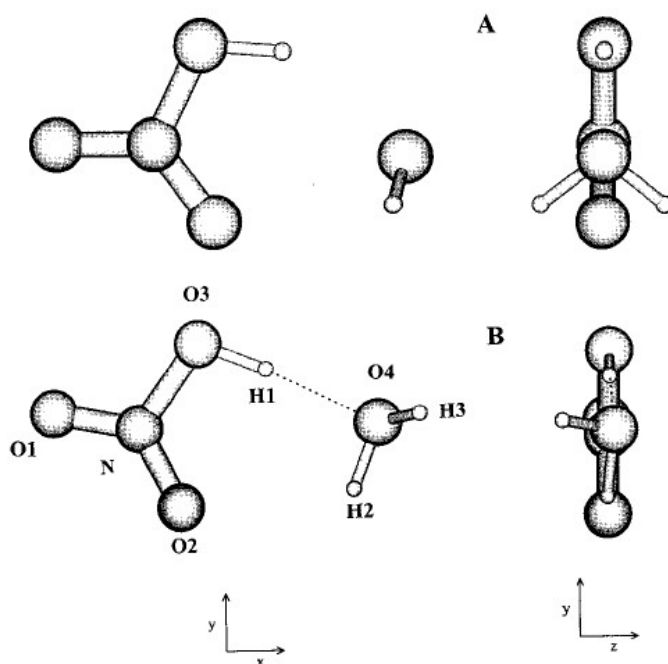
A Car-Parrinello módszer alkalmazása egy legkörkémiai kérdésnél [s17]

Mainzi időmben az egyik szomszédos intézetben légkörkémiai kutatások folytak Paul Crutzen vezetésével. Egy előadásuk alatt jött az ötlet, hogy a sztratoszférikus ózon bontásának egy részkérdését vizsgáljam Car-Parrinello szimulációval [Car&Parrinello 1986]. A mainzi idő lejártával ezzel a témával pályáztam a Magyar Zoltán posztdoktori ösztöndíjra, a kutatás helyeként az ELTE Elméleti Kémiai Tanszéket megjelölve.

A téma a salétromsav-hidrátok szerepének kvantumkémiai vizsgálata volt az ózon sztratoszférikus bomlásának folyamatában. Kristályos és amorf salétromsav-trihidrát és kisebb részben monohidrát a fő alkotója az ún. első típusú sztratoszférikus felhőnek [Graedel&Crutzen 1993, Koop et al. 1997]. Képződésük hőmérséklete a mérések szerint 195 K és felhőként két módon vesz részt a sztratoszférikus ózombontásban. Az első hatása az ózon bontásban az, hogy felhők alsóbb rétegekbe süllyedhetnek és ezzel jelentős mennyiségű salétromsav kerül ki a sztratoszférából. Ezzel csökken a salétromsavból kialakuló NO_x gyökök mennyisége, amelyek fontos szerepet játszanak a légköri aktív klórtartalmú molekulák (Cl, ClO, HOCl) dezaktiválásában. A másik szerepe, hogy a felhők felületén történik a klór tartalmú és légkörben levő HCl és ClONO_2 molekulák átalakulása aktív vegyületekké.

Munkám rögtön egy kompromisszummal kezdődött, mivel a salétromsav-trihidrát kristály elemi cellája 56 atomot tartalmaz, míg a monohidráté csak 32 atomot [s17]. Emiatt végül az összes számolást monohidrára végeztem el.

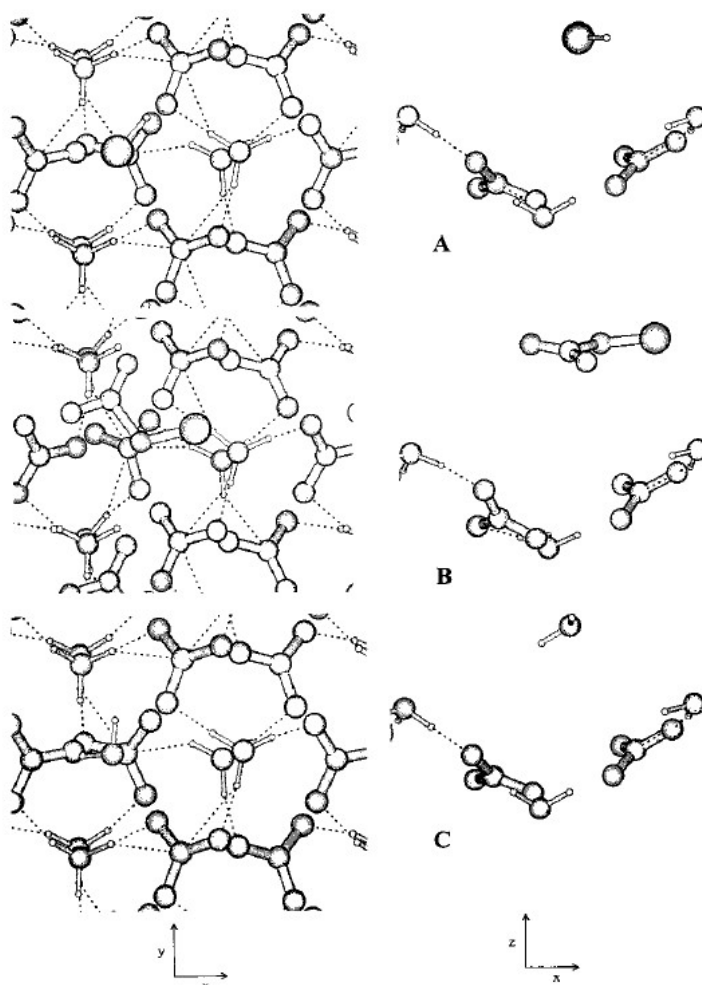
Első körben a salétromsav-víz bimolekuláris komplexet vizsgáltam szemiempirikus kvantumkémiai szinten statikusan, valamint a fragmens MNDO Monte Carlo szimulációinkkal is a 100-300 K hőmérséklet tartományban. Hartre-Fock, MP2 és DFT számolásokat végeztem a Gaussian szoftverrel [Gaussian 1993] viszonylag jónak tekinthető bázis függvényekkel. Mindemellett statikus számolásokat végeztem DFT+síkhullám bázissal pszeudopotenciálok mellett, ehhez a CPMD kódot [CPMD 1996] használtam. A CPMD-t használtam továbbá a rövid molekuláris Car-Parrinello dinamikai számolásokra is. A részletes eredményekből csak azt emelném ki, hogy megfelelően nagy bázisfüggvény és haladóbb módszerek esetén más szerkezetet kaptam az energia minimumra, mint pl. a szemiempirikus módszerekkel (4.4. ábra). A bimolekulás rendszerben nem kaptam stabil energiaminimumot arra az esetre, ha sav protonálta a vizet. A periodikus szupercellába helyezett bimolekuláris rendszerre, ha nem is egyeztek tökéletesen a geometriai és spektroszkópiai értékek a hagyományos gázfázisú számolásokkal, de nem voltak azoktól messze, vagyis úgy gondoltam, hogy megpróbálhatom a kristályos tömbfázist és annak egy felületét érintő számolásokat a CPMD kóddal.



4.4. ábra Salétromsav és vízmolekula gázfázisú optimált szerkezete, 2-2 projekcióban. Felül: szemempirikus számolás Alul: DFT számolás [s17]

A salétromsav-monohidrát kristály tömbfázisára és az (100) felületére statikus számolásokat végeztem, meghatároztam az egyensúlyi szerkezeteket és rezgési spektrumokat. Ezekben az esetekben a stabil forma az ionos, vagyis nitrát és oxónium ionok alkotják a rendszert. Mindkét rendszerre Car-Parrinello molekuláris dinamikai szimulációkat is végeztem, ahol sikerült a szimuláció során spontán reverzibilis protonátmenetet megfigyelnem a két ion között, ahol időszakosan visszaállt a semleges sav és a vízmolekula forma.

A kutatás harmadik részében azt vizsgáltam, hogy talállok-e kémiai folyamatokat elindító adszorpció helyeket a klór rezervoár molekuláknak a salétromsav (100) felületen, illetve van-e specifikus víz adszorpció. Az adszorpciót sikerült kimutatni (4.5. ábra), a kötési energiákat is meghatároztam, de az adszorbeálódó molekulák szerkezete nem változott meg. Vagyis az adott hőmérsékleti körülmények között nem sikerült megtalálnom, hogyan képződnek az aktív klór vegyületek az inaktív rezervoár molekulákból.



4.5. ábra. Optimált adszorpciós geometriák salétromsav-monohidrát (100) felületén A) HCl, B) ClONO_2 C) H_2O molekulákra. [s17]

A kutatás az elején is említett számítástechnikai korlát miatt „feldobott egy labdát”: a számolások megismétlését a relevánsabb salétromsav-trihidrát rendszerre. Két éven belül meg is jelent cikk amerikai szerzőktől, de ők se találták meg az aktiválódási reakciót [Sullivan et al. 1999]. Valójában az alapállapotú molekulák vizsgálata mellett a gerjesztett állapotokat is célszerű lett volna vizsgálni, de ezekre a rendszerekre ez még ma is komoly számítástechnikai kihívás lenne. A rendszer ma is foglalkoztatja mind a kísérleti és elméleti kutatókat, pl. [Verdes 2019].

4.3 Az inverz probléma alapjai saját tapasztalatokkal

Henderson tétele a mi szemünkkel [s24]

Az RMC módszer egyértelműségével kapcsolatban már említettem a 2000-ben Baranyai Andrással megjelent rövid összefoglaló cikkünket. Pár kérdést ott nem részleteztem ebből a cikkből, mivel tematikailag azok szorosabban kapcsolódnak ide, a potenciálmeghatározás kérdésköréhez.

Henderson 1974-ben közölte a tételét [Henderson 1974], amely szerint egy párkölcsönhatással leírható klasszikus mechanikai rendszer esetén létezik egy elméleti funkcionál pár, $F(u(r))$ és $F^{-1}(g(r))$, amelyek kölcsönösen egyértelműen összekötik a két függvényt:

$$\exists F(u(r)) \rightarrow g(r) \text{ és } F^{-1}(g(r)) \rightarrow u(r) \quad 4.3. \text{ egyenlet}$$

A 4.3. egyenletben nem jelöltem a potenciálban az energia nullpontját önkényesen definiáló konstans. Ez az elmélet tulajdonképpen klasszikus mechanikai analógja a Hohenberg-Kohn elmélet I. tételének. Az elméletet többen is tárgyalták, vagy bizonyították (Evans 1990, Zwicker&Lovett 1990, s24, Jain et al. 2006). [s24]-ben részleteztünk egy új elméleti levezetést, ahol egy kis perturbációt alkalmazva a kölcsönhatásra egy egyenlőtlenségre jutottunk, ami a funkcionál első rendű Taylor sorához vezetett és bizonyította a létezését az $F^{-1}(g(r)) \rightarrow u(r)$ inverz funkcionálnak.

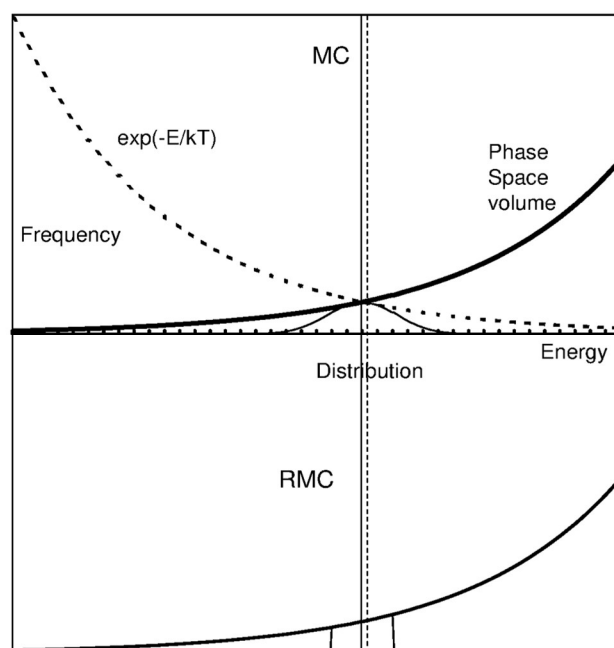
Az elmélet tovább fejleszthető párkölcsönhatásnál nagyobb tagokra is bezárólag olyan értelemben, hogy a teljes $u_2...u_n$ -nel leírható rendszer esetén a teljes $g_2...g_n$ -nel hozható kapcsolatba a fenti funkcionál. Fontos, hogy itt a hagyományos teljes változókészletű függvényekről van szó. Molekuláris esetekere akkor érvényes, ha pl. merev molekulák leírása az Euler konvenció szerint történik. A gyakorlatban sokszor ún. site-site kölcsönhatásokkal írják le a molekuláris rendszereket, vagyis kölcsönhatási pontokra bontottan és nem egyben kezelve a molekulát pl. szögváltozókkal kiegészítve. Ebben az esetben az egyértelműséghez szükséges további kényszerek figyelembevétele mind a potenciális mind az eloszlásfüggvények terében szükséges (pl. két intramolekuláris hidrogén atom egy vízben), különben információhiányos a kapcsolat. Többkomponensű rendszereknél mind parciális, mind totál függvényekre igaz, ha azok száma teljes, vagy felbonthatóságának ismert súlyokkal egyértelműen megoldható.

A Henderson-féle összefüggés érvényes bármely termodinamikai sokaságon, ha a megfelelő állapotjelzőket rögzítjük. Nem érvényes azonban arra az esetre, ha nem egy termodinamikai, pl. nem kanonikus sokaságról van szó. Mint a 3. fejezet eredményei jelezték, az RMC által generált sokaság nem mondható kanonikusnak. Vagyis az egész Henderson-féle elmélet nem igazolása az RMC módszernek, csak valószínűsíti, hogy akár működhet is, ha legalább egyensúlyi rendszerről van szó.

A fordított Monte Carlo módszer fázistérbeli mintavételezése [s32]

Az RMC anomáliáját akkor észleltük először, amikor pl. a g_2 - g_3 függvények fluktuációját, azok entrópia járulékát, vagy a konfigurációs hőmérsékletet vizsgáltuk. Még határozottabban észleltük az eltérést, amikor a Born-Green-Yvon egyenletet alkalmaztuk MC és RMC konfigurációkra (részleteiben lásd 4.4 fejezetben). A kérdés egy sematikus ábráját az [s32] publikációban mutattuk be. A 4.6. ábrán látszik, hogy a kanonikus sokaság esetén a növekvő energiaértékekhez egyre több lehetséges konfiguráció tartozik, de ezek a Boltzmann faktor révén egyre kisebb súlyt kapnak. A kialakuló eloszlás az ábrán egy maximummal rendelkező, nagy részecskeszám esetén a központi határeloszlás tétele miatt közelítőleg

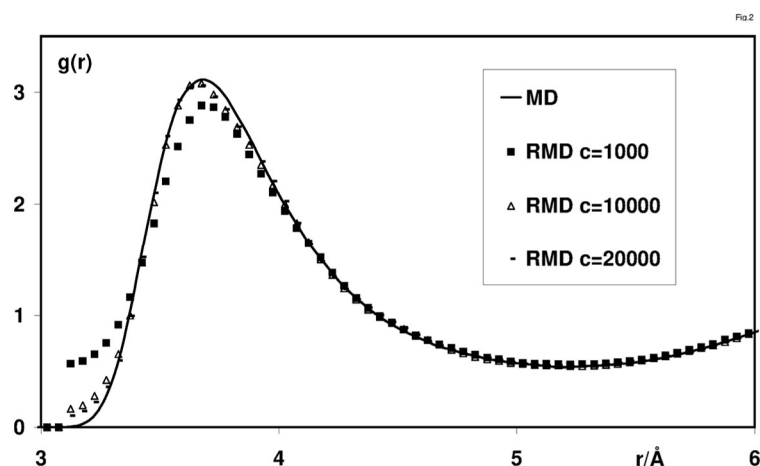
Gauss eloszlás jellegű az energia tengelye szerint. Ez persze nem mindig igaz, mivel a lehetséges konfigurációk száma a különböző rendszerek esetén eltérően változik az energia tengelye szerint, sőt, akár az unimodalitás is sérülhet. Az ábrán egy sematikus exponenciális növekedés szerepel. Az RMC esetében a $\Delta\chi^2$ kritérium viszont mindkét oldalról zárja az intervallumot, pontosabban az átlag szerkezeti függvénytől való eltérés mind a két irányban egyformán hat az energia tengelyen. A kisebb és a magasabb energiájú konfigurációk ugyanúgy kaphatnak kisebb vagy nagyobb súlyt. A kialakuló eloszlás alakja megint lehet Gauss eloszlás jellegű, de mivel több nagyobb energiájú jut be a kiválasztottak közé a nagyobb számosságuk miatt, ezért a végső átlagos energia a kanonikus átlagnál mindenképpen magasabb lesz.



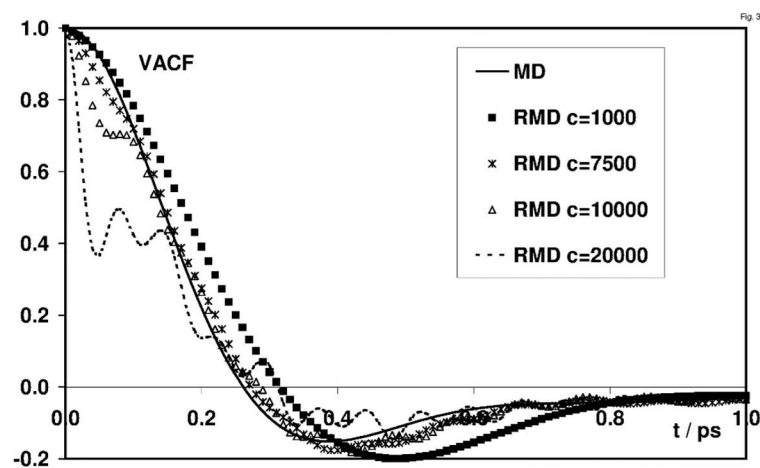
4.6. ábra. A fázistér eltérő súlyozása a klasszikus és a fordított Monte Carlo módszereknél. A klasszikus konfigurációk átlagát a folytonos függőleges vonal, míg az RMC-ét a szaggatott vonal jelöli. [s32].

A fordított molekuláris dinamika és a valós dinamika [s32]

Az RMC-vel kapcsolatos első dinamikai vizsgálatokat [s32]-ben publikáltuk. χ^2 potenciálnak neveztük el az RMC szerkezeti eltérésre utaló energia analóg tagját és bemutattuk, hogy ez a potenciáltér más dinamikát jelent: hasonló $g(r)$ mellett egészen más dinamikát kaphatunk (4.7. és 4.8. ábrák).



4.7. ábra. Lennard-Jones rendszer párkorrelációs függvényeinek első csúcsa klasszikus MD szimulációból, illetve az ezen szimulációból számolt szerkezeti függvényre végzett RMD szimulációkból. Az RMD esetében az erőterben egy c állandó szerepel, amely az erőt skálázza, de felfogható egy hőmérsékleti skálázásnak is. Nagy c értékkel szinte tökéletesen reprodukálható a párkorrelációs függvény. [s32]



4.8. ábra A 4.7. ábrán szereplő MD és RMD szimulációk normalizált sebesség autokorrelációs függvényei. A hagyományos MD dinamika még jó szerkezeti reprodukció esetében se reprodukálható RMD-vel. [s32].

Tehát kb. 2005 környékén jutottam odáig a megértésben, hogy két különböző módszerrel szimulált rendszert akkor tekintsek azonosnak, ha mind a szerkezeti, mind a dinamikai tulajdonságai azonosak. Kicsit előre ugorva az időben, ehhez 2007-ben még egy rendszerjellemzőt hozzátettem, az energetikáját, első körben a belső energiát. Ezt a hármas megközelítést ma is élőnek tartom, tehát úgy gondolom, hogy pl. két szimulációs módon előállított rendszer minden szerkezeti, dinamikai és energetikai tulajdonságának azonosnak kell lenni. Ezzel kapcsolatban indítottam el az egyik utolsó kutatásomat a témában egy szakdolgozó bevonásával 2008-ban. Mindezek az összefüggések azonban eddig az inverz probléma témakörében kerültek általam kutatásra. Az inverz probléma pedig egy rendkívül szűk körben érdeklődést felkeltő téma volt. Nem kapcsolódtak hozzá önálló konferenciák, a világban 4-5 kutató reflektált az addigi eredményeimre. 2010. környékén feladtam a kutatásokat ezen a területen. Bár ismertem a coarse graining problémakört, sőt láttam annak az inverz problémával való kapcsolatát, valamiért nem bíztam abban, hogy ez a nagyon szűkkörű terület valaha is fontos lehet. Az RMC módszer határainak felismerésével elindult egy kutatási irányom, ahol az inverz problémát a potenciálmeghatározás irányából közelítettem meg. Az első munkában még Baranyai Andrással és kapcsolódóan az RMC-hez, majd többnyire egyedül vagy szakdolgozók bevonásával. Összesen négy módszert dolgoztam ki a téma megoldására, jelentősen eltérő kiindulási ötletekkel.

4.4 Saját módszerek az inverz probléma megoldására

A Born-Green-Yvon hierarchia és a fordított Monte Carlo módszer kombinálása [s26]

A Born-Green-Yvon egyenlet egzakt módon köti össze egy párkölcsönhatásokkal leírható rendszer itt $\varphi(r)$ -rel jelölt párkölcsönhatását, párkorrelációs függvényét, a $g(r)$ -t, és a triplett korrelációs függvényét, a $g(r,s,t)$ -t [Hill 1956].

$$-\frac{d}{dr}(\ln g(r) + \beta\phi(r)) = \beta\pi\rho \int_0^\infty ds \frac{d\phi(s)}{ds} \int_{|r-s|}^{r+s} dt \frac{t(r^2 + s^2 - t^2)}{r^2} \frac{g(r,s,t)}{g(r)} \quad 4.4. \text{ egyenlet}$$

A folytonos egyenlet diszkrét függvényekre is átírható, ha közös ekvidisztáns távolságfelosztásokat használunk az r , s , és t pártávolságokra, azokat R , S és T -vel jelölve és Δ értékeket úgy értelmezve, hogy az adott R diszkrét érték pl. az $[R-\Delta, R]$ intervallumot jelenti [s23][s26].

$$\begin{aligned} & -\frac{\ln g(R+\Delta) - \ln g(R-\Delta)}{2\Delta} - \beta \frac{\phi(R+\Delta) - \phi(R-\Delta)}{2\Delta} \\ & = \beta\pi\rho \sum_{S=R_{\min}}^{R_{\max}} \left(\frac{\phi(S+\Delta) - \phi(S-\Delta)}{2\Delta} \sum_T F(R,S,T) \frac{g(R,S,T)}{g(R)} \right) \quad 4.5. \text{ egyenlet} \\ & g(R) \equiv 3N_R / 4\pi(R^3 - (R-\Delta)^3). \end{aligned}$$

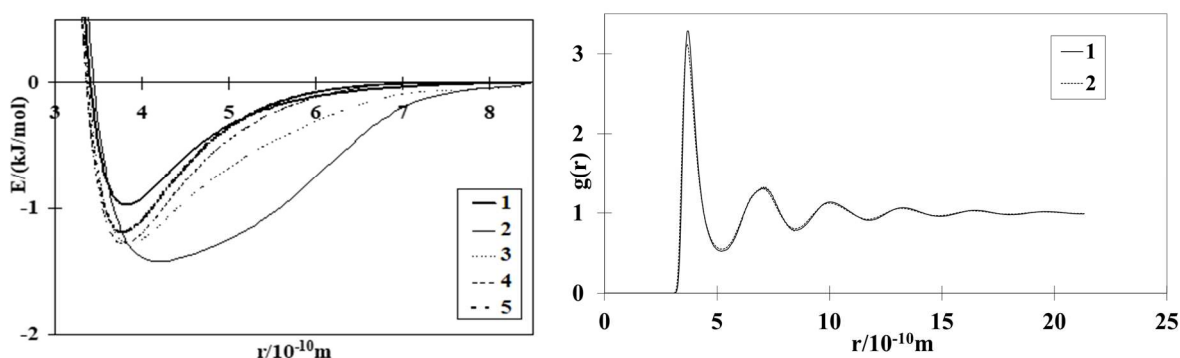
A matematikailag munkaigényes feladat F meghatározása, mivel az r, s, t értékeknek teljesítenie kell a háromszögegyenlőtlenséget, vagyis csak ott értelmezhető egy adott R, S, T esetén az integrál, ahol bármelyik kettő összege nagyobb a harmadiknál.

$$F(R, S, T) = \int_{S-\Delta}^S ds \int_{T-\Delta}^T dt \frac{v^2 + s^2 - t^2}{v^2} t \quad 4.6. \text{ egyenlet}$$

ahol $v = R - \Delta/2$. $F(R, S, T) = 0$, ha $T < \max(R - S, -\Delta)$ vagy $T > R + S$.

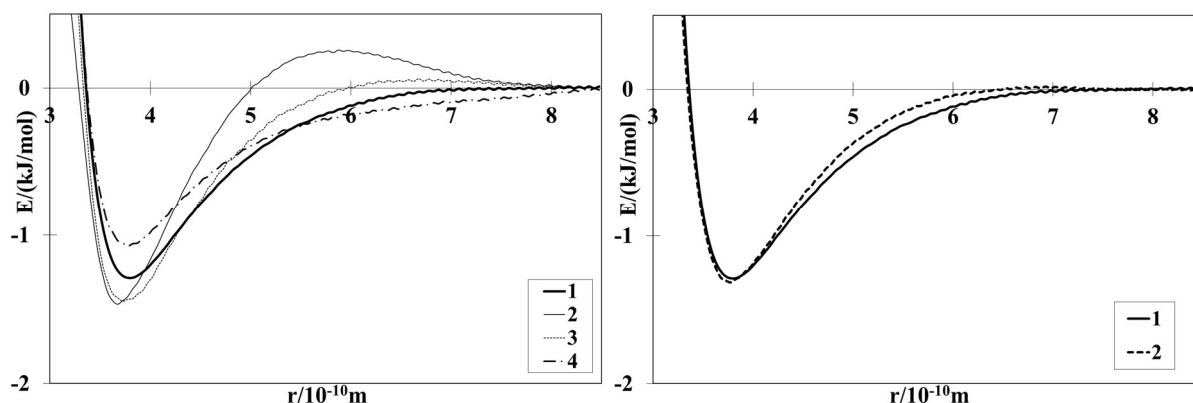
F meghatározása a három változó aszimmetriája miatt sokkal több alesetre esik szét, mint a pl. a $g(r, s, t)$ entrópia járulékanak a számításánál. Számomra emlékezetes, hogy két hónapig tartott, amíg papíron elvégeztem a paraméteres integrálást F összes esetére. Az eredmény programozása után a próbaszámolásoknál derült ki, hogy valamit elnéztünk Baranyai Andrással a levezetésnél. Újabb két hónap helyett megtanultam az akkor még újdonságnak számító Mathematica szoftvert és azzal két nap alatt megkaptam a javított összefüggésre a képleteket.

A fenti egyenlet a potenciálra egy lineáris egyenletrendszert jelent. A BGY egyenlet numerikus bemutatását és a Δ felosztás hatását úgy ellenőriztük, hogy Lennard-Jones rendszerre klasszikus MC szimulációból jó statisztikával kiszámoltuk a $g(r)$ és $g(r, s, t)$ függvényeket, majd ezek ismeretében meghatároztuk a hozzájuk tartozó $\varphi(r)$ kölcsönhatási potenciált. Jól látható az 4.9. ábrán, hogy Δ -t viszonylag kicsire kell venni, hogy jó eredményeket kapjunk. Ennek egyik fő oka, hogy olyan felosztást kell elérni, ahol egy-egy adott intervallumra már elfogadható közelítés az állandó érték mind a potenciál, mind a korrelációs függvények esetén. Ugyanakkor kis Δ esetén a diszkrét $g(r, s, t)$ elfogadható statisztikájú értékeihez nagyon nagy mintára és számábrázolási pontosságra van szükség. Az akkori számítástechnikai lehetőségek mellett nem tudtunk tökéletes egyezést elérni. A felosztás memória igénye, a szükséges statisztika és az ahhoz kapcsolódó szokásos duplapontosságú számábrázolás nem tette ezt lehetővé.



4.9. ábra.: Bal oldal: Párpotenciál meghatározásának felosztásfüggése. 1. eredeti LJ potenciál 2-5. diszkrétizált BGY egyenlettel kapott potenciálok 2. $\Delta = 0.2 \cdot 10^{-10}$ m, 3. $\Delta = 0.1 \cdot 10^{-10}$ m, 4. $\Delta = 0.05 \cdot 10^{-10}$ m, 5. $\Delta = 0.0333 \cdot 10^{-10}$ m. Jobb oldal: Az eredeti MC és az általunk akkor elérhető legkisebb felosztással kapott potenciállal végzett MC szimuláció párkorrelációs függvényei. [26]

Az RMC módszer újdonsága a korábbi diffrakciós kiértékelési megoldásokhoz képest az volt, hogy az egyváltozós párkorrelációs vagy szerkezeti függvény alapján háromdimenziós szerkezeteket generált. Ezekre a szerkezetekre kiszámolható a pár- és a triplett korrelációs függvény. Vagyis, ha az RMC által generált sokaság megfelelne pl. a kanonikus sokaságnak, akkor a diszkretizált párkölsönhatás meghatározható lenne ezen az úton diffrakciós szerkezeti adatokból. Több RMC variációt is vizsgáltunk, hogy melyik szolgáltat jobb kölcsönhatást. A szokott módon klasszikus Lennard-Jones MC szimuláció $S(q)$ -jára és/vagy $g(r)$ -jére végeztünk RMC szimulációkat, majd ezekre a konfigurációkra oldottuk meg a BGY egyenletet. Sajnos az alapesetekben nem kaptunk jó potenciálokat, de létezett olyan kombináció, ami kielégítő eredményt adott (4.10. ábra). Talán ez volt az a pont, amikor egyértelműen eljutottam arra a következtetésre, hogy az RMC sokaságát úgy lehet közelíteni a kanonikus sokasághoz, ha minél több input adatot használunk, lehetőleg mind a q és az r térből, mivel a különböző terek mérési pontjai eltérő súlyfaktorokkal jellemezhetőek. Ez a szemlélet kb. összeköthető a statisztikus termodinamikával, hogy sok apró hozzájárulás kiegyensúlyozottabb megoldást ad. Elnagyolt analógia lehet a double descent (kettős lecsengés) jelenség is a neurális hálózatokkal kapcsolatban [Vallet 1989, Spigler 2019, Belkin 2019].



4.10. ábra. Bal oldal: BGY-RMC közelítéssel kapott párpotenciálok. 1. eredeti Lennard-Jones 2-4. RMC szimulációk konfigurációi alapján, 2. $S(q)$ eltérésnégyzet alapján, q független hibával és 0,45 elfogadási arány mellett, 3. az előző 0,2 elfogadási aránnyal 4. $g(r)$ eltérés a 4. hatványon alapon r -függő hibával 0,2-es elfogadási aránnyal. Jobb oldal: Párkorrelációs függvények 1) eredeti MC 2) RMC konfigurációkból, ahol kombináltan $S(q)$ négyzetes és $g(r)$ 4. hatványos célfüggvény volt q független és r -függő hibával 0,2-es teljes elfogadási aránnyal. [s26]

A módosított inverz Monte Carlo eljárás [s27] [s29]

Lyubartsev és Laaksonen 1995-ben publikálták módszerüket [Lyubartsev&Laaksonen 1995], amelyben egy általános levezetést követően az r -térben azonos felbontást használva a párkorrelációs függvényre

és a párpotenciálra bevezettek egy iteratív sémát az utóbbi meghatározására. Az eljárásukban egy-egy iterációs lépésben lineáris egyenletrendszer közelítéssel finomították a kezdeti potenciált (ismeretlenek és független egyenletek száma azonos volt). Minden iterációban egy-egy szimulációt kellett elvégezni. Az egyenletrendszer felállításához szükséges parciális deriváltakat egy fluktuációs séma segítségével számolták ki. Módszerük azóta az egyik leggyakrabban használt eljárás a coarse graining potenciálok előállításánál, és IMC rövidítéssel szokták jelölni.

Módszerüket 2001-ben továbbfejlesztettem két irányban. Bemutattam, hogy a módszer alkalmazható $S(q)$ függvényre is, valamint az egyértelműen megoldható egyenletrendszer helyett túlhatározott egyenletekre vonatkozó megoldási sémákat javasoltam [s27, s29].

Egy tetszőleges B_j fázistérbeli mennyiség a következő átlagolással számolható a fázistér pontjainak ismeretében:

$$\langle B_j(\Gamma, \underline{P}) \rangle = \frac{\int B_j(\Gamma, \underline{P}) \exp(-\beta H(\Gamma, \underline{P})) d\Gamma}{\int \exp(-\beta H(\Gamma, \underline{P})) d\Gamma} \quad 4.7. \text{ egyenlet}$$

,ahol $\underline{P}$ a potenciális energiafüggvény paramétereit jelenti, Γ a fázisteret. B_j értékének változása felírható első közelítésben a $\underline{P}$ megfelelő változói szerinti parciális deriváltakkal:

$$\Delta \langle B_j(\Gamma, \underline{P}) \rangle = B_j^{exp} - \langle B_j(\Gamma, \underline{P}) \rangle^{simulated} = \sum_i^{np} \frac{\partial \langle B_j(\Gamma, \underline{P}) \rangle}{\partial p_i} \Delta p_i + o(\Delta P^2) \quad 4.8. \text{ egyenlet}$$

,ahol a változást lehet úgy is értelmezni, mint egy szimulált és egy kísérleti görbe értékei közti eltérést. A Laaksonen és Lyubartsev féle fluktuációs forma segítségével a deriváltak számolhatóak, pl. adott $S(q)$ értéket tételezve fel B_j helyett:

$$\frac{\partial \langle B_j(\Gamma) \rangle}{\partial p_i} = \frac{\partial \langle S_j(\Gamma) \rangle}{\partial p_i} = -\beta (\langle S_j(\Gamma) D_i \rangle - \langle S_j(\Gamma) \rangle \langle D_i \rangle) \quad 4.9. \text{ egyenlet}$$

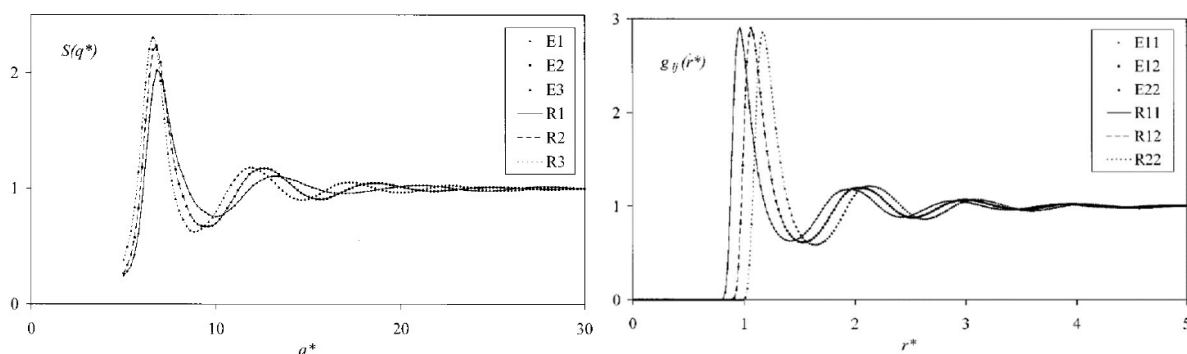
D_i a potenciáltól függően más és más, pl. Lennard-Jones kölcsönhatás esetén:

$$D_k = \sum_i^N \sum_{j>i}^N 4 \left(\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right) \quad \text{ha } p_k = \varepsilon \text{ és}$$

$$D_k = \sum_i^N \sum_{j>i}^N 4\varepsilon \left(12 \frac{\sigma^{11}}{r_{ij}^{12}} - 6 \frac{\sigma^5}{r_{ij}^6} \right) \quad \text{ha } p_k = \sigma \quad 4.10. \text{ egyenlet}$$

A 4.8. egyenlet valójában egy lineáris egyenletrendszer, amely a Δp_i ismeretlen értékekre megoldható egyértelműen, vagy amennyiben B pontjainak száma nagyobb, mint a p_i -k száma, akkor a túlhatározott egyenletrendszer szabályainak megfelelően. Első munkánkban a megoldást SVD algoritmus segítségével végeztem, a túlhatározott egyenletrendszerek általánosított inverzet alkalmazó alakjával. A módszert leellenőriztem egy- és kétkomponensű Lennard-Jones rendszereken, ez utóbbinál a kétkomponensű Lennard-Jones hagyományos MC szimulációira számolt három totál $S(q)$ függvény volt a bemeneti adata az a módosított IMC eljárásomnak. Egy komponensű rendszere 2-3, a kétkomponensű esetében 8-10 iteráció alatt 0,1 %-on belül elértem Lennard Jones paraméterek elvi

értékét, az illesztett teljes $S(q)$ függvények és a parciális $g(r)$ -ek szinte tökéletes egyezést mutattak az eredetiekkel (4.11. ábra).



4.11. ábra. Bal oldal: Kétkomponensű Lennard Jones rendszerre számolt három totál $S(q)$ reprodukciója a módosított IMC módszer segítségével. Jobb oldal: A három eredeti parciális párkorrelációs függvény (E) és a módosított IMC eljárással kapott potenciállal (R) végzett szimulációké. [s27]

A módszert közvetlenül argon röntgendiffrakciós mérési eredményére is vizsgáltam, ahol az adott méréssel konzisztensebb Lennard Jones paraméter értékeket kaptam vissza, mint a szokásos, de szélesebb állapotjelző tartományra optimalizált paraméter készlet.

A fenti egy iterációs lépésre alkalmazott lineáris egyenletrendszer megoldása helyettesíthető a nemlineáris paraméterbecslés más módszereivel is, pl. Gauss-Newton-Marquardt-Levenberg módszerrel:

$$\underline{P}^{(k+1)} = \underline{P}^{(k)} + (J^T W J + \lambda^{(k)} I)^{-1} J^T W (\underline{B}^{exp} - \underline{B}^{sim}) \quad 4.11. \text{ egyenlet}$$

, ahol a Jacobi mátrix elemei:

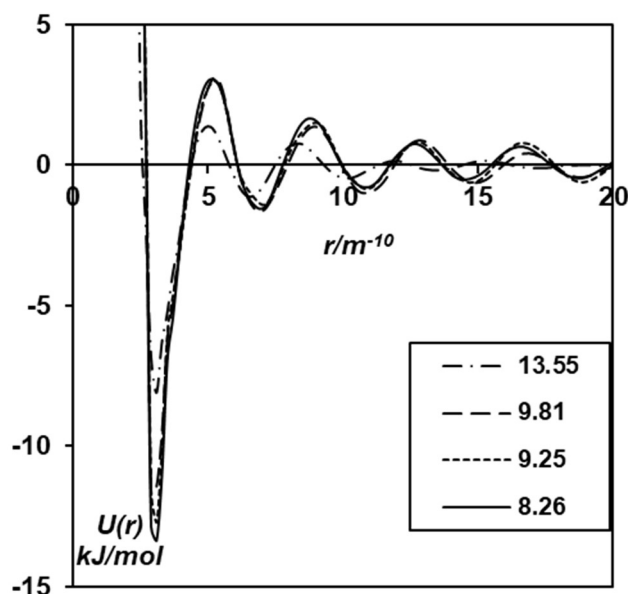
$$J_{ji} = \frac{\partial \langle B_j(\Gamma, \underline{P}^{(k)}) \rangle}{\partial p_i}$$

és

$$W = \begin{bmatrix} w_1 & 0 & \cdot & 0 \\ 0 & w_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & w_{ne} \end{bmatrix} \quad 4.12. \text{ egyenlet}$$

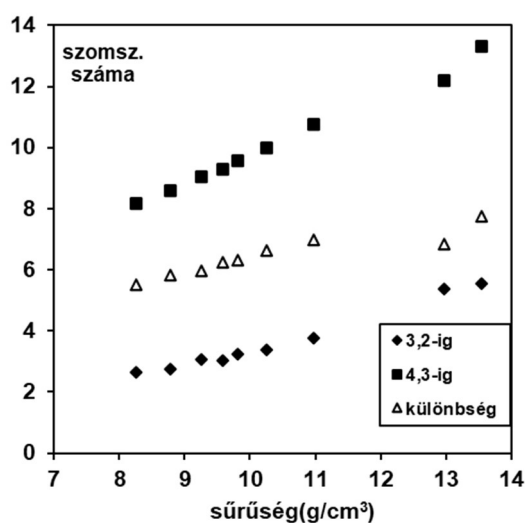
tartalmazza a kísérleti pontok súlyát.

A megoldás itt is a 4.8. egyenlet fluktuációs formulájával számolható. Ennek a változatnak a szemléltetését egy folyékony higanyra készült sorozatmérés kiértékelésével mutattam be [s29]. 10 különböző hőmérsékletű Hg-ra vezettem le így párpotenciált, és elemeztem a végső szimulációs eredményeket (4.12. ábra).



4.12. ábra. Tabulált kölcsönhatási potenciál folyékony Hg-ra különböző sűrűségeken a módosított IMC módszer segítségével. [s29]

A potenciálok jól mutatták a fémek esetében ismert ún. Friedel oszcillációt. A kiértékelés célja volt az is, hogy megerősítsünk korábbi feltevéseket, pl., hogy a higany sűrűségfüggő vezető-félvezető átmenete kapcsolatba hozható a közeli, 3.2 Å-on belüli szomszédok számának a csökkenésével (4.13. ábra). Erre az első sejtést Franz számolásai adták [Franz 1986], amit RMC szimulációk is valamennyire alátámasztották [Arai&McGreevy 1998], pedig ott sokkal rendezetlenebb szerkezeteket kaptak mint én.

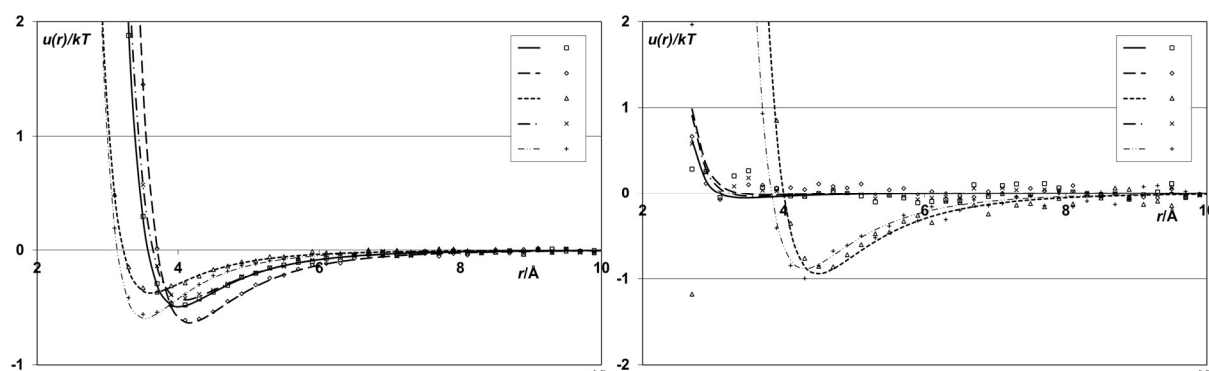


4.13. ábra. Átlagos szomszédsszámok különböző folyékony higany sűrűségeknél 3,2 Å és 4,3 Å távolságokon belül, illetve a különbségük. [s29]

Számos numerikus nehézség adódott a tabulált potenciál meghatározása során, ilyen volt pl. a kis r tartomány, ahol a mért $S(q)$ tartomány nem ad releváns információt. Kérdéses volt a szomszédos tabulált potenciál értékek harmonizálása és a legnagyobb értéknél a 0-hoz konvergálás is. A problémát Golay-Savitzky simítás közbeiktatásával kezeltem az új tabulált potenciálra, trükköket alkalmazva az elejére és a végére, pl. ez utóbbit 2 extra 0-n tartott potenciálpont hozzáadásával a simítás előtt. Az első rész illesztése különösen kritikus volt, mivel ez a rész van a legnagyobb hatással a rendszerek nyomására. A Hg szimulációim számolt nyomásértékei nagyon fluktuáltak az iteráció során és a végső potenciálok se lettek tökéletesek nyomás szempontjából.

Az inverz probléma megoldása mesterséges ideghálóval [s31]

Numerikus matematikai oktatási feladataim révén elméletben megismertem a mesterséges ideghálókat, többek között mint egy eszközt ismeretlen alakú függvények/funkcionálok közelítésére. 2005-ben megjelent cikkünkben az $S(q) \rightarrow u(r)$ kapcsolatot próbáltuk meg felderíteni mesterséges idegháló alkalmazásával. 2200 molekuláris dinamikai szimulációt hajtottunk végre Lennard-Jones, Morse és Buckingham párkölcsönhatásokkal egykomponensű rendszereken. Viszonylag széles hőmérséklet és sűrűsértományú rendszereket választottunk. A háló bemenő adatai a sűrűség mellett az $S(q)$ értékek voltak, 53 diszkrét q pontban szimulációként. A kimenő adat a tabulált potenciál volt, 62 diszkrét r értéknél. A potenciál típusai szerint külön-külön és különböző kombinációkban közös modelleket is létrehoztunk. Teszt készletnek 200 szimulációt választottunk. Különböző felépítésű mesterséges ideghálókkal próbálkoztunk, a legjobb eredményeket egy rejtett rétegű és abban 60 logisztikus aktivációs függvényű neuront tartalmazó rendszerrel kaptuk. A részletes elemzés helyett itt csak néhány eredményt mutatok (4.14. ábra):



4.14. ábra. Bal oldal: Az eredetivel a legjobb egyezést mutató 5 párpotenciál a teszt halmazból. Jobb: Az eredetivel a legrosszabb egyezést mutató 5 párpotenciál a teszt halmazból. [s31]

Ezek alapján nem tűnne túl jónak a módszer, hiszen a legrosszabb potenciálok vizuálisan is eléggé gyenge eredményeket mutattak, nem csak az általunk használt numerikus hibafüggvényénél. Azonban pl. a Lennard-Jones kölcsönhatásokra kiderült, hogy néhány esetre nagyon rossz eredményeket kapunk, de ha pl. a hiba felől nézve kb. 10 %-os percentilisének levő eredményeket nézzük, azok már kifejezetten jók, vagyis az eredmények kb. 90%-a használható.

A módszer rövid ideig a neten is szerverszolgáltatásként elérhető volt, Kozma Dániel egy TDK munka keretében írta meg a szerveren futtatható programot. A feltöltött $S(q)$ és sűrűség esetén a program szinte azonnal visszaadta a kész tabulált kölcsönhatást.

Egy izraeli konferencián mutattam be ezeket az eredményeimet. Egy fizikai elméleti problémának a fekete doboz megoldása nagy vitát váltott ki. Akkor, 2005. táján a hallgatóság pozitívan értékelte a teljesen szokatlan ötletet, de minden felszólaló kiemelte, hogy az nem válhat a jövő útjává, hogy fizikai alapokat nélkülöző matematikai módszerekkel oldjunk meg egy elméleti kérdést. Kíváncsi lennék, hogy az akkor jelenlevők közül hányan adtak be mostanában olyan pályázatot, ahol a mesterséges intelligenciát próbálják meg alkalmazni fekete dobozként hasonló feladatokra. Az inverz probléma gépi tanulási megoldása azóta se történt meg, bár vannak próbálkozások, pl. egy dimenziós esetre [Kampa et al. 2025], vagy a folyadékok integrál elméletében szereplő híd (bridge) függvény közelítését próbálták meg hasonló módon [Goodall&Lee 2021]. Tomaz Urbic szlovén kollégával mi is kísérletet tettünk 2023-ban arra, hogy tovább lépve a régi ötletből $g(r)$ -ek és $g(r,s,t)$ -k ismeretében neurális hálókval reprodukáljuk az $u(r)$ -ket, majd a neurális hálók interpretációja segítségével igazoljuk vagy továbbfejlesztjük az egyes zárási megközelítéseket a folyadékok integrálegyenlet elméletében, de első körben eddig nem sikerült megfelelően teljesítő gépi tanulás alapú modellt létrehozunk.

Párpotenciál meghatározása potenciális energia illesztésével [s36]

A korábbi évek kutatásai, különösen az RMC módszer vizsgálata vezetett el ahhoz, hogy egy rendszer leírását két szimulációban akkor tekintsem azonosnak, ha mind a szerkezetet, mind a dinamikát sikerül reprodukálni. Bár a Henderson-féle elmélet bemutatja, hogy az összes parciális párkorrelációs függvény ismerete elvileg elegendő egy parciális párpotenciállal leírható rendszer kölcsönhatási függvényeinek meghatározásához, de mint az RMC módszernél láttuk, a párkorrelációs függvények egyezése mögött más dinamika is lehet, ha nem azonos sokaságnak megfelelő mintából lettek számolva. A dinamika eltérése jelentős, ha csak a rendszer egy részére történik már a modellezés, pl. coarse graining céllal. A negyedik inverz elméletre kifejlesztett megoldó módszerem [s36] egy harmadik szempont, az energia oldaláról közelítette meg a feladatot. Egy kétkomponensű rendszer potenciális energiája párkölcsönhatások esetén felírható a párkorrelációs függvény ismeretében, akár egy konfigurációra is, ahol k az adott konfigurációt jelöli az alsó 1-2 indexek pedig mutatják, hogy melyik parciálisról van szó:

$$E^{(k)} = E_{11}^{(k)} + E_{22}^{(k)} + E_{12}^{(k)} = \frac{1}{2} \sum_i^{Nr} V \frac{4\pi\rho_1^2}{3} (r_i^3 - r_{i-1}^3) g_{11}^{(k)}(r_i) u_{11}(r_i)$$

$$+ \frac{1}{2} \sum_i^{Nr} V \frac{4\pi\rho_2^2}{3} (r_i^3 - r_{i-1}^3) g_{22}^{(k)}(r_i) u_{22}(r_i) + \sum_i^{Nr} V \frac{4\pi\rho_1\rho_2}{3} (r_i^3 - r_{i-1}^3) g_{12}^{(k)}(r_i) u_{12}(r_i) \quad 4.13. \text{ egyenlet}$$

Ha ismeretlennek tekintjük a kölcsönhatási potenciált és annyi konfigurációról van adatunk, mint ahány N_r számú a $g(r)$ felosztása szorozva a parciálisok számával, akkor felírható egy lineáris egyenletrendszer:

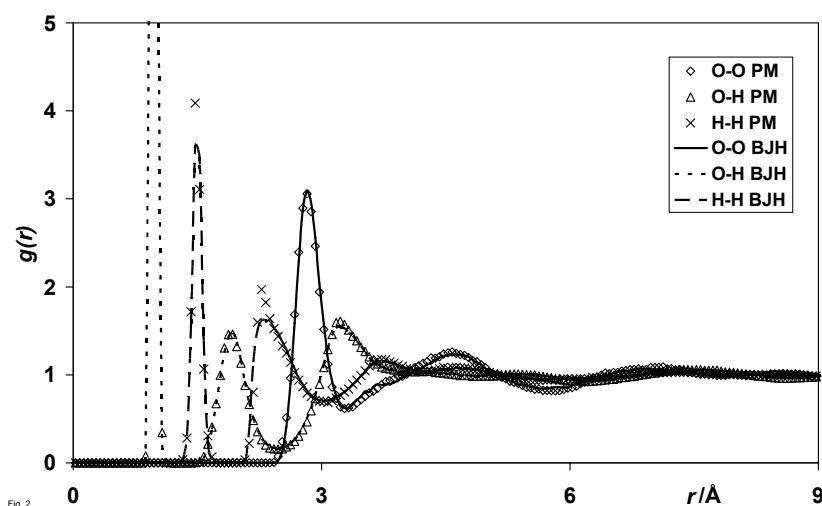
$$A\mathbf{u} \approx \mathbf{e} \quad 4.14. \text{ egyenlet}$$

,ahol A sorai tartalmazzák a parciális $g(r)$ -ket egymás után írva egy sorba konfigurációként, megszorozva az egyenletben előtte levő tagokkal, $\mathbf{e}$ tartalmazza a potenciális energiákat vektorban, $\mathbf{u}$ pedig a parciális párkölcsönhatási függvényeket, ez utóbbiakat mint ismeretleneket. Amennyiben több konfigurációról van adatunk, mint ahány párkölcsönhatási értéket keresünk, akkor túlhatározott az egyenletrendszer és annak megfelelő megoldási sémát kell használni.

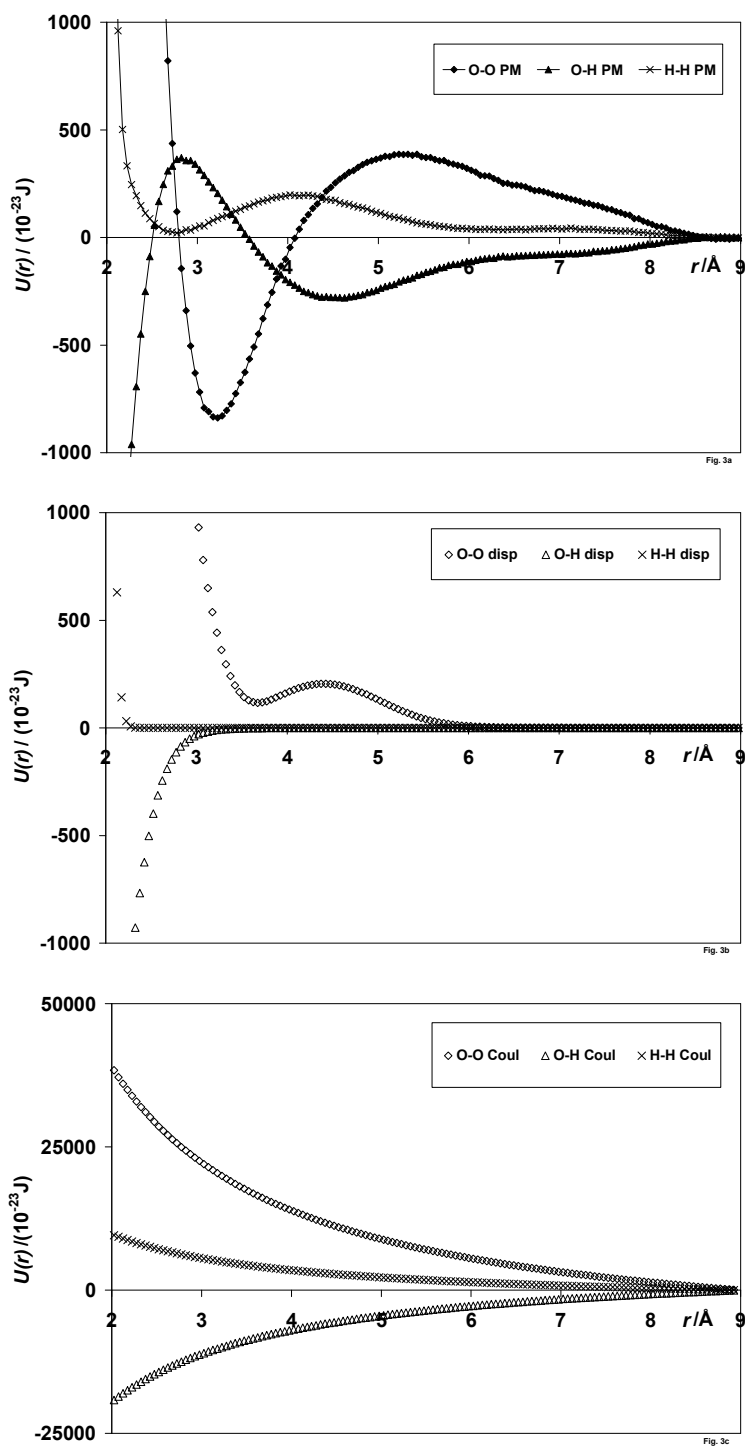
Ez a tiszta elméleti megközelítés a gyakorlatban azt jelenti, hogy az inverz probléma ebben az esetben nem az egyensúlyi rendszer átlagos $S(q)$ -jához vagy $g(r)$ -jéhez kapcsolódó $u(r)$ meghatározását jelenti, hanem egy olyan esetét, amikor ismerjük a részecskék egy-egy konfigurációjából számítható párkorrelációs függvényeit. A módszer működhet analitikus potenciál formára is, lineáris paraméter meghatározása esetén lineáris egyenletrendszerként, de nemlineáris paraméterbecslés formájában bármilyen paraméterre. A gyakorlatban a módszer inkább a coarse graining esetben használható, pl., ha az összes atomot tartalmazó szimulációból olyan potenciál meghatározása a cél, amit redukált, pl. oldószermentes szimulációban szeretnénk használni. Persze ebben az esetben célszerű más szimulációs módszert választani az új rendszerhez, pl. Langevin-dinamikát. Csak a szerkezet reprodukálásához jó a hagyományos MC vagy MD szimuláció is. Egy másik lehetőség, hogy „on fly” kvantumkémiai módon számolt konfigurációk alapján egyszerűbb klasszikus mechanikai potenciálokat vezetünk le a módszerrel. Ebben a témában létezik egy alternatív, gyakran használt módszer, Izvekov módszere, a „force matching” [Izvekov et al. 2004]. Más matematikai elven, de ott a részecskére ható erő leképezése történik meg valamilyen klasszikus formára.

A módszer működőképességét olyan eseteken mutattam be, ahol túlhatározott egyenletrendszereket kellett megoldani, minimum tízszeres túlhatározottsággal. A konfigurációkból számolt párkorrelációs függvények azonban nagyon hasonlóak voltak, hasonló sorokat eredményezve A -ban. Így az invertálandó mátrix többnyire kvázi-szinguláris volt. A hagyományos megoldások, pl. szinguláris érték felbontással, vagy a sajátérték egyenletre való visszavezetéssel nem tudtak eredményt adni. Végül egy iteratív megoldási módszert használtam, amelyik az adott programnyelv lineáris algebrai könyvtárának része volt.

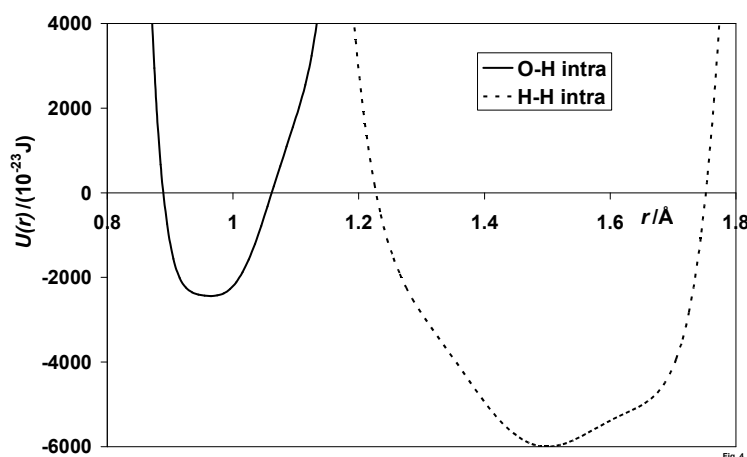
A „potential matching” (PM) módszer tökéletesen működött Lennard-Jones rendszerre egy és két komponensű esetekben, utóbbinál, ha mindhárom parciális kölcsönhatást használtuk az új szimulációban is. Érdekes eset volt a BJH vízmodell kölcsönhatásainak vizsgálata (4.15. ábra), itt tökéletesen sikerült reprodukálni a parciális párkorrelációs függvényeket, míg a potenciál nem volt összehasonlítható az eredetileg diszperziós és coulomb kölcsönhatásokat is tartalmazó intermolekuláris kölcsönhatással. Sikerült az intramolekuláris kölcsönhatást is meghatározni. Eredményeink alapján is elmondható, hogy molekuláris rendszereknél a klasszikus felbontás diszperziós, elektrosztatikus intermolekuláris és valamilyen, pl. harmonikus intramolekuláris részre nem reprodukálható, mivel ezek a tagok a molekuláris belső kötöttség miatt csatolódnak, amit a PM algoritmus máshogy szed szét párkölcsönhatásra, mint az eredeti szétbontás. (4.16. és 4.17. ábrák)



4.15. ábra. BJH vízmodell és a PM módszerrel levezetett kölcsönhatásokkal kapott parciális párkorrelációs függvények. [s36]



4.16. ábra. A PM módszerrel levezetett parciális párpotenciálok (fent), a BJH vízmodell diszperziós párkölcsönhatási része (középen) és elektrosztatikus része (lent). [s36]



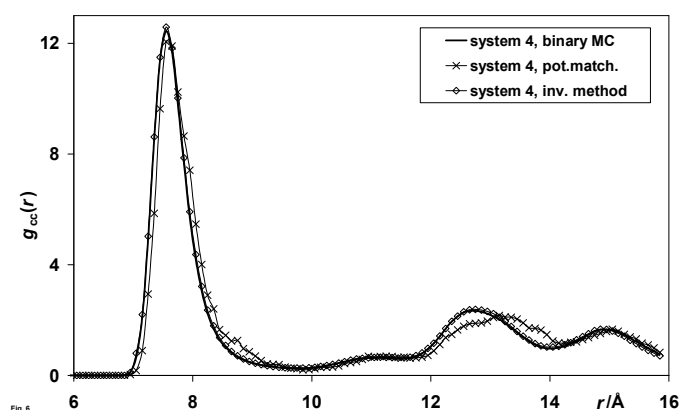
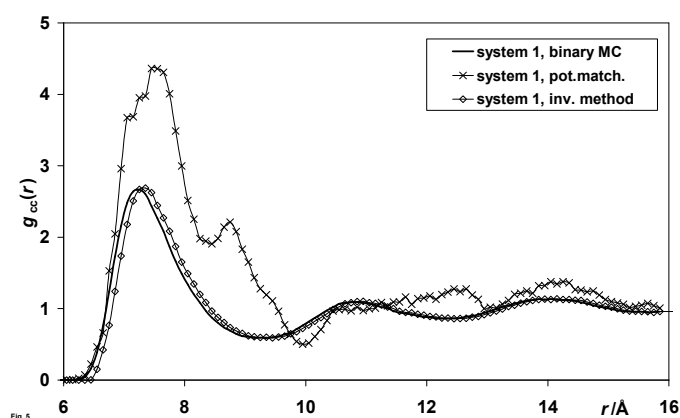
4.17. ábra. A PM módszerrel a BJH vízmodellre levezetett intramolekuláris párpotenciálok. [s36]

Ebben cikkben viszonylag részleteztem a coarse graining eseteket is, vagyis amikor pl. kétkomponensű Lennard-Jones rendszerből csak az egyik (pl. a nagyobb) atomra vezetem le kölcsönhatási potenciált. 4 rendszeren vizsgáltam a PM és egy klasszikus $g(r)$ -ből induló inverz módszer teljesítőképességét, ha csak az egyik részecske típusra egyszerűsítettem a rendszert (4.3. táblázat). Minden esetben a megtartott részecskék kétszeres átmérőjűek voltak az elhanyagolthoz képest. Vizsgáltam a szerkezet és az energia reprodukálását. A szerkezet alapú potenciál illesztés minden esetben jó párkorrelációs függvényt adott (4.18. ábra), de az eredeti szimulációk és a PM potenciállal végzett szimulációk energiája csak azoknál a rendszereknél korrelált jól, ahol jelentősen nagyobb volt a nagyobb részecskék közötti kölcsönhatás aránya, mint a többié. Vagyis a lokális rendet a nagy részecskék között erősen befolyásolta a nagy-nagy részecskék direkt kölcsönhatása, az oldószer hatása inkább térkitöltő és entrópiikus volt. A PM eljárás is csak ennél a két rendszernél működött jól.

A cikk végén viszonylag részletesen tárgyalom a coarse graining feladat és az inverz módszerek közti kapcsolatot. Az inverz problémához hasonlóan coarse graining esetében is csak akkor beszélhetünk két szimulációs modell azonosságáról, ha mind a három faktort, a szerkezetet, a dinamikát és az energetikát is azonosnak kapjuk. Szándékosan nem használom a termodinamika kifejezést itt, mert valójában csak a potenciális energiát vizsgáltam az eddig tárgyalt cikkekben, illetve a kinetikus energiát valamennyire a konfigurációs hőmérséklet segítségével.

4.3. táblázat A kétkomponensű Lennard-Jones rendszerek adatai és összehasonlításuk az egykomponensűre redukált rendszerekkel, ha azok párpotenciálja energia alapon (PM módszer) vagy szerkezet alapon (IT [Schommers 1973]) lett meghatározva. c a nagy részecskét, s a kicsit jelenti, ez utóbbi nem szerepel az egykomponensű szimulációkban. [s36]

sorszám	parciális LJ ε paraméterei		potenciális energia eloszlása a parciálisok között %			a konfigurációk potenciális energiáinak korrelációs együtthatója a különböző módszerekkel		
	$\varepsilon_{cc}/\varepsilon_{ss}$	$\varepsilon_{cs}/\varepsilon_{ss}$	E_{cc}	E_{ss}	E_{cs}	PM-binary	IT-binary	PM-IT
1	1	1	3	63	34	0.30	0.13	0.45
2	5	2.24	10	35	55	0.48	0.30	0.73
3	10	3.16	18	24	58	0.94	0.91	0.97
4	10	1	44	45	11	1.00	1.00	1.00



4.18. ábra A megtartott (nagyobb) részecskék párkorrelációs függvénye az eredeti kétkomponensű rendszerben és az egyszerűsített egykomponensű rendszerben PM módszerrel (pot.match) vagy szerkezet alapú inverz módszerrel (inv. method [Schommers 1973]) levezetett párpotenciál alapján. Fent az 4.3. táblázat 1. rendszerére, Lent a 4. rendszerére. [s36]

4.5 Az inverz probléma későbbi vonatkozásai

A 2007-es összefoglaló cikkem utáni kutatási irányok [s37]

Az inverz problémával kapcsolatos munkáimat és utóéletüket érdemes a 2007-ben írt áttekintő cikkem [s37] szerkezetében értékelni, röviden említve a különböző megközelítések mai helyzetét is.

A Henderson-féle elmélettel kapcsolatban nem csak az összefoglaló munkában tárgyalt cikkekben foglalkoztak, hanem későbbi elméleti munkák is születtek annak igazolására és egyes iteratív módszerek eseténben a konvergenciára. A 2010-es éve második felében a téma egyik aktív elméleti kutatója Martin Hanke lett [pl. Rosenberger et al. 2016, Hanke 2017].

A coarse graining és a multiscale szimulációk megjelenésével felpörgött az inverz módszerek gyakorlati alkalmazása. Coarse graining szimulációk esetében általában egy első közelítésnek, vagy egy a megmaradó részecskék direkt kölcsönhatását modellező módszernek használják az inverz módszereket. Leggyakrabban az IMC módszert és az eredetileg Schommers által kifejlesztett majd Soper által továbbfejlesztett módszereket használják, ez utóbbi családot inverse Boltzmann iteration néven szokták emlegetni, IBI rövidítéssel. Az általam fejlesztett módszerek közül néhányat újra felfedeztek, pl. BGY módszerünkkel teljesen azonos bemutatás történt a miénkre rá néhány évre. Hanke és munkatársai pedig újra kidolgozták az IMC általunk továbbfejlesztett változatait, ugyanazon célból, hogy az eredeti IMC-nél stabilabb iterációt kapjanak, ugyanúgy a regularizált Gauss-Newton-Marquardt paraméterbecsléssel [Rosenberger et al. 2017]. Ugyanígy elkezdtek egy korábbi integrálegenletes folyadék elméleten alapuló módszer újrafeledzését, itt a témát újragondolták, amikor az összefoglaló cikkemben megtalálták a régi módszert. Kialakult egy európai tudományos centrum a Darmstadt-Mainz régióban, ahol a coarse graining problémakör minden területével, egymással szorosan összefogva dolgoznak olyan nemzetközileg ismert kutatók, mint Friederike Schmied, Jürgen Gauss, Nico van der Vegt.

Mint már írtam, az inverz problémával való kutatásokat 2011-ben felfüggesztettem, mivel a területen szinte semmi visszajelzést nem kaptam, látszólag senkit se érdekelt akkor ez a terület. Az utolsó munkát Bősz Attila szakdolgozó kezdte el. A potenciális energia illesztés sémáját dolgoztam át úgy, hogy párhuzamosan, vagy váltakozva történt volna eredetileg kétkomponensű rendszer megmaradó típusú részecskéjére a kölcsönhatási potenciál iteratív finomítása három célfüggvény irányában, amelyeket teljes kétkomponensű szimulációkból kaptunk: $g(r)$, E és sebesség autokorrelációs függvény, vagyis szerkezet, energetika és dinamika párhuzamos illesztésével. A szakdolgozati munkában nem sikerült befejezni a szimulációkat. A korábbi kutatásaim az inverz probléma és a coarse graining egyik mai legismertebb USA-beli fejlesztőjét (William Noid) megihlette, és egy több cikkes PhD dolgozat készült a kutatásokból, egy olyan módszert létrehozva, amit Bősz Attilával szerettünk volna elérni. Az [s36]-os publikációban bevezetett célfüggvényemet „energy-matching variational functional”-nak nevezi és egy

azóta is rendszeresen használt módszerük alapjának tekintik. Korábbi eredményeimet és sejtéseimet részletes vizsgálatnak vették alá és kiterjesztették [Lebold&Noid 2019a, 2019b]. A coarse graining ilyen megközelítésének egyik legjobb összefoglalója mostanában jelent meg [Noid et al. 2024].

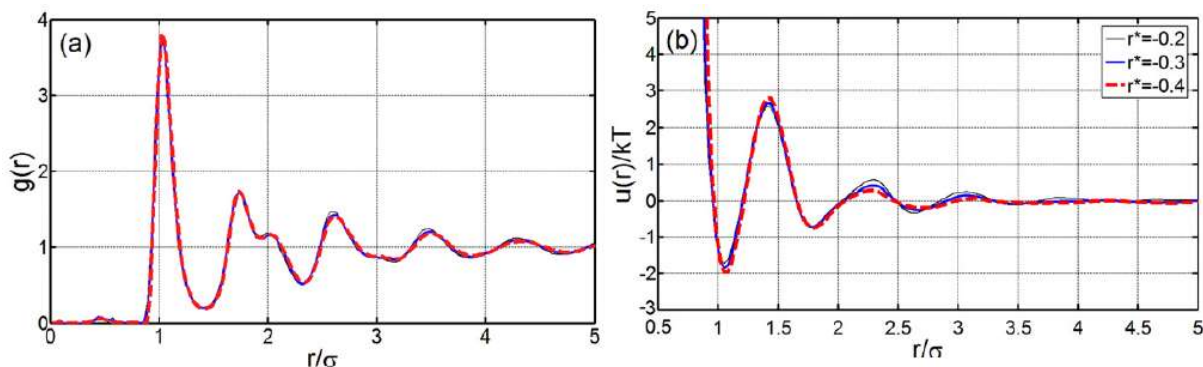
A neurális hálós megoldás egy másik iránya ma egy teljes kutatási irányt ad, az ún. gépi tanulásos potenciálokat kvantumkémiai számolások egyszerűsítésére, illetve kiegészítőjeként sokan használnak. De történtek direkt alkalmazási próbálkozások is az inverz problémára, pl. a már említett közelítése a híd függvénynek [Godall&Lee 2021]. Kampa 2025-ös cikkében számos hasonló próbálkozás bibliográfiája található meg [Kampa et al. 2025].

A coarse graining témakörnek azonban csak egy része az inverz probléma. A mai kutatások homlokterében az általánosított Langevin egyenlet Mori-Zwanzig féle formalizmusában levő megoldásokat keresik. A kulcs szó itt az ún. memóriefüggvény meghatározása. Király Péterrel 2020-ban részt vettünk online egy bécsi konferencián, aminek a témája pusztán ez a memória függvény volt. Poszterünkéből végül cikket nem készítettünk, mivel az általunk kemometriai módszerekkel minden részecskét tartalmazó (all atom) szimulációkból meghatározott memóriefüggvény nem szolgáltatott alkalmazható minőségű eredményeket. Ugyanakkor sikerült kidolgoznunk, hogy milyen információkat kell a coarse graining során időfüggően tárolni, hogy arra valamilyen numerikus vagy gépi tanulásos módszerrel majd memóriefüggvényt illeszthessünk. Ennek a rövid visszatérésnek az inverz elmülethez/coarse graininghez volt a következménye, hogy 2022-ben ötletadója és társzervezője lettem a „Szimulációk több skálán időben és térben” veszprémi workshopnak.

Egy alkalmazás: párkölcsönhatás meghatározása fázismezőkristály-modellezés eredményeire [s43] [s45]

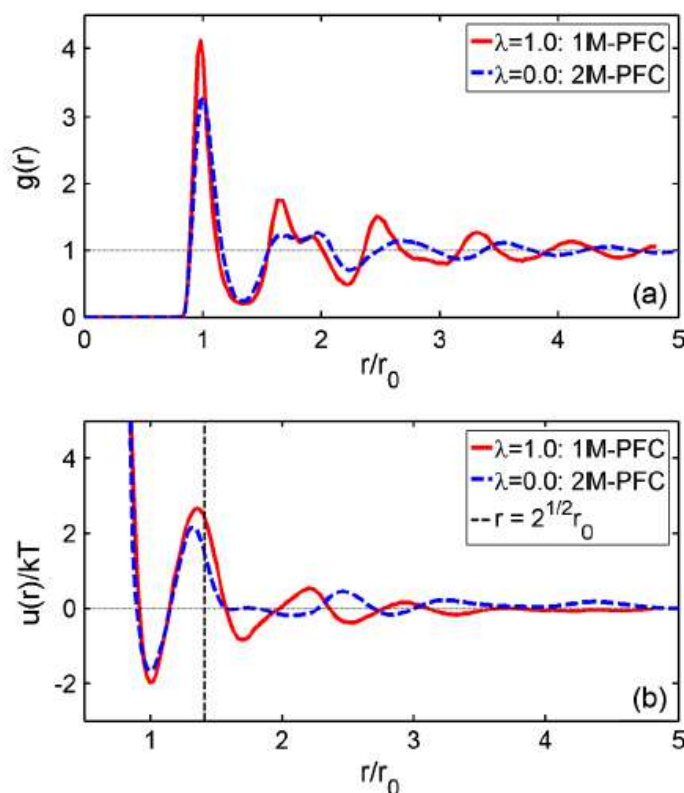
A kvantummechanikai sűrűségfukcionál-elméletnek léteznek klasszikus mechanikai analógjai. Gránásy László csoportja dinamikus szimulációkat végzett az ún. fázismezőkristály-modellezéssel (phase-field crystal model, PFC). Számos rendszert vizsgáltak, az évek során atomi felbontástól egészen mezoskopikusig. 2010-ben a polimorfiát, a kristály nukleációt és a kristálynövekedést szimulálták diffúzió kontrollált rendszerekben, mind két- és háromdimenziós esetekre. A szimulációik során a kialakuló kristályok és azok időbeli fejlődését követték, fázisdiagramokat készítettek, meghatározták a fázisok kialakulásának feltételeit is. A kristályokra számolható volt a párkorrelációs függvény, ami különös figyelmet kapott az amorf szerkezetek és a kristályosodás prekurzorainak esetére. Egyik alapkérdés az volt, hogyan magyarázható ezeknek a prekurzoroknak a jelenléte. Irodalmazáskor bukkantak rá a lehetőségre, hogy párkorrelációs függvényből effektív párpotenciál határozható meg. Az effektív potenciál a szerkezetet visszaadhatja akár tökéletesen is, ha csak párkölcsönhatásokkal leírható a rendszer, de csak egyensúlyi rendszerekre és párkölcsönhatásokkal leírható rendszerekre igaz ez elméletileg. Két munkájukban vállaltam el részfeladatot. Az inverz problémát megoldó

Schommers-féle módszerre [Schommers 1973] programot írtam a számukra, amellyel néhány ezer részecskét tartalmazó Monte Carlo szimulációk segítségével iteratív módon határoztuk meg az adott párkorrelációs függvényekhez tartozó effektív párpotenciálokat. Hozzájárulásom a két publikációban az inverz problémát megoldó program elkészítésére és azokkal rövidebb próbaszámolások végzésére korlátozódott. Mindkét cikkben szerepeltek inverz módszerrel meghatározott párpotenciálok, az első esetben különböző redukált hőmérsékletű rendszerek lett összehasonlítva (4.19. ábra).



4.19. ábra. Bal oldal: 3D PFC szimulációk párkorrelációs függvénye három különböző szimulációs paraméterrel kapott amorf szerkezetre. Jobb oldal: a Schommers-féle módszerrel kapott párkölcsönhatások a három $g(r)$ alapján. [s43].

Ebben az esetben az effektív párpotenciálokban minimális eltérések voltak, de a függvények alakját a szerzőtársaim összevetették olyan kölcsönhatási klasszikus potenciálokkal, amelyek ismertek üvegképző jellegükről. A második esetben az úgynevezett egymódusú és kétmódusú fázismezőkristályszimulációkat hasonlítottuk össze. A kapott effektív kölcsönhatásnál jól azonosítható volt a $\sqrt{2}r_0$ értéknél levő csúcs, ami a más potenciálokkal való összehasonlítások alapján nagy valószínűséggel felelős az amorf prekursorok létrejöttéért, mivel a csúcs jelenléte gátolja a szorospakolású (FCC, HCP) kristályszerkezetek létrejöttét (4.20. ábra). Még egyszer hangsúlyozom, hogy ezeknél a kutatásoknál nem vettem részt a PFC szimulációkban, érdemi hozzájárulásom az inverz probléma megoldásánál használt szoftver elkészítése és tesztelése volt.



4.20. ábra. Egy- és kétmódusú PFC modellezéssel kapott amorf szerkezetek. Fent: párkorrelációs függvények. Lent: a párkorrelációs függvények alapján meghatározott párpotenciálok. [s45]

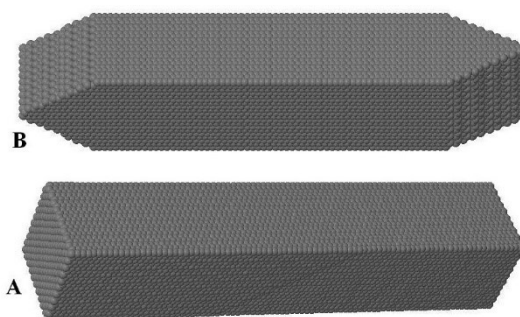
4.6 Lépések a kinetikai Monte Carlo irányában

Karbamid kristályosodása [s39]

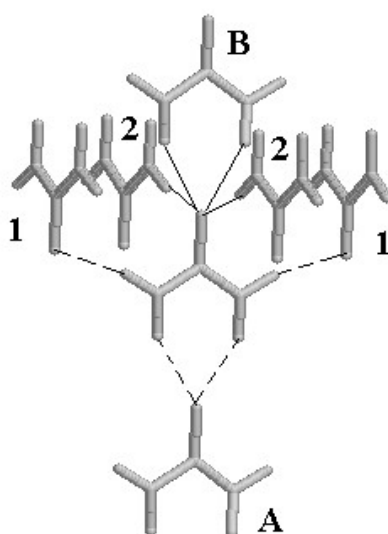
A coarse graining témakör szorosan kapcsolódik a számítógépes kémia egyik fő problémájához, hogy a minél egzaktabb leírásra törekvéshez nagyon pontos és részletes elméleti módszereket kell használnunk, míg a jelenségek leírásához egyre nagyobb rendszert. A kutatói hozzáállás egyik legfontosabb kérdése ezen a területen, hogy az adott kérdésre milyen módszerrel és mekkora rendszerrel, illetve szimuláció esetén milyen időtartományban keressük a megoldást. A 2000-es évek közepén akadtak kezembe Piana és Gale közleményei [Piana&Gale 2005, Piana et al. 2005, Piana&Gale 2006], ahol az idő és méret skála átlépését úgy oldották meg, hogy molekuláris dinamikai szimulációkkal karbamid kristályok képződését vizsgálták és kinetikai sebességi állandókat határoztak meg különböző karbamid kristályfelületek vizes és metanolos oldataiban a kristályosodás és az oldódás folyamataira. A következő lépésben ezeket a kinetikai sebességi állandókat kinetikai Monte Carlo (KMC) szimulációkban használták, ahol sikeresen reprodukáltak a makroszkopikus kristályformákat. A 3.2 fejezetben részleteztem azokat a kísérleteimet, ahol diffrakciós adatokból összetett részecskék belső szerkezetére próbáltam meg következtetéseket levonni. A fényszóródási mérések és a kisszögű

mérések esetén elmondhatom, hogy egy nem túl részletes mérési információ alapján próbáltam meg 3D szerkezeti információkat kinyerni, pontosabban a mérési görbéknek nem ellentmondó szerkezeteket generálni. Piana és Gale munkáit olvasva felmerült bennem, hogy meddig csökkenthető a bemenő információ. Lecsökkenthető-e ez annyira, hogy ismert elemi cellájú kristály makroszkopikus morfológiájából visszakövetkeztessünk a kristályképződés mikroszkopikus dinamikájára? Bemutatható-e olyan példa ezen a területen, ahol a Henderson analóg kapcsolat nem más, mint egy makroszkopikus kristály alakja, a kölcsönhatás pedig egy sebességi állandó készlete a kristályosodás folyamatainak különböző felszíneken?

A kérdést a Piana és Gale által vizsgált karbamid kristályosodáson vizsgáltam. Vízben és metanolban két jellegzetes, de eltérő kristályforma képződik, ugyanazzal az elemi cellával (4.21. ábra). Az elemi cellája ismert a kristálynak, minden karbamidot 6 másik vesz körül (4.22. ábra). Első szomszéd közelítéssel és a szimmetriákat is figyelembe véve ez 36 különböző lehetőséget jelent arra, hogy egy adott karbamidnak melyik szomszédos pozíciójában található egy-egy karbamid molekula vagy az adott pozíció betöltetlen.

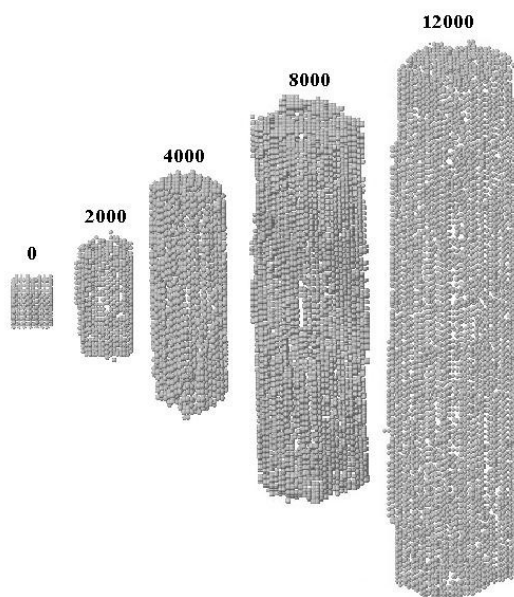


4.21. ábra. Karbamid kristály sematikus alakja A) vízben B) metanolban. [s39]



4.22. ábra. Kristályos karbamidban a legközelebbi 6 másik molekula elhelyezkedése a hidrogénkötéseket is jelölve. [s39]

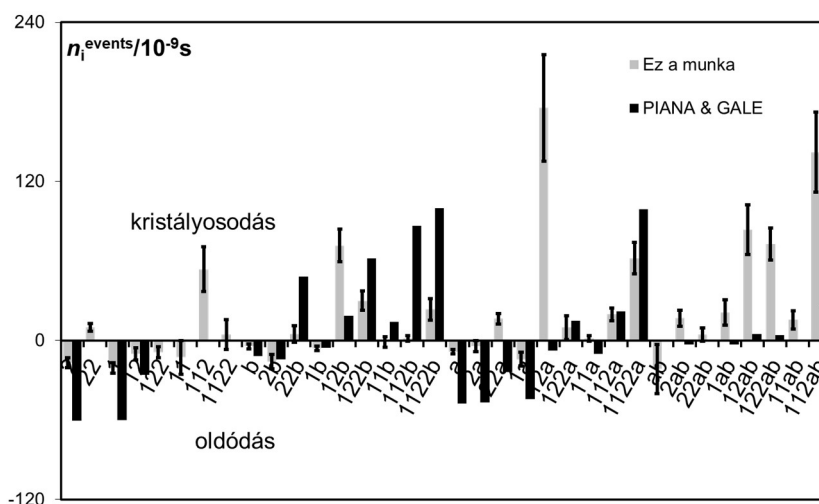
A kinetikai Monte Carlo szimuláció során egy nagy szupercellából indultam ki, amelyet az elemi cella többszörös eltolásával hoztam létre. A cella közepén pár pozíciót betöltöttnek jelöltem, a többit üresnek. Ezek után a KMC elveinek megfelelően a kinetikai sebességi állandók arányában változásokat generáltam véletlenszerű pozíciókban. Annyi megszorítást tettem, hogy mind oldódási mind kristályosodási folyamatot csak a már létező góc határfelületén tettem lehetővé. Egy növekedési példa a 4.23. ábrán látható.



4.23. ábra Kristálynövekedés a KMC szimulációban. Egy-egy gömb egy-egy karbamid molekulát jelöl. A szimuláció során az általunk vizes oldatra meghatározott kinetikai paramétereket használtuk. Az időskála önkényes és a jobb láthatóság miatt csak a felületi gömbök láthatóak. [s39]

A szimulációk célja az volt, hogy csak a makroszkopikus szerkezet alapján olyan kinetikai állandó készletet határozzak meg, amely jól reprodukálja a makroszkopikus formákat. Az első körben véletlen kinetikai állandó készlettel szimuláltam, majd ezt a készletet genetikus algoritmus segítségével optimalám olyan irányba, hogy a KMC szimulációban kapott kristály minél jobban hasonlítson a makroszkopikus formához. Az algoritmus technikai részleteit itt most nem sorolom fel, a lényeg, sikerült olyan kinetikai sebességi állandó készleteket létrehoznom, amelyek visszaadták a makroszkopikus formát. Piana és Gale munkája bemutatta, hogy klasszikus molekuláris dinamikai szimulációból elindulva a reális kinetikai állandók kaphatóak, amelyek helyes makroszkopikus kristályformát adnak. Pontosabban, még korábbról indultak, hiszen a klasszikus mechanikai MD szimulációk potenciáljainak egy részét kvantumkémiával határozták meg. A fordított irányból indulva az én kutatásom eredménye ugyanilyen kinetikai állandóknak a meghatározása volt a makroszkopikus

forma felől indulva. A két megközelítés hasonló jellegű kinetikai állandókat szolgáltatott. Az 4.24. ábrán egy adott pozícióra a kristályosodási és az oldódási folyamat állandóinak a különbsége szerepel, mivel az én megközelítésemben részben nem szerepelt semmilyen a valósághoz skálázható időfaktor illetve kontroll az ellentétes folyamatpárok két részére külön-külön.



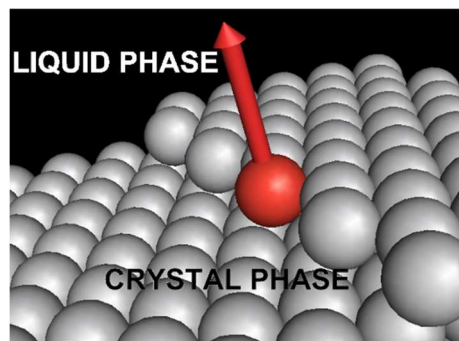
4.24. ábra Karbamid kristályképződési és oldódási sebességi állandói közti különbség különböző felületi környezetekben. A Piana és Gale által MD szimulációkban meghatározott állandók és a makroszkopikus morfológia alapján KMC szimulációkkal meghatározott értékek, ez utóbbiaknál a hibaszáv a genetikussal kapott párhuzamos értékek terjedelmét mutatja. [s39]

Szubjektív értékelésem szerint ez az egyik kedvenc munkám, a feltett kérdés és az arra kapott válasz, valamint a megoldás módjára máig is büszke vagyok. Egy elvi kérdésre tudtam egy numerikus választ adni.

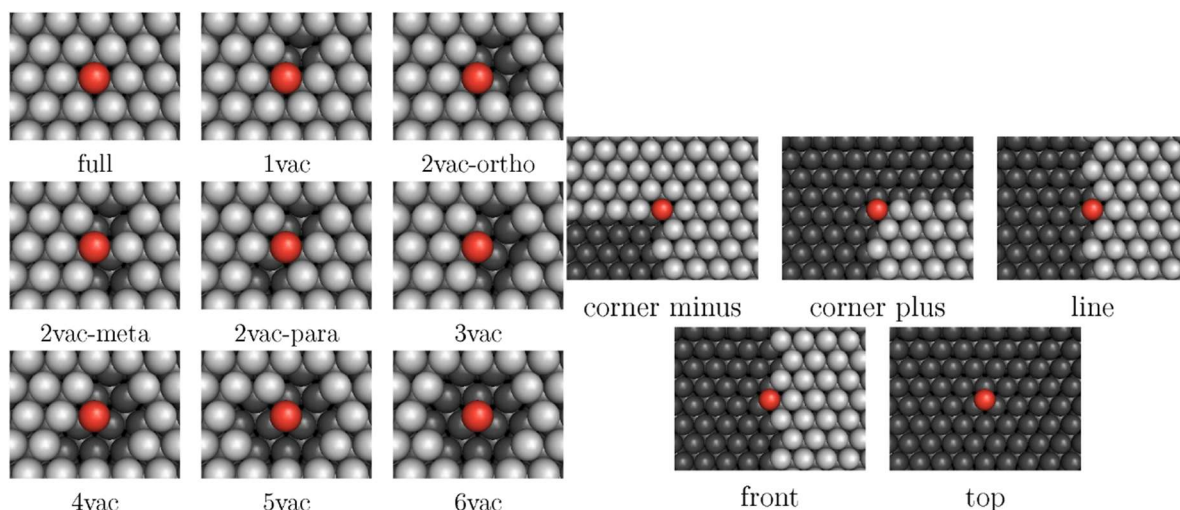
Argon kristályosodás sebességi állandói [s46]

Ritka események esetén a Piana és Gale által használt klasszikus molekuláris dinamika időskálája nem elegendően hosszú kellő számú átmenet szimulációjához kinetikai paraméter kinyerésére. Ilyenkor célszerű valamilyen a ritka eseményekhez kapcsolható módszert használni, pl. valamilyen szabadenergia meghatározására kifejlesztett. Ezeknek a módszereknek elterjedése adta az ötletet, hogy megpróbáljam pl. Lennard-Jones kölcsönhatással modellezett argon kristályosodásának teljes kinetikai paraméterkészletét ilyen módon meghatározni. Meg kell jegyeznem, hogy a téma valamennyire feldobott labda volt, párhuzamosan a Kozma Dániel szakdolgozóval végzett munkánkkal két neves kutatói csoport is meghatározott sebességi állandókat kristályosodásra, de egyik esetben se olyan teljes készleteket, amelyek aztán mezoszkopikus KMC szimulációkban használhatóak lettek volna.

Különbözö mikroszkopikus felületi konstrukciókban vizsgáltuk az argon kristály/folyadék határfelületet. Metadinamika [Laio&Parrinello 2002] segítségével egy-egy kristályos (kötött) atomot távolítottunk el különböző kristálypozíciókból (4.25. és 4.26. ábrák).



4.25. ábra Argon kristályosodási sebességi állandóinak meghatározásához végzett metadinamikai szimulációk alapötlete. A metadinamika során a segédpotenciál fokozatosan lökte ki a kiszemelt részecskét a kristályból. [s46]



4.26. ábra. A metadinamikával vizsgált különböző elrendeződések és azok angol elnevezése. [s46]

A kapott szabadenergiagát magasságai alapján sebességi állandókat számoltunk az átmenetiállapot-elmélet segítségével. Tudtuk, hogy az általunk használt sűrűség/hőmérsékletértékek mellett a prefaktorra (transzmissziós koefficiensre) durva közelítés az átmenetiállapot-elmélet ilyen összetett folyadék/szilárd rendszer esetén. Ennek ellenére relatív értelemben teljesen jó közelítés, mivel nem várható, hogy lényegesen különbözne a prefaktor a különböző szerkezeti helyekre. A kapott sebességi állandókat a 4.4. táblázat tartalmazza, ahol a kristályosodás hajtóereje a folyadék és a szilárd fázis hőmérsékletkülönbsége a koegzisztencia vonal két oldalán. Az állandók dimenziója $(\text{mol}^{-1}\text{s}^{-1})^a$ hipotetikusán feltételezett unimolekuláris reakciónak megfelelően szerepel (4.4. táblázat).

4.4. táblázat $\Delta T=5$ K által kristályosodási irányban indukált sebességi állandók. A pozíciókban a 4.26. ábra jelölései szerepelnek. [s46]

	$k_{\text{krist}} T=82 \text{ K}$	$k_{\text{old}} T=77 \text{ K}$
	($\text{mol}^{-1}\text{s}^{-1}$)	($\text{mol}^{-1}\text{s}^{-1}$)
full	$9,61 \cdot 10^{32}$	$5,73 \cdot 10^{32}$
1 vacancy	$1,21 \cdot 10^{33}$	$7,33 \cdot 10^{32}$
2 vac.-orto	$1,99 \cdot 10^{33}$	$1,25 \cdot 10^{33}$
2 vac.-meta	$1,99 \cdot 10^{33}$	$1,25 \cdot 10^{33}$
2 vac.-para	$1,99 \cdot 10^{33}$	$1,25 \cdot 10^{33}$
3 vac. side by side	$2,56 \cdot 10^{33}$	$1,63 \cdot 10^{33}$
4 vac. side by side	$3,11 \cdot 10^{33}$	$2,00 \cdot 10^{33}$
5 vacancies	$4,86 \cdot 10^{33}$	$3,22 \cdot 10^{33}$
6 vacancies	$8,20 \cdot 10^{33}$	$5,63 \cdot 10^{33}$
corner minus	$2,42 \cdot 10^{33}$	$1,53 \cdot 10^{33}$
line	$6,18 \cdot 10^{33}$	$4,16 \cdot 10^{33}$
front	$8,23 \cdot 10^{33}$	$5,65 \cdot 10^{33}$
corner plus	$5,37 \cdot 10^{34}$	$4,17 \cdot 10^{34}$
top	$5,32 \cdot 10^{34}$	$4,12 \cdot 10^{34}$

A koefficiensek segítségével ezek után KMC szimulációt végeztünk, amellyel szoros pakolású, de nem karakterisztikus morfológiájú argon kristályokat kaptunk.

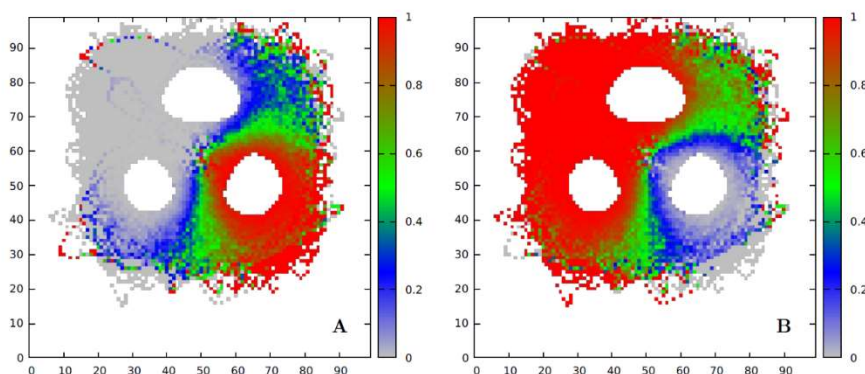
Több minimumú rendszerek kommittor függvényei [s54]

A kinetikai Monte Carlo alapfeltevése, hogy a rendszerünkben diszkrét állapotokat definiálunk és ezen állapotok között sebességi állandókat. A diszkrét állapotok definiálása lehet szemléletes, pl. egy reakció kiindulási vagy termék állapota, vagy az átmeneti állapot. Egy összetettebb rendszernél más lehetőségek is vannak, pl. egy fehérje szimulációjakor a szerkezetet feloszthatjuk idő alapon is, egy hosszabb szimulációban akár több ezer állapotot is definiálva [Meng et al. 2016]. Amennyiben sikerül reális sebességi állandókat meghatározni az állapotok között, akár elfogadva azt is, hogy számos esetben ez nulla, akkor áttérhetünk egy kinetikai Monte Carlo megoldásra. Ilyen módon hosszabb időigényű események (pl. fehérje feltekeredése) is reálisan szimulálható KMC segítségével. A téma elméleti hátterét többen is részletesen kidolgozták, mi a Vanden-Eijden féle elméletet [Vanden-Eijden 2006] használtuk egy következő munkánk során, pontosabban annak diszkretizált változatát, a DTPT (Discrete Transition Path Theory) elméletet [Metzner et al. 2006, Metzner et al. 2009].

A hagyományos $A \rightarrow B$ reakció sémában gyakran előfordul egy köztes átmeneti állapot megjelenése. Ha ez kellően stabil, akkor tekinthetjük a rendszerünket úgy is, hogy három energiaminimummal rendelkezik a rendszer. 2017-ben kezdtünk bele egy kutatási témába, ahol ilyen több minimumú rendszereknek a DPTP elméletét vizsgáltuk. A témához több alapfogalmat tisztázni kellett. Egy ilyen fogalom, hogy mit is tekintünk diszkrét állapotnak. Ez lehet maga a termék, reaktáns vagy átmeneti állapot is, de lehet egy akár idő vagy térbeli diszkrétizálása a rendszer által bejárható fázistérnek. Ez utóbbi esetben ezeket az állapotokat csoportosíthatjuk reaktáns, termék, átmeneti osztályok szerint. A másik alapfogalom a kommittor, ami azt mutatja, hogy egy adott állapotból induló trajektóriák milyen valószínűséggel vezetnek adott osztályba. Többnyire ezt úgy számolják, hogy a konfigurációs tér alapján diszkrétizált rendszer adott állapotából véletlen sebességekkel indítják el a szimulációt és számolják az termékbe vezető trajektóriák arányát, amennyiben csak két osztály van jelen, a reaktáns és termék osztályok. A témához szükséges még a rendparaméter, a kollektív változó és a reakciókoordináta definiálása. A rendparaméter csak az adott osztály megkülönböztetésére szolgál, pl. reaktáns vagy termék jellegű-e az adott pont. A kollektív változó egy olyan kombinációja vagy abból számolható tulajdonságok kombinációja bármilyen fázistérfüggő tulajdonságnak, amely segítségével megkülönböztethetők az osztályok és esetlegesen egy trajektória is ábrázolható. Ezek akár vektorváltozók is lehetnek. A reakciókoordináta egy olyan többnyire kollektív skalár vagy vektorváltozó, ami összekapcsolható a reakció előrehaladásával is. A reakciókoordináta teljesítőképessége kommittor analízissel végezhető el, ideális reakciókoordináta szerint a kommittor függvény unimodális és az átmeneti állapotban az értéke 0.5 [Dellago&Bolhuis 2009, Peters 2017]. A munkánk első részében egy háromminimumú egyrészecske rendszert vizsgáltunk Langevin dinamikai szimuláció segítségével. Feltételes valószínűség segítségével definiáltuk egy több minimumú rendszer egyedi állapotokra és egyedi reakcióra vonatkozó előremenő kommittor függvényeit a B állapotra az i -dik felosztási pontról:

$$q_{B,i}^+ = \begin{cases} \frac{\sum_{\alpha} n_{\alpha B,i}}{\sum_{\alpha\beta} n_{\alpha\alpha,i} + n_{\alpha\beta,i}}, & \forall i \in \Gamma \setminus (A \cup B \cup C) \\ 0, & \forall i \in (A \cup C) \\ 1, & \forall i \in B \end{cases} \quad 4.15. \text{ egyenlet}$$

, ahol $n_{\alpha\beta,i}$ pl. a száma az összes beütésnek az i -dik felosztási pontban amelyik része egy $\alpha \rightarrow \beta$ trajektóriának és Γ jelöli az összes felosztási pontot. Az így definiált kommittor értékek jól ábrázolhatóak:

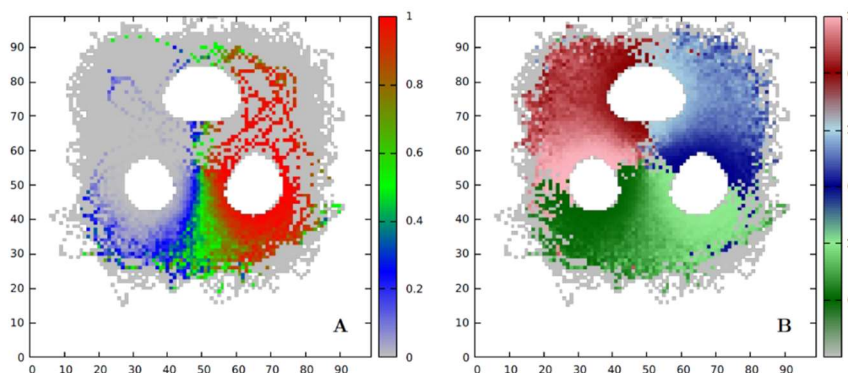


4.27. ábra. Három minimumú rendszer kommittorjainak ábrája. A három minimum osztálya a három fehér folt. A) $q_{B,i}^+$ előre menő kommittor B) $q_{B,i}^-$ hátra menő kommittor. [s54]

Az így definiált előre menő és hátra menő kommittor értékeket felhasználva, a három minimumú (A, B, C) rendszerben feltételes valószínűséggel definiáltuk az A→B átmenethez tartozó kommittort:

$$q_{AB,i}^+ = \frac{q_{B,i}^+}{q_{C,i}^+} = \frac{q_{B,i}^+}{q_{A,i}^+ + q_{B,i}^+} \quad 4.16. \text{ egyenlet}$$

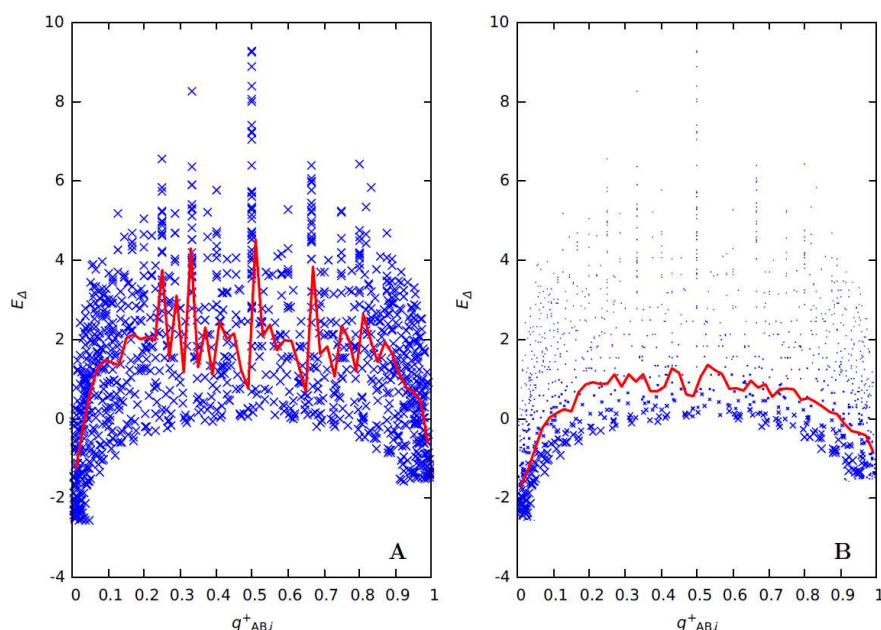
Bemutattuk, hogyan lehet több ilyen feltételes reakció kommittort szemléltetni, illetve akár egy ábrán is bemutatni a három előre menő kommittorra szűrési szabályok alkalmazásával (4.28. ábra).



4.28. ábra. A) $q_{AB,i}^+$ reakció kommittor B) $q_{AB,i}^+$, $q_{BC,i}^+$ és $q_{AC,i}^+$ előre menő reakció kommittorok közös ábrázolása [s5]

Levezettük azokat a köztes mennyiségeket, amelyek segítségével a reakció kommittor és a trajektóriák figyelembevételével számolható az útvonalak valószínűségi sűrűségi eloszlása. Külön foglalkoztunk azzal is, hogy egy átmenet során a megszokott minimális energia útvonal helyett milyen súlyozott

formán mutatható ez be, miképpen készíthető informatív ábra egyszerre a minimális energia útvonalról, az útvonalak átlagos energetikájáról és az egyedi trajektóriák sűrűségéről (4.29. ábra).



4.29. ábra. Az $A \rightarrow B$ részreakció trajektóriáinál bejárt diszkrét állapotok energiaeioszlása a kommittor értékek függvényében. A) súlyozatlan eset B) súlyozott eset az eredeti közleményben $m_{t,AB,i}$ -vel jelölt sűrűség figyelembevételével. A piros vonal az adott ábra átlagos értékeit mutatja. A legelső pontok összekötésével lehetne a minimális energiájú útvonalat szemléltetni. [s54]

A cikk rengeteg színes ábrát tartalmaz a különböző esetekre, pl. néztük, hogyan befolyásolja az ábráinkat, ha pl. a C állapotot csak átmeneti állapotként kezeljük, illetve egy valós molekula, a 2-klór-bután esetében annak 3 minimumára is alkalmaztuk a számolásokat.

A munka fő érdemének azt tartom, hogy a feltételes valószínűségek segítségével sikerült definiálnunk a több minimumú rendszerekre a szükséges elméleti egyenleteket, amelyek mindenképpen szükségesek a mai reakciórendszerek kinetikai Monte Carlo rendszerű vizsgálatához. Mindemellett több olyan ábrázolást is kifejlesztettünk, amelyek jól áttekinthetően mutatják egy rendszer átalakulási lehetőségeit a kollektív változók, esetleg reakciókoordináták függvényében.

Ennél az alfejezetnél nem tértem ki néhány olyan kutatásra, amelyekből nem készült nemzetközi publikáció, de esetleg pl. szakdolgozat formájában azért foglalkoztunk egy-egy ötlettel. Ilyen kísérlet volt a kalcit különböző morfológiáival kapcsolatos próbálkozás, ahol végül nem jutottunk el olyan teljes sebességi állandó készlethez, ami KMC szimulációban használható lett volna [Irsik 2010]. Víz párolgási mechanizmusának vizsgálatát is elvégeztük több vízmodellre [Kiss 2013], illetve számoltuk a párolgás

toth.gergely_304_24

szabadenergiaváltozását [Kohut 2014]. Daru János pedig módszert dolgozott ki a kommittor térképek egy effektív számolására és reprezentálására Gauss függvények segítségével [Daru 2015].

5. Kemometria:

5.1 Szeriálás [s41] [s53] [s63]

Egy a Kémiai Intézetet érintő belső költözés során Szepesváry Pál nyugdíjas kollégával kerültem közös szobába. Pali bácsi a magyar kemometriai közélet doyenje volt abban az időben, így nem csak a szakterület tanulhattam meg tőle, hanem viszonylag hamar aktív résztvevője lehettem a hazai kemometriai közösségnek. Szepesváry Pállal való közös munkáink szinte mindig ugyanúgy kezdődtek: egyikünk megpróbált valami érdekeset elmagyarázni, amit a másik nem értett. A rosszul sikerült magyarázat vagy félreértés elindított egy gondolat kísérletet, néha egy-egy kisebb numerikus kísérletet, amelyeknek az eredménye több esetben is cikk lett.

Közösen hallgattunk meg egy előadást, ahol az Elbából vett vízminták analitikai kémiai eredményeit mutatta be egy német vendég. Az előadás után jött számunkra az ötlet, hogy az eredetileg folyamkilométerek szerint rendezett táblázatot nem lenne-e érdemes inkább a folyó melletti ipar szerint csoportosítani és miképpen lehetne egy ilyen átrendezését az adatmátrixnak matematikai formába önteni. Az első cikkünket úgy publikáltuk, hogy nem ismertük az ezzel az adatrendezési feladattal kapcsolatos terminus technicust, a szeriálást. 2009-ben megjelent cikkünk [s41] után 1-2 évvel bukkantunk rá, hogy maga a feladat a tudomány számos területén önálló módszereket eredményezett és a kérdés szeriálás néven ismert. Az általunk javasolt megközelítést azonban a számos módszer ellenére senki se alkalmazta, ez vonatkozott mind arra az adatszerkezetre, amelyet mi megcéloltunk, mind az általunk kidolgozott matematikai rendezőelvre. Első cikkünk után 8 évvel újra elővettem a témát és számos új, kémiai szempontból érdekes adatra vizsgáltuk meg a módszert [s53], míg 2023-ban leközlöttünk egy újabb rendezőelvet, amelyet már kettőnél több dimenziós adattömbökre is alkalmaztunk [s63].

Mi az a szeriálás? Általánosan minden olyan adatrendezési folyamatot így lehet nevezni, ahol az eredeti adatok sorait, oszlopait, vagy kettőnél több dimenziós adattömbök esetében a dimenziókhoz kapcsolódó sorrendet permutáljuk azzal a céllal, hogy az objektumok és változók új sorrendje a táblázatokban és a vizualizált ábrákon segítse azok értelmezését, pl. hasonlóságok, trendek, csoportok, osztályok felismerését. Története a XIX. századig nyúlik vissza, amikor archeológiai táblázatokat rendeztek át [Petrie 1899]. A módszerek első interdiszciplináris áttekintése Liiv nevéhez fűződik [Liiv 2010], aki összefoglalta a különböző területeken elért eredményeket, valamint rendszerezte a módszereket. Bevezette a mód fogalmát (mode), ami azoknak a vektoroknak a számát jelentette, amelyek szerint egy adott adattömb valamilyen módon permutálható/sorba rendezhető volt. Az út (way) fogalmát pedig úgy definiálta, hogy a dimenziója a rendezett adattömbnek, vagy a rendezéskor használt az eredeti adattömbből levezetett tömbnek. Egy hagyományos adatmátrixot, ahol a sorok az

objektumokat, az oszlopok pedig a tulajdonságokat jelentik, permutálhatjuk csak a sorok szerint, csak az oszlopok szerint, vagy mindkettő szerint egyszerre. A táblázat eredetileg kétdimenziós adattömb, így az első két lehetőség egy módú két utas, míg a harmadik két módú két utas rendezést jelent. Ha az eredeti adattömbből távolság vagy korrelációs mátrixot számolunk, ha ugyanazt a sorrendet írjuk elő mind a két tengelyen, akkor egy módú két utas rendezésről van szó. Eltérően is rendezhető sorrend esetén két módú két utas rendezés is lehetséges, annak ellenére, hogy mindkét tengelyen vagy az objektumok, vagy a változók szerepelnek.

Saját szeriáló módszerek és alkalmazásuk [s41] [s53] [s63]

A szakirodalomban számos módszer található, beleértve a biológiai rendezéseknél is használt „biclustering” módszert is [Cheng&Church 2000]. Az általunk kifejlesztett módszerek azonban egyediek mind a célzott adatszerkezetek, mind az objektum és a változótengelyek összekapcsolása szempontjából. Alapfeltevésünk az volt, hogy olyan adatmátrixokat szeretnénk rendezni, ahol az objektumok hasonlóságát egy-egy objektumpárra vagy csoportra néhány tulajdonság határozza csak meg, míg a többi tulajdonság értékei nem relevánsak, azok eltérőek is lehetnek. Célunk az volt, hogy a rendezés után leolvasható legyen, hogy melyik objektumok a hasonlóak és hogy az adott pár vagy csoport hasonlóságát melyik tulajdonságok okozzák. Rendezési módszereink két matematikai részből állnak, az első egy lokális távolság tömb, vagy azzal egyenértékű lokális hasonlóság tömb számolása. A második konstrukció pedig egy célfüggvény, amelyet a lokális távolság/hasonlóság tömb alapján számoltunk ki és az eredeti adattömb megfelelő permutációjával maximálva vagy minimálva jutunk el a szemléletesen vizualizálható, már rendezett adattömbhöz.

Az alábbiakban hagyományos kétdimenziós objektum-változó adattömbre mutatom be, hogyan definiáltuk a lokális távolságot, illetve a lokális hasonlóságot (local distance or similarity matrices), valamint a 2009-ben [s41] és 2017-ben használt diagonális célfüggvényt (diagonal measure) és a 2023-as cikkben szereplő folt jellegű célfüggvényt (patch function). A magasabb dimenziós adattömbökre vonatkozó egyenletek a 2023-as cikkben találhatóak [s63].

Egy adott az $A^{N \times M}$ adatmátrixra a lokális hasonlóságot a következőképpen definiáltuk:

$$l_{ij} = \frac{\sum_{k=i-1}^{i+1} \sqrt{\sum_{l=j-1}^{j+1} (a_{il} - a_{kl})^2}}{2} \quad 5.1. \text{ egyenlet}$$

l_{ij} az adott i -dik objektum (sor) átlagos eukleidészi távolságát mutatja az $i-1$ -dik és $i+1$ -dik objektumoktól, de nem a teljes változó térben, hanem csak a $j-1, j, j+1$ változók terében.

A lokális hasonlóságot a következő egyenletekkel definiáltuk:

$$l_{i,k,j} = \sqrt{\sum_{l=j-1}^{j+1} \left(\frac{a_{kl} - a_{il}}{\text{diff}_{\max,l}} \right)^2} \quad 5.2. \text{ egyenlet}$$

$$s_{ij} = \left(\sum_{k=i-1, i+1} 1 - \frac{l_{i,k,j}}{D_{col,j}} \right) / D_{row,i} \quad 5.3. \text{ egyenlet}$$

, ahol l_{ikj} háromdimenziós skálázott lokális távolság tömb egy eleme, mivel a két szomszédos objektumon külön-külön történik a hasonlóság számítása az 5.2. egyenlet szerint és ezek kerülnek átlagolásra. A különböző osztók és normálók ($diff_{max}, D_{col,j}, D_{row,i}$) segítségével kezeltem le a szélső objektumok és a szélső változók eseteit, valamint a hasonlóság $[0;1]$ intervallumba való skálázását. A pontos definíciójuk az [s63]-as publikációban szerepel.

A hasonlóságra való áttérést azért tettük meg, mert numerikusan stabilabb olyan számokkal dolgozni, amelyek a $[0,1]$ intervallumba esnek, mint a $[0,\infty)$ intervallumba eső távolságokkal. Nem részletezem itt, de több változatot is kidolgoztunk, pl. az előbb említett konstansokkal a táblázat szélső elemeinek és hiányzó adatok kezelésére, speciális esetben a nulla értékeknek a távolság számításból való kihagyására is [s63].

A két első cikkben a D globális célfüggvényt (diagonális mérték) használtuk az alábbi definícióval:

$$D = \sum_i^N \sum_j^M t_{ij} |l_{ij}| \quad 5.4. \text{ egyenlet}$$

, ahol t_{ij} az adott i - j elemnek egy hipotetikus távolságát mutatta a diagonálisától az $N \times M$ mátrixban:

$$t_{ij} = 1 + \frac{\left| \frac{j-0.5}{M} - \frac{i-0.5}{N} \right|}{\sqrt{\frac{1}{N^2} + \frac{1}{M^2}}} \quad 5.5. \text{ egyenlet}$$

A 2023-ban bevezetett foltfüggvény definíciója:

$$P = \sum_{i=1}^{n-1} \sum_{j=1}^m 2(s_{ij}s_{i+1,j})^q + \sum_{i=1}^n \sum_{j=1}^{m-1} 2(s_{ij}s_{i,j+1})^q \quad 5.6. \text{ egyenlet}$$

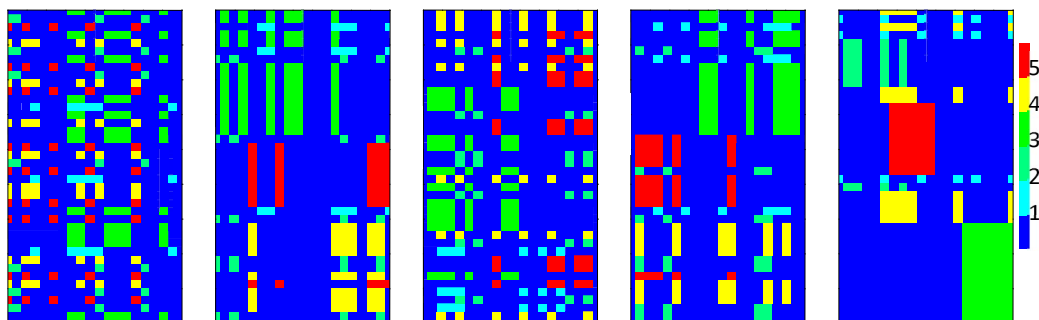
, ahol q empirikus paraméter volt, próbálkozással többnyire 2-3 értékekre állítva.

Lokális távolságmátrix és diagonális mérték használata esetén az utóbbi maximálása vagy minimálása más-más célból készült elrendezést ad. Ha úgy permutáljuk a sorokat és az oszlopokat, hogy a lokális távolságmátrix nagy elemei a diagonálistól minél távolabbra kerüljenek, vagyis D -t maximáljuk, akkor a közeli objektumok egymás mellé kerülnek úgy, hogy a hasonlóságukért az adott diagonális részhez közeli tulajdonságok a felelősek.

Lokális hasonlóság és a P foltfüggvény használata esetén a nagy lokális hasonlóságok, amelyek ugyanazoknál a változóknál szerepelnek egy-egy objektum csoportnál olyan elrendezést adnak P maximálása esetén, hogy az eredeti adattáblázatban foltok alakulnak, ahol egy folt mutatja a hasonló objektumokat és a hasonlóságot okozó változókat. Ezek a foltok tetszőleges pozíciókban kialakulhatnak, nincsenek kötve a diagonális elrendezéshez, mint a D célfüggvény esetén volt.

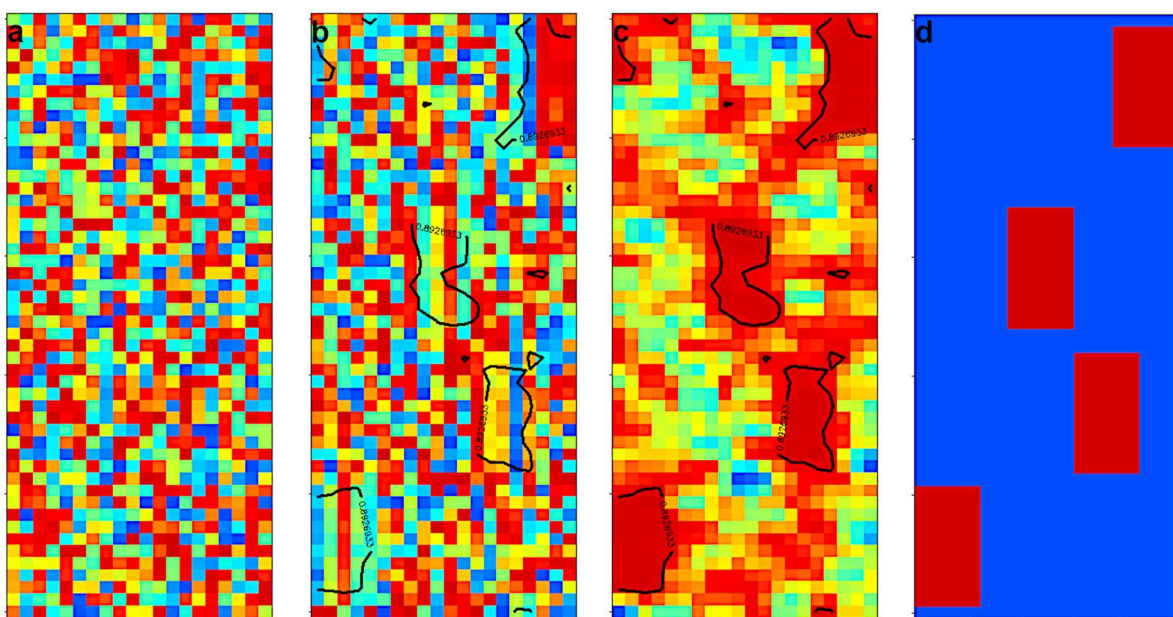
A diagonális módszer rendezési elve és összehasonlítása az egy módú, illetve független két módú rendezésekkel az 5.1. ábrán látható. Szimulált adatsorról van szó, ahol 40 objektum volt 5 csoportba osztva úgy, hogy a 20 változóból néhány karakterisztikusan hasonló volt egy-egy csoporton belül, de a többi változó véletlenszerű volt. Csoportonként más-más változók voltak a jellemzőek. Az ábrán látható egy kiindulási véletlen sor-oszlop permutált eset, a színkód az adott csoporthoz rendelt

változókra azonos. Vagyis a kéktől eltérő színek csak annyit jelentenek, hogy egy adott változó egy adott csoportnál karakterisztikusan felelős a hasonlóságért. Balról a második ábrán egy olyan rendezés található, ahol csak az objektumok sorrendje lett optimálva, a középső ábrán csak a változók sorrendje, míg a következőn parallel mind a sorok, mind az oszlopok, de nincs a két módnál a sorrendiséget összekötő matematikai összefüggés. A jobb oldali ábra mutatja a diagonális rendezést.



5.1. ábra. Diagonális szeriálás szimulált adatsorral. Az ábrán a színkód az adott csoportra jellemző változókat mutatja. Bal oldal: véletlen sorrend, Balközép: objektumok hasonlósága alapján rendezve Középen: változók hasonlósága alapján rendezve, Jobbközép: sor és oszlop permutációval rendezve kapcsolat nélkül a két módra, Jobb oldal: diagonális szeriálás. [s41]

Folt szeriálás esetében is egy szimulált adatsorra mutatom be a módszer teljesítőképességét, itt 50 objektum volt 20 változós térben, 4 csoportba tartoztak az objektumok 5-5 karakterisztikus változóval, illetve 10 objektum nem tartozott egyik csoportba se. Jól látható, hogy a szeriálás tökéletesen megtalálta a négy csoportot (5.2. ábra).

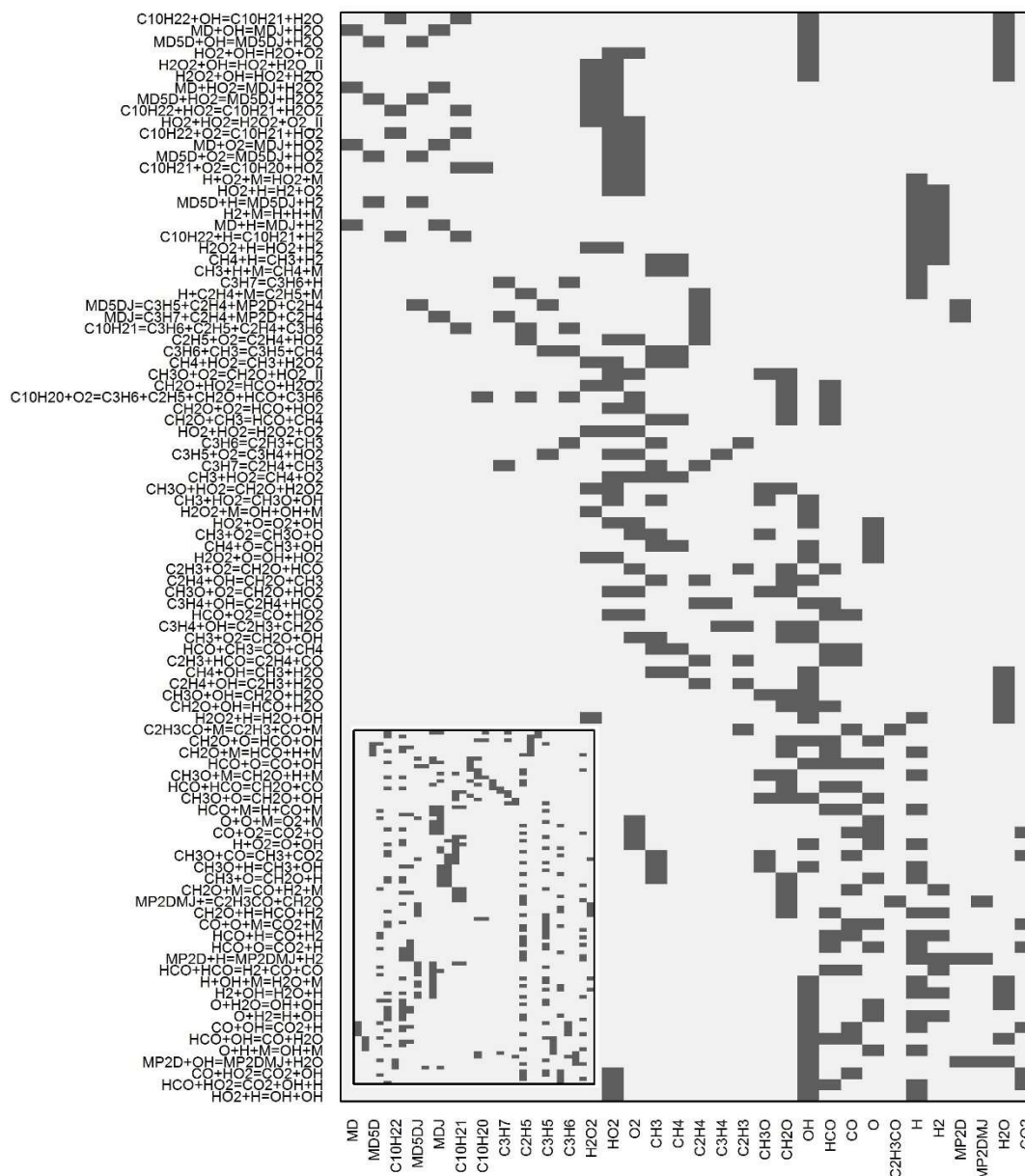


5.2. ábra. a) Random sor-oszlop permutációval az eredeti adatmátrix színkódos értékekkel. b) A folt szeriált adatmátrix. A fekete szintvonalak azokat a pozíciókat mutatják az adatmátrix értékeinek színkódjai felett, ahol a lokális hasonlóság értéke 0,85-nél nagyobb. c) A lokális hasonlóság mátrixa a rendezett esetben d) Az adatgenerálásból ismert, de a szeriálás során nem használt csoportinformáció a szeriált mátrixban. [s63]

Az 5.2.b ábra egy összetett ábrázolás, ahol az eredeti értékek színkódos ábrázolásán felül szerepel egy szintvonalas rész is azokkal az esetekkel, ahol a lokális hasonlóság értéke nagyobb mint 0,85 (értékeit lásd a c ábrán). Ezzel az általam kidolgozott kombinált ábrával könnyen azonosíthatóak az objektumklaszterek és a hozzátartozó változók. Az [s63] cikkben részletesen összehasonlítottuk a módszerünket más szeriálási eljárásokkal. Egyik másik szeriálási módszer se volt képes megtalálni ezt a rejtett csoport információt, még megközelítően se olyan jól, mint a foltseriálás lokális hasonlósági alapon.

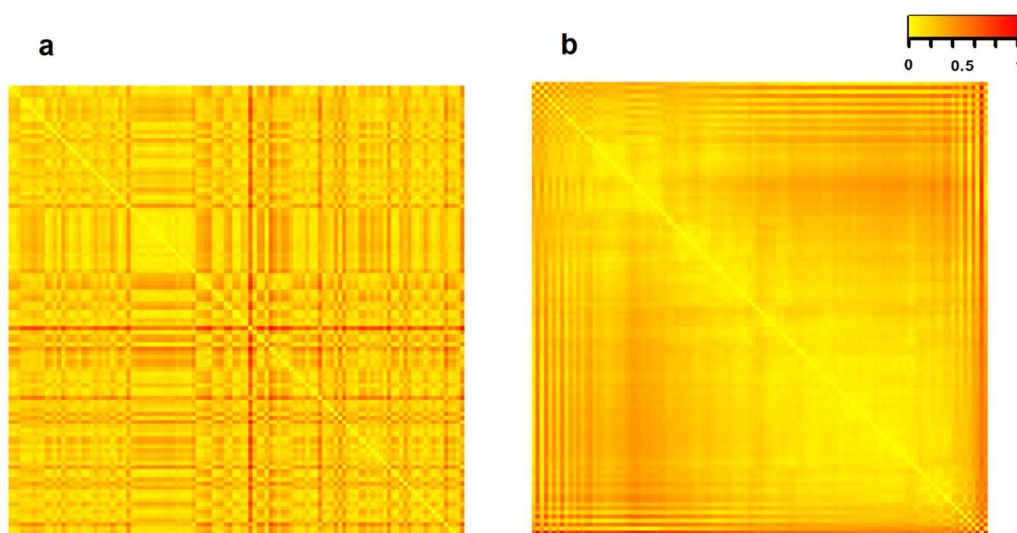
A diagonális szeriálásra egy biodízel reakció egyenletrendszer példája látható a következő ábrán (5.3. ábra) [Juhász 2015]. A komponensek az x, a reakciók az y tengelyen találhatóak. A fekete színezés jelenti, hogy egy adott komponens részt vesz-e egy adott reakcióban. Az ábrán a diagonális mérték minimalizálva volt, illetve kényszerként szerepelt, hogy a három kiindulási reaktáns az első három, a víz és széndioxid termékek az utolsó oszlopokban szerepeljenek. Az adott sorrend nem kapcsolható össze valós reakcióúttal. Az égéskinetikai területen más gráf jellegű megközelítések azóta széleskörűen

elterjedtek (cf. 4.6 fejezet és Kawka et al. 2020). Turányi Tamás csoportja is használja ezeket, akikről a szeriált reakciórendszer származott.



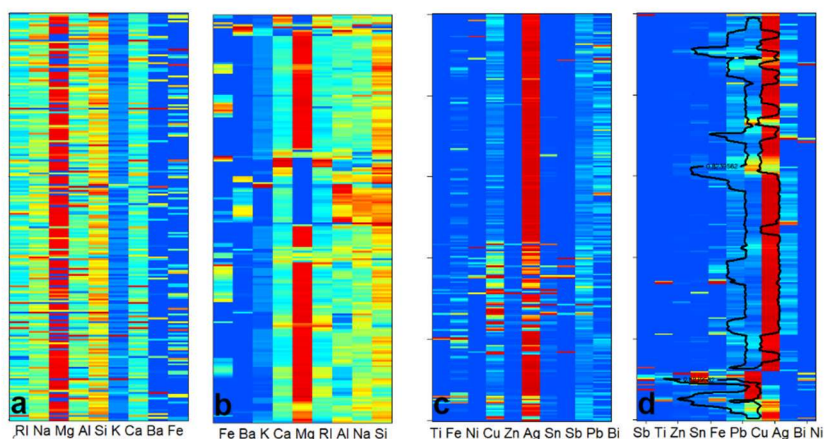
5.3. ábra. Biodízel égési modellezés reakció rendszerének és komponenseinek szeriálása diagonális célfüggvénnyel. A kis ábrán a szeriálatlan sorrend látható. A feketével jelölt mátrixelemek jelölik, ha az adott komponens szerepel az adott reakcióban, akár kiindulási anyagként, akár termékként. [s53]

A diagonális rendezést távolság- és korrelációs mátrixokon is kipróbáltuk. A távolságmátrixos példa a következő ábrán toxikológiai adatok mért és QSAR becsült adatai alapján készült [Arthur et al. 2017]. Jól látható, hogy a szeriált ábra sokkal áttekinthetőbb, a hasonló toxikus értékekkel rendelkező molekulák egymás mellé kerültek (5.4. ábra). A hagyományos csoportosítási eljárásoktól eltérően egy folytonos átmenet figyelhető meg, nem diszkrét a csoportok megjelenése.



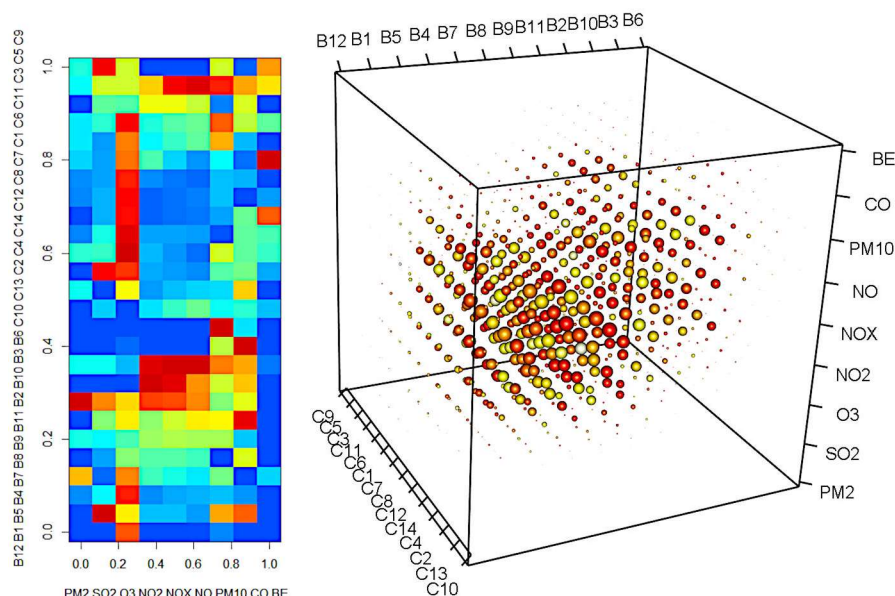
5.4. ábra Molekulák változótérbeli távolságmátrixának szeriálása diagonális mértékkel Artur et al. adatain a) véletlen sorrend b) a távolságmátrix egy módú két utas szeriálás után. Mindkét tengelyen az adatsor molekulái szerepelnek. A színkód a $[0;1]$ intervallumba skálázott távolságokat mutatja. [s53]

A 2 módú - 2 utas rendszerekre vonatkozó foltfüggvényes szeriálást a következő ábra demonstrálja. Mindkét esetben kémiai elemek alkották a változókat. A baloldali két ábra üvegek elemösszetételét [German 1987] mutatja, a jobboldali X-XIII. századi magyar érmék [Rácz et al. 2013, Christie et al. 2014] elemanalízisét (5.5. ábra). A szeriálással tisztább színkódos ábrákat kapunk, sokszor jól értelmezhető hasonlóságot kapva az objektumokra és azok hasonlóságát okozó változókra.



5.5. ábra. Két módú két utas folt szeriálások skálázott adatmátrixai a-b) üvegösszetételek [German 1987] a) véletlen sorrend b) szeriált c-d) érmék elemanalízise [Rácz et al. 2013, Christie et al. 2014] c) véletlen sorrend d) szeriált sorrend. Mindkét adatsornál függőlegesen az objektumok szerepelnek, a-b esetben az üvegminták, c-d esetben az érmék, míg vízszintesen a kémiai analízis elemei. [s63]

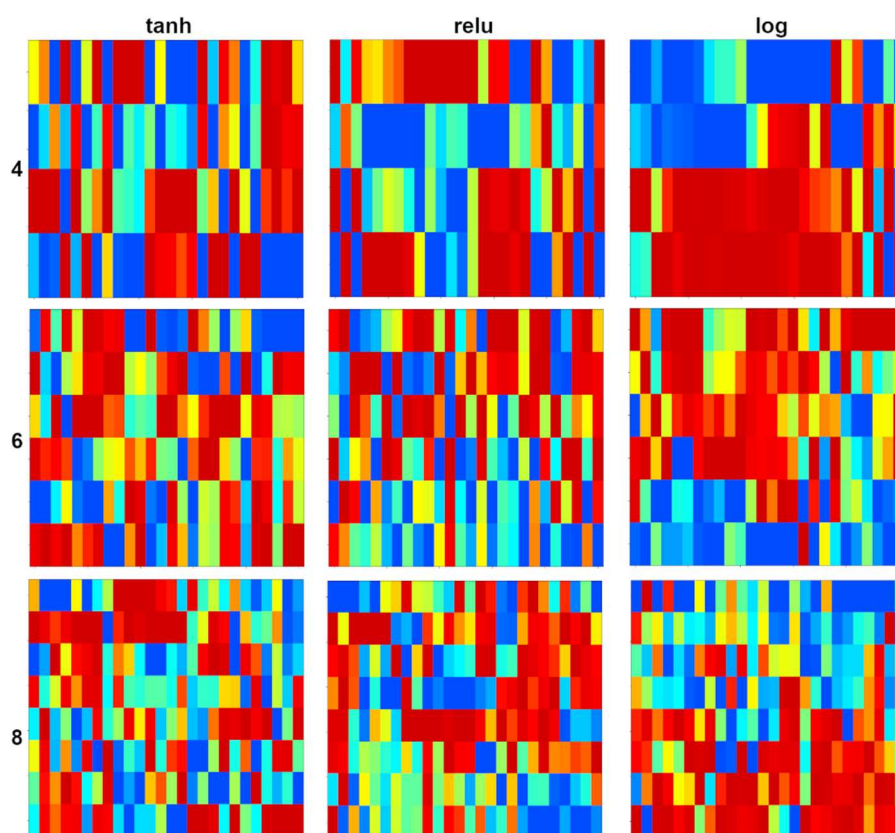
Három módú – három utas szeriálás esetén az eredmények bemutatása nehezebb. A példán egy olyan eset szerepel, amikor légszennyezőket mértek budapesti és vidéki monitoring állomásokon [HAQN 2017]. A szeriálás a háromdimenzióban számolt lokális hasonlóság mátrix alapján történt, külön tengelyként a szennyezőkre, a budapesti és a vidéki állomásokra. Az eredmény azonban visszatranszformálható az eredeti kétdimenziós adatmátrixba (bal oldal), illetve ábrázolható maga a lokális távolságmátrix is (5.6. ábra).



5.6. ábra. Három módú három utas szeriálás légszennyezők éves adatára [HAQN 2017, s63] Bal oldal: szeriált adatmátrix két dimenzióba visszatranszformálva, (budapesti (B1-B12) és vidéki (C1-C14) állomásokon mért értékek (y irány két részre bontva) és x-irányban a légszennyezők (ábrán PM2 = PM2.5 koncentráció, BE = benzol koncentráció) Jobb oldal: A háromdimenziós lokális hasonlósági mátrix a három permutált tengelyre (budapesti és vidéki állomások, légszennyezők) Részletek az [s63] publikációban találhatóak.

Nemcsak mért vagy számolt adatmátrixok vizualizációját segítheti elő a szeriálás, hanem pl. a választást különböző modellek között, illetve azok interpretációját. A következő ábrát lobbanáspontokra vonatkozó QSAR adatsorra készült [Tetteh et al. 1999] egy rejtett réteget tartalmazó mesterséges ideghálók belső szerkezetének a vizualizációjára használtuk. 4, 6 és 8 rejtett rétegbeli neuronnal és háromfajta átviteli függvénnyel készült 9 modellt hasonlíthatunk össze a szeriálás után a következő ábrán, ahol az objektumok a rejtett réteg neuronjai, a változók pedig az adott neuronhoz tartozó súlyok a bemeneti réteg felől (5.7. ábra). Az ábrára kerülő modellek kiválasztásánál csak annyi feltételt szabtuk meg, hogy mind a tanító, mind a teszt halmazra 0,9 feletti R^2_{teszt} értékűeknek kellett lenniük.

Jól látható, hogy a kevés neuront tartalmazó modellek jobban interpretálhatóak, illetve a legkönnyebb ez logisztikus átviteli függvény esetén.



5.7. ábra. Lobbanáspont modellek szeriált tulajdonságai [Tetteh et al. 1999, s63]. Az objektumok a neurális hálók rejtett rétegében levő neuronokat jelentik (4, 6, 8 neuronos modellek), a változók pedig a bemeneti változók súlyainak a hatása a modellekben a neuronokra. Három aktivációs függvényre: tangens hiperbolikus, relu és logisztikus (szigmoid). [s63]

Összegezve, mire is jó a szeriálás? Olyan nem destruktív, vagyis adatvesztéssel nem járó permutációs eljárás, ahol az új formájában jobban áttekinthető adatstruktúrákat kapunk. A szeriálás önmagában ritkán éri el a hagyományos, de külön-külön alkalmazott statisztikai/kemometriai módszerek adatbányászati teljesítményét, de önmagában is jó ötleteket ad az adatokban szereplő mintázatok, trendek, eloszlások felismerésére. Az általunk kidolgozott módszerek speciális szerkezetű mátrixokra készültek, ahol egy-egy csoportba sorolható objektumok a változó tér egy-egy készleténél nagy hasonlósággal rendelkeznek, míg a többi változónál levő értékek nem relevánsak a csoportosítás szempontjából. Módszerünk egyedi megközelítése az is, hogy a több tengely szerinti rendezések csatoltak a lokális távolság vagy hasonlóság mátrix rendezése miatt. Az eredmények két alakzat szerint értelmezhetőek, diagonális, vagy folt jellegű formában. A módszereink által preferált adatstruktúra

esetén más szeriálási módszerek teljesítőképessége kifejezetten gyenge (részleteket lásd [s63] kiegészítő anyag 4. pontjában levő táblázatban).

5.2 In memoriam Szepesváry Pál: trendvizsgálat, normálosztás kérdései

Trendvizsgálat varianciaanalízissel [s42] [s47]

Ebbe és a következő alfejezetbe négy olyan cikk kutatásait vettem be, amelyekben vagy résztvevőként, vagy a téma elindításában Szepesváry Pál is részt vett.

Egy külső oktatásra való felkészülés során Szepesváry Pál a közös dolgozószobánkban heteken át a Neumann-féle trend tesztet [Neumann et al. 1941] tanulmányozta. A trendtesztnek létezik egy másik ismert neve is, az Abbé teszt. A teszt alapjában a $0 \leq d_N \leq 4$ mennyiséget számolja összegezve N darab sorrendiséggel rendelkező x_i változókra:

$$d_N = \frac{\sum_{i=2}^N (x_i - x_{i-1})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad 5.7. \text{ egyenlet}$$

Hipotézisei:

H_0 : nincs fokozatos trend (tendencia) az adatokban

H_1 : fokozatos trend található az adatokban

Valójában normális eloszlású trendmentes adatokra $d_N=2$, ennél kisebb érték esetén pozitív autokorreláció van az egymást követő adatoknál, 2-nél nagyobb esetben negatív. A tesztnek nem triviálisan számolható, a mérettől és szignifikancia szinttől függő kritikus értékek tartoznak, pl. pozitív autokorreláció esetén akkor vetjük el H_0 -t, ha $d_N < d_{Ncrit}$, pl. $N=30$ $\alpha=0,05$ -nél $d_{Ncrit}=1,418$. A teszt ismert, de ritkán használt. A teszt alapfeltételezése, hogy az adatok normál eloszlásúak.

A tesztnek egy változata a Durbin-Watson teszt [Durbin&Watson 1950, 1951], ahol ugyanez r_i illesztés utáni reziduálisokra vonatkozik, alapértelmezésben lineáris regressziót követő maradékokra:

$$d_{DW} = \frac{\sum_{i=2}^N (r_i - r_{i-1})^2}{\sum_{i=1}^N r_i^2} \quad 5.8. \text{ egyenlet}$$

A teszt külön alkalmazható pozitív és negatív autokorreláció értelmezésére, szokásosan három válaszra tagolt skálán, ahol pl. pozitív korreláció esetén $d_{DW} < d_{DW,lowcrit}$ szignifikáns pozitív autokorreláció van, $d_{DW,lowcrit} < d_{DW} < d_{DW,upcrit}$, nincs szignifikáns döntés és $d_{DW,upcrit} < d_{DW}$ esetén nincs pozitív korreláció.

Az általam fejlesztett egytényezős varianciaanalízis alapú (ANOVA) megfelelője a két tesztnek a hagyományos varianciaanalízis formulát használta [s42], vagyis

$$F = \frac{\frac{\sum_j^G \left(\frac{\sum_i^n y_{ij}}{n_j} \right)^2}{N} - \left(\frac{\sum_j^G \sum_i^n y_{ij}}{N} \right)^2}{\frac{\sum_j^G \sum_i^n y_{ij}^2 - \sum_j^G \left(\frac{\sum_i^n y_{ij}}{n_j} \right)^2}{N-G}} \quad 5.9. \text{ egyenlet}$$

, de a csoportokat a Neumann-féle módon egymás utáni párokból választotta. Vagyis Y mátrix oszlopai egy-egy véletlenül választott egymás melletti párt jelentettek, x_i -t és x_{i+1} -et, és összesen n_j ilyen párról volt szó. Mivel az ANOVA-nál függetlenséget kell feltételezni a csoportok között, ezért nem célszerű a létező összes párosból egy-egy csoportot képezni. Ezt elkerülendő véletlen mintavétellel számoltam a csoportokat, $N/4$ darabot minden alkalommal, visszatevés nélkül módon. A végső eredményként több ilyen mintavételezés átlagát vizsgálatam az ANOVA-ban használt F -eloszlással való összevetéssel. Ismert, hogy az ANOVA is normál eloszlást feltételez, de valamivel gyengébb szinten, mint a Neumann vagy a Durbin-Watson teszt. Itt elég a csoportokon belül a normál eloszlás feltételezése, nem a teljes adathalmazra kell feltételezni.

A kidolgozott teszt hipotézisei mind a Neumann, mind a Durbin-Watson sémában felírhatóak, attól függően, hogy az eredeti adatokra vagy illesztés utáni reziduálisokra alkalmazzuk [s42]:

Neumann esetben:

H_0 (Neumann): Nincs graduális trend az adatsorban

H_1 (Neumann): Graduális trend van az adatsorban

H_0 elutasításra kerül, ha $1-p(<F>)$ kisebb, mint az α szignifikancia szint, vagy más néven hibaszint.

Durbin-Watson esetben:

H_0 (DW): Nincs autokorreláció az adatban

$H_{1,low}$ (DW): Pozitív autokorreláció van

$H_{1,up}$ (DW): Negatív autokorreláció van

H_0 (DW) elutasításra kerül $H_{1,low}$ irányában, ha $1-p(<F>)$ kisebb mint α .

H_0 (DW) elutasításra kerül $H_{1,up}$ irányban, ha $1-p(<F>)$ nagyobb mint $(1-\alpha)$.

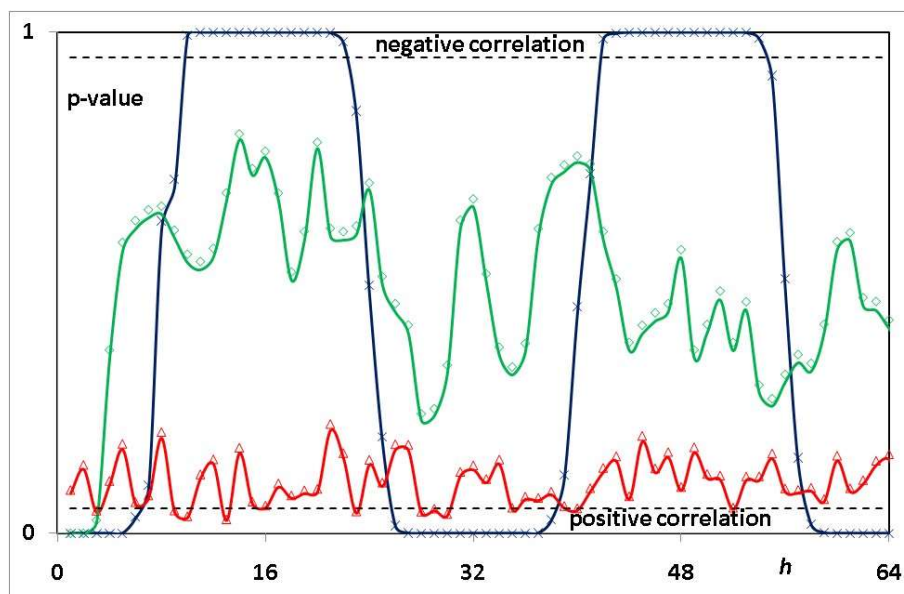
Ugyanebben a cikkben még egy kérdést tárgyaltam, általánosítottam mind a Neumann és Durbin-Watson, mind a saját tesztet arra az esetre, ha a párok számolásánál nem egymás utáni, hanem h -val eltoló értékekkel számoltam. Ennek szerepe az volt, hogy a megszokott időfüggő $C(h)$ autokorrelációkra is számolható legyen a trendvizsgálat:

$$C(h) = \frac{\sum_i (x_i - \bar{x})(x_{i+h} - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad 5.10. \text{ egyenlet}$$

Modellszámolásokkal bemutattam, hogy a régi és az ANOVA teszt közel azonos detektálási készséggel rendelkezik. Ugyanakkor nem szükséges a nehezen számolható Neumann vagy Durbin-Watson kritikus értékeket keresgélni az interneten. Az F -eloszlás értékei szinte minden statisztikai, táblázatkezelő szoftverben és programnyelvben elérhetőek, illetve a beépített ANOVA függvények is alkalmazhatóak, ha Y -t feltöltöttük a megfelelő párokkal.

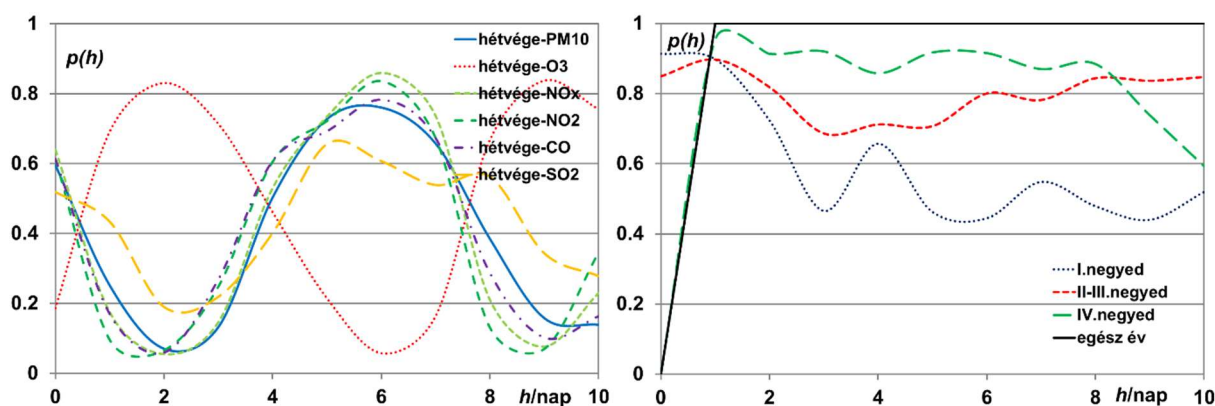
Az időeltolódásos változatot szimulált adatsorokon teszteltem, majd légszennyezési adatok elemzésére használtam. A szimulált adatokra végzett számolásokról látszik az 5.8. ábrán, hogy az időeltolódásos

ANOVA verzió egyenértékű az általunk javasolt Neumann(h) módszerrel és sokkal érzékenyebb, mintha csak az autokorrelációs függvények alapján keresnénk a trendeket.



5.8. ábra. Három szimulált függvény ANOVA(h) (vonalak) és Neumann(h) (jelölők) féle trendtesztrel számolt p -értékei Függvények: Kék keresztek: $x_i=0,4*\sin(i*2*\pi/32)+0,6*Err$ Piros háromszögek: $x_i=0,4*i/1000+0,6*Err$ Zöld négyszögek: $x_i=0,4*(x_{i-1})+0,6*Err$, $i_{\max}=1000$, Err: standard normális eloszlásból vett minta. [s42]

A fenti egyenletek kis módosítással átírhatóak úgy is, hogy nem egy x változó önmagával való eltoltságait vizsgáljuk, hanem az x változó egyik értékét egy másik y -nal jelölt változó későbbi értékéhez hasonlítjuk [s47]. Ebben az esetben a módszer valójában keresztkorreláció vizsgálatát teszi lehetővé, szinte teljesen hasonló elven. A gyakorlatban azonban nem egyedi adott h -értékre vonatkozó tesztként alkalmaztuk a módszert, hanem a h függő p -értéket ábrázoltuk és abból vontunk le következtetéseket. A módszert légszennyezési és kapcsolódó meteorológiai adatokon alkalmaztuk [HAQN 2009-2011, Salma 2023]. A módszerrel vizuálisan könnyen értelmezhető adatokat kaptunk, pl. a légszennyezés és a munkanap/hétvége kapcsolatára, a szezonális hatására a napsütésnek és a nedvességnek (5.9. ábra). A klasszikusan tanított és ismert légszennyezőkkel kapcsolatos trendek, azok egymással és az időjárási adatokkal való korrelációi részletesen szerepelnek a cikkünkben [s47].



5.9. ábra. Bal oldal) Légszennyezők keresztkorrelációja a hétvége/hétköznap adattal az év első negyedében. A hétvége 1-gyel, a hétköznap 0-val volt jelölve. Jobb oldal) A levegő nedvességének korrelációja a napsütés intenzitásával az év különböző szakaszaiban. 2007-es adatok alapján [s47].

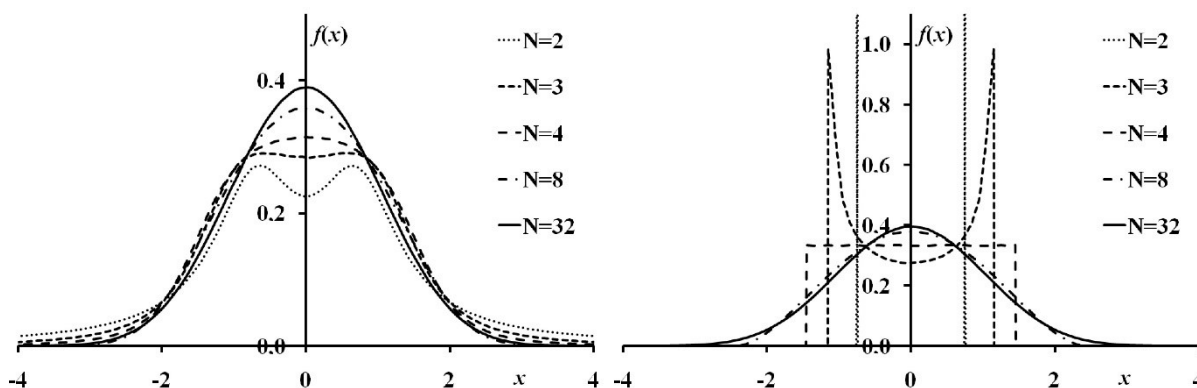
Kérdések a normál eloszlással kapcsolatban [s44] [s51]

Egyik oktatási feladatomhoz normál eloszlású véletlenszámokat generáltam és azok négyes csoportjain próbáltam bemutatni a standardizálás módját és az eredő standardizált változók eloszlását. Legnagyobb csodálkozásomra a kapott változók eloszlása egyenletes eloszlású volt! A téma alapos körbejárása után hat lehetséges kombinációját találtam a skálázásnak és a centrálásnak, attól függően, hogy azt elvi vagy mintából becsült várható értékkel és szórással tesszük, illetve, hogy az egyedi vagy a minta átlag eloszlását vizsgáljuk. A 6 lehetőséget a következő táblázatban foglaltam össze (5.1. táblázat).

A táblázat eredményei többnyire ismertek voltak a szakirodalomban, de érdekes módon a szisztematikus összefoglalásuk és értelmezésük sehol se volt elérhető. Valamiért a legtöbb helyen megelégedtek a t -eloszlás értelmezésénél. Ugyanakkor van olyan eset (3-as), ahol se mi, se máig senki se határozta meg az elméleti eloszlást, illetve ahol mi adtuk meg először a zárt formát (4. eset). A 3. és 4. eseteket néhány mintaelemszámra itt is bemutatom, a sűrűségfüggvények alakja igazi tudományos érdekesség.

5.1. táblázat. Különböző módon centrált és skálázott adatok eloszlásai. Az eredő X változó Y változókból lett számolva, ahol Y N elemű minta $\mathbf{N}(\mu; \sigma^2)$ normál eloszlásokból. $\bar{Y}$: a minta átlagát, s : a szórását jelöli. μ és σ a sokaság várható értékét és szórását jelöli, a t és β indexek a student eloszlásra és a β -eloszlásra utalnak. [s44]

Típus	$X_i =$	$f(X) =$
1	$\frac{Y_i - \mu}{\sigma}$	$\mathbf{N}(0;1)$
2	$\frac{Y_i - \bar{Y}}{\sigma}$	$\mathbf{N}\left(0; 1 - \frac{1}{N}\right)$
3	$\frac{Y_i - \mu}{s}$	$f(X) = \frac{1}{\pi\sqrt{2}} \frac{(X^2 + 1)}{(X^4 + 1)}, \text{ ha } N=2; \quad ? \text{ ha } N>2$
4	$\frac{Y_i - \bar{Y}}{s}$	$f_{\beta}\left(X^2; \frac{N}{(N-1)^2}; \frac{1}{2}; \frac{N-2}{2}\right) * X ^{\frac{N}{(N-1)^2}}, \text{ ha } N>2$
5	$\frac{\bar{Y} - \mu}{\sigma / \sqrt{N}}$	$\mathbf{N}(0;1)$
6	$\frac{\bar{Y} - \mu}{s / \sqrt{N}}$	$f_t(X; N-1)$



5.10. ábra. Az 5.1. táblázat 3. (bal) és 4. (jobb) sorának megfelelő X változók sűrűségfüggvénye néhány kiválasztott mintaméret esetén [s44].

Ennek a témának a kidolgozásakor futottam bele, hogy bár a t -eloszlás levezetésekor normál eloszlású mintán való számolásáról van szó, az ott szereplő s egy olyan módon torzított becslése a szórásnak, aminek az irodalomban létezik a torzításmentes változata. Az ún. $c_4(N)$ korrekcióval tehető rendbe, ahol $c_4(N)$ gamma függvényekből számolható és a torzításmentes becslt szórás a következőképpen adható meg:

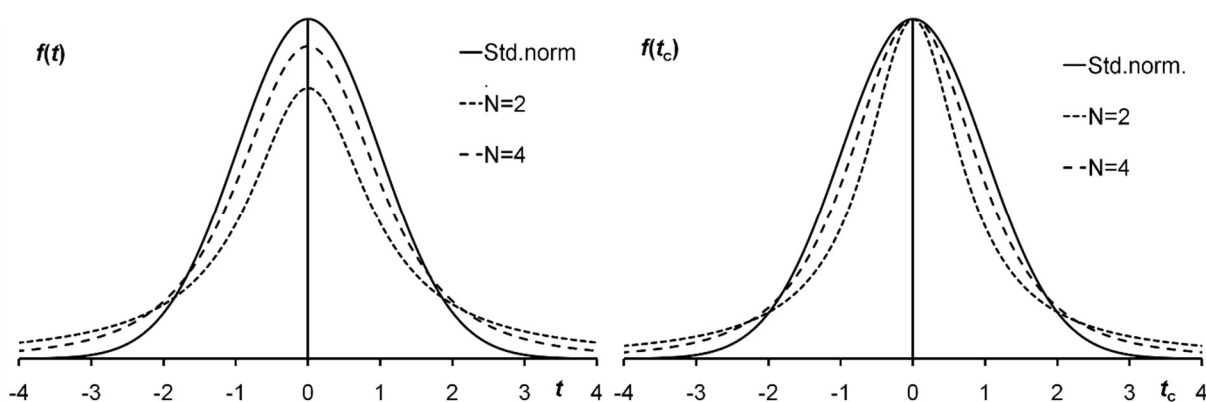
$$c_4(N) = \sqrt{\frac{2}{N-1}} \frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})} \quad 5.11. \text{ egyenlet}$$

$$s_c = \frac{s}{c_4(N)} \quad 5.12. \text{ egyenlet}$$

Számomra máig se világos, hogy mióta létezik ez a korrekció, nem sikerült visszafejtenem a szakirodalmat, de [s51]-ben több mint százéves, illetve közel százéves cikkekig jutottam. Amennyiben ezt a korrekciót használjuk a t változó kiszámolásához,

$$t_c = \frac{\bar{y} - \mu}{s_c / \sqrt{N}} \quad 5.13. \text{ egyenlet}$$

, és ennek az eloszlását ábrázoljuk, akkor a korrigált t -eloszlás sűrűségfüggvénye nem mutatja a jól ismert középen lenyomott formáját a normál eloszlásnak.



5.11. ábra. Hagyományos módon (bal oldal) és torzítatlan(jobb oldal) s -sel számolt t -eloszlások $N=2$ és $N=4$ méretű minták esetén [s51]

Mivel ezt a sűrűségfüggvényt sehol se találtuk meg az irodalomban, ezért megpróbáltuk leközelíteni. Természetes jeleztük, hogy ez csak egy elvi kérdés, mivel ha s -t vagy az s_c -t használjuk a képletekben, a végső változókat pl. konfidencia intervallum meghatározására használva, a konfidencia intervallumok határai az eredeti x -re vonatkozóan ugyanazok lesznek, illetve statisztikai próbákban is ugyanazt kapjuk. A cikket több újság is visszautasította, hogy vagy ismert dolognak tekintették, vagy teljesen érdektelennek. Ugyanakkor továbbra se tudtuk meg, hol közölték ezt, ha ismert. A közlési próbálkozásokat végül feladtuk, de rá fél évvel az egyik, általunk több körben zaklatott matematikai újság egyik szerkesztője megjelentetett egy cikket, amelynek a fele erről szólt...

2017-ben egy veszprémi szerkesztésű újságba Liszi János emlékszámot terveztek. Szepesváry Pali bácsi jól ismerte Liszi Jánost, mert ő is közel 20 éven át Veszprémben lakott és dolgozott. Úgy gondoltam, hogy a témát felmelegítem és a Liszi emlékszámba írok egy Szepesváry posztumusz cikket. Az eredeti témát kiegészítettem egy újabb példával, ahol bemutattam, hogy a $c_4(N)$ korrekcióval rendbe tehető a Geary-féle kurtózis (csúcsosság) számolás méretfüggése [Geary 1936]. Az eredeti számolásban levő

szórás helyett a $c_4(N)$ -nel korrigált szórást kell alkalmazni. A következő táblázat mutatja az eredeti Geary-féle kurtózis N -függését (5.2. táblázat).

5.2. táblázat. A hagyományos Geary-féle kurtózis méretfüggése. A $c_4(N)$ -nel korrigált Geary-féle kurtózis minden mintaméretnél a végtelen nagy mintára vonatkozó határértéket adja, vagyis az utolsó sor $\sqrt{2/\pi}$ értékét. Az értékeket 10^5 db normáeloszlásból véletlen mintavételezés átlagaként kaptam. [s51]

N	Geary féle kurtózis
3	0.9004
4	0.8659
5	0.8489
6	0.8385
7	0.8319
8	0.8267
9	0.8233
10	0.8200
20	0.8085
50	0.8021
100	0.7999
∞	0.7979

Megvizsgáltuk számos normalitás eldöntésére vonatkozó statisztikai teszt elsőfajú hibáját is. Ezen tesztek egy része kifejezetten torzít kis mintaméreteknél, használatuk itt nem is javasolt (pl. Anscombe-Glynn, Bonett-Seier, Jarque-Bera tesztek [Anscombe&Glynn 1983, Bonett&Seier 2002, Jarque&Bera 1980]) [s51]. Amennyiben a tesztben használt kurtózisokat (Geary és Pearson féle) a $c_4(N)$ korrigált szórásokkal segítségével számoljuk, a tesztek alkalmazhatósága javul a kis mintaméreteknél.

5.3 Az allokáció és a modellvalidálás kérdései

A modellalkotás kérdéseinek két központi témája az allokáció és a validálás. Allokáció alatt a modellezéskor azt értjük, hogy milyen változó tartományokból, vagy akár pontosan milyen értéknél és milyen stratégia alapján jelöljük ki a különböző halmazok, pl. a tanító és a teszt készletek elemeit. Validálás alatt pedig egy adott modell minősítését értjük annak érdekében, hogy a modell teljesítőképességét pontosan határozzuk meg.

Periodikus adatok allokációja kemometriai számolásokban [52]

A témában az értekezésben elsőként bemutatásra kerülő munka valamennyire az allokáció témakörébe tartozik, de egy speciális jellegű adatstruktúrát érint, ahol a változók vagy a változók egy része periodikus. Klasszikus molekuláris dinamikai és Monte Carlo szimulációs háttérrel számomra ismerős volt a kérdés, hogy periodikus adatok esetében a leíró statisztikai értékek és a mintázatkeresés speciális módon történik. Az elvi megoldás kémia adatok esetében is a végtelen periodikus térbe vagy tóruszra helyezés [pl. Jáklí et al. 2012] lenne, de meglepő módon a standard csoportosító, osztályozó és modellező szoftverekben ezek a lehetőségek nincsenek beépítve, még olyan szinten se, hogy pl. egy távolságmátrix számolása a periodicitást figyelembe véve történjen. A másik furcsaság számomra az volt, hogy rengetegszer közöltek ilyen periodikus adatokra is átlagot, szórást vagy akár kovarianciát is az irodalomban, nem törődve azzal, hogy ezek eltolva a periodikus térben változnak és akár értelmüket is veszítik. Tipikus ilyen változók a teljes 360 fokot felölelő szögváltozók, a periodikus időbeli változók, pl. óra, nap, hónap, évszakok. Az adatelemzési módszerek azonban gyakran használják az átlagot, varianciát és kovarianciát, pl. főkomponens elemzéskor vagy Mahalanobis távolság [Mahalanobis 1936] számolásakor.

Mivel a standard szoftverek nem tudják kezelni a periodikus változókat, ezért felmerült, hogy létezik-e olyan általános elvi alapon kijelölhető periódusa a változótérnek, amely nem valamilyen konvenció, hanem az adatszerkezet alapján optimális.

Ha elképzelünk egy periodikusan ismétlődő n pontból álló mintázatot egy számegegyenesen, akkor a variancia számolásakor n különböző eredményt kapunk, ha pontosan egy periódust veszünk figyelembe. Ezt úgy kell elképzelni, hogy az egy periódusnyi adatazok eltolásakor egy ideig ugyanazt a varianciát kapjuk, egészen addig, amíg egy az egyik oldalon kikerülő adat a túloldalon be nem kerül. Vagyis az eltolás n szakaszra osztható, ahol egy-egy szakaszon belül állandó a variancia. Az ablakok közül mi annak a választását javasoltuk, ahol az adott ablak likelihood függvénye maximális.

n darab y_i pontnak $N(\mu, \sigma^2)$ normál eloszlást feltételezve a likelihood függvénye:

$$L = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right). \quad 5.14. \text{ egyenlet}$$

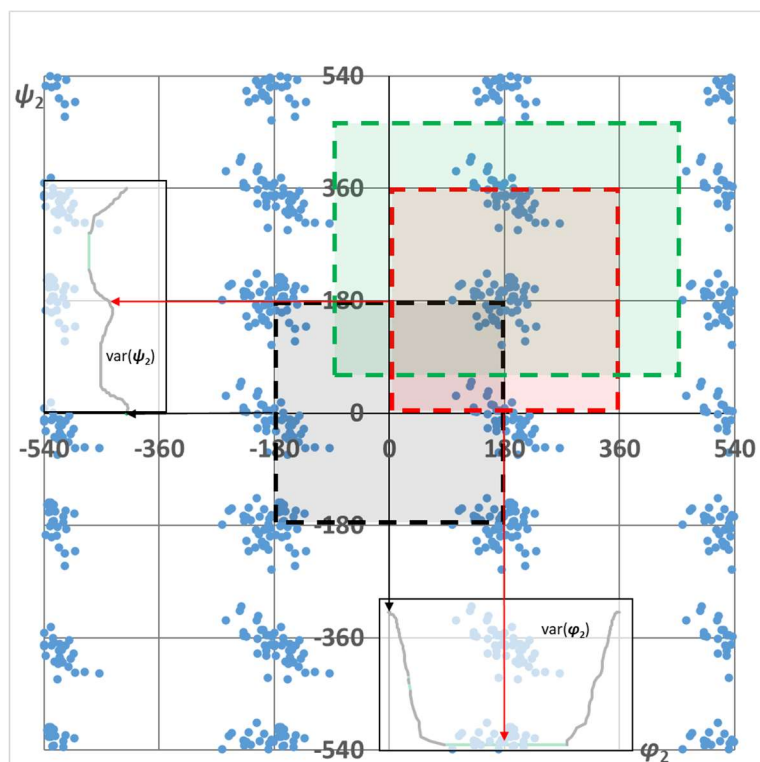
Ha μ -t and σ^2 -t

$$\bar{y} = \sum_{i=1}^n y_i / n \text{ és } s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \quad 5.15. \text{ egyenlet}$$

az átlaggal és a becsült varianciával helyettesítjük, akkor a log-likelihood függvény csak s^2 -től és n -től függ:

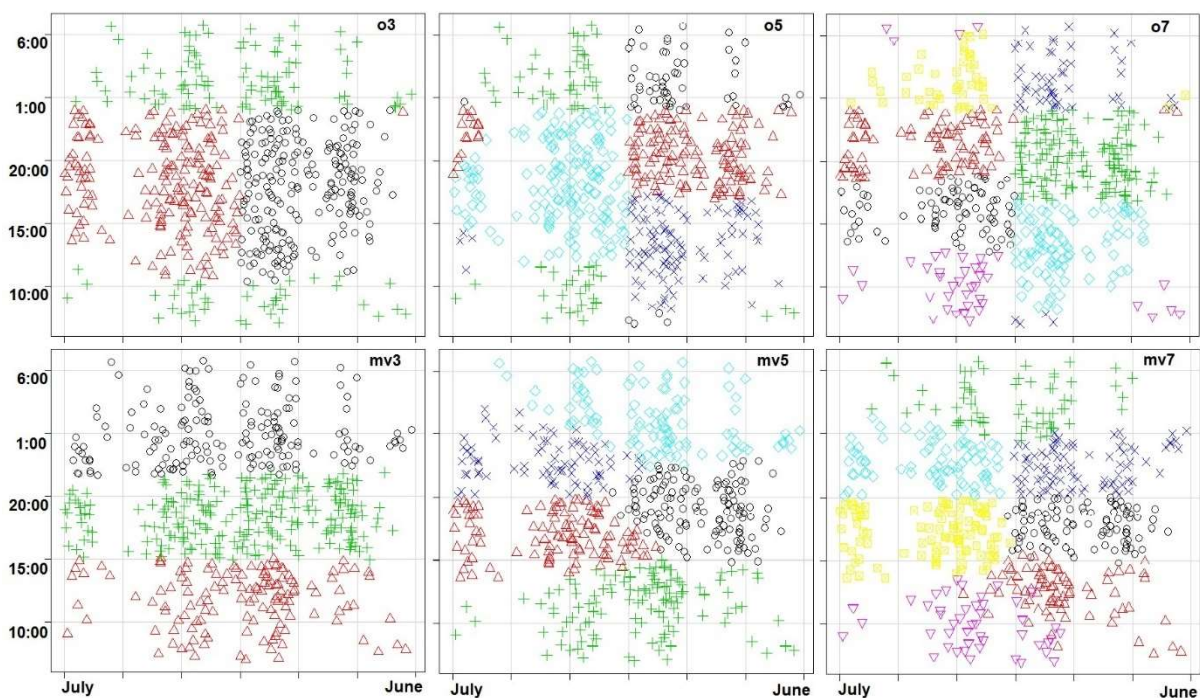
$$\ln L = (-n/2) \ln(2\pi s^2) - \frac{1}{2s^2} \sum_{i=1}^n (y_i - \bar{y})^2 = (-n/2) \ln(2\pi s^2) - \frac{1}{2} (n-1) \quad 5.16. \text{ egyenlet}$$

Az 5.16. egyenletből látszik, hogy a legnagyobb likelihood érték a legkisebb varianciához fog tartozni. A javaslatunk működését két példával illusztráltuk. A következő ábrán triglicin konformereinek Ramachandran szögeértékei láthatóak a C-terminális torziós értékekre (5.12. ábra). Jól látszik, hogy a IUPAC szögtartományban [IUPAC 1970] 6 csoport alakul ki, a $[0-2\pi]$ tartományban három, míg a minimális varianciájú periodikus tartományt választva csak a két valójában eltérő csoportot kapunk. Az ábrán a zöld tartomány nagyobb, mivel mind a két szög szerint az invariáns szakaszok egészére ábrázoltuk, ebből lehet tetszőleges 2π periódusnyit választani.



5.12. ábra. Periodikusan kiterjesztett Ramachandran diagramja a triglicin C terminális torziós szögeinek. Adatablakok: IUPAC – fekete négyszög, $[0, 2\pi] \times [0, 2\pi]$ piros négyszög, minimális varianciájú ablak – zöld négyszög. A kis ábrák a két szög (φ , ψ) függvényében a variancia változását mutatják. [s52]

Hasonló tisztulás figyelhető meg a PM10 szennyezőre [HAQN 2017] a hónap/óra projekción (2012-2016 közötti adatok a Budapest, Széna téri állomáson). A projekciónál az adatvezérelt ablakválasztással elérhető, hogy egyik csoport se legyen szétvága k -közép klaszterezés során, mindegyik csoport egy-egy tartományban ábrázolható (5.13. ábra).



5.13. ábra. k -közép csoportosítási eredmények nagyértékű PM10 légszennyező adatokra a hó-óra projekcióban. A k -közép klaszterszámait 3-ra, 5-re és 7-re állítottam be. „o” utal az eredeti adatsorban levő adatablakra, „mv” a minimális varianciájúra. Az eredeti adatablak adatai is a minimális varianciájú intervallummal szerepelnek a felső sorban, hogy az összehasonlítás lehetséges legyen, de a kilógó színű klaszterek mutatják, hogy ezek csoportosítása az eredeti periódus szerint történt. [s52]

Az allokáció hatása validációs és kísérlettervezési paraméterekre [s58]

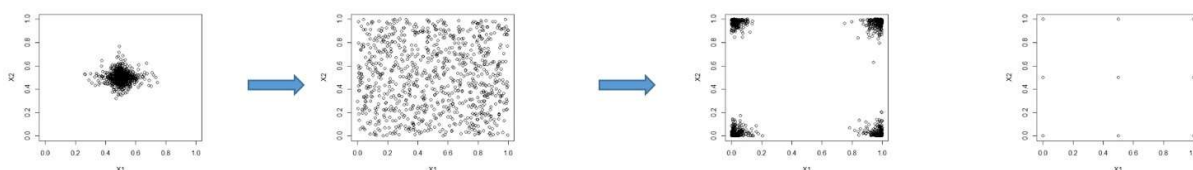
2015-ben Papp Gábor szakdolgozóval elkezdtünk egy kutatást, ahol a numerikus integrálás fixpontos kvadratura módszereinek (pl. Gauss kvadratura) analógiáján kiindulva azt vizsgáltuk, hogy létezik-e olyan allokációja a független változóknak, amelyik kiemelkedően jó többváltozós lineáris regressziót ad. Pozitívnak tűnő eredményeinket egy munkabizottsági ülésen be is mutattam, de pár hétre rá kiderült, hogy az általunk írt R kódban [R 2013] hiba volt és valójában tévesen találtunk valamiért kiemelhetően jó megoldásokat. A témával utána 4 évig nem foglalkoztam, de a validációs kutatásaim elindulásakor újra elővettem. Új célt definiáltam: a modellezés első fázisában szereplő kísérlettervezéssel végzett allokáció és a modellezés utolsó fázisában levő validálás kapcsolatának vizsgálatát. Pontosabban, a kísérlettervezés különböző sémái, optimálisai és a validáció különböző paraméterei közötti kapcsolatot vizsgáltuk.

Az allokáció a modellalkotás egyik első lépése, tervezett formában két módon lehet végezni. Az egyik esetben az ún. kísérlettervezés (design of experiment, DOE [pl. Eriksson et al. 2008]) során valamilyen matematikai függvények (egy betűs „optimális”) vagy változótérbeli sémák segítségével választjuk ki a független változók azon értékeit, ahol a méréseket vagy a számítógépes/kísérleti adatgyűjtést

elvégezzük. A kísérlettervezést itt nem részletezem, legtöbbször a modell paramétereinek minél kevésbé torzított és kisebb szórású (hatékony, efficiens) meghatározása a cél. A másik esetben nagyszámú már meglévő kísérleti adatból szeretnénk kiválasztani egy olyan halmazt, amelyik megfelelően reprezentálja a modellezendő jelenséget (representative subset selection). Itt a fő cél, hogy a modell illesztése során átláthatóbb módon és ésszerű számítógépes kapacitással tudjuk a modell paramétereit kiszámolni, pl. elkerüljük az akár kombinatorikus robbanáshoz vezető nagy mintaszámokat. A kísérlettervezés optimálisai is alkalmazhatóak a reprezentatív halmaz kiválogatására, de ezen a területen néhány másik módszer is ismert [pl. Chaudhuri et al. 1994a, 1994b, Daszykowski et al. 2002, Rodionova&Pomerantsev 2008].

Tehát a DOE és a reprezentatív halmaz kiválogatása a modellezés első lépései közé tartoznak, de a hatásukat igazából csak a modellparaméterek meghatározása után tudjuk számszerűsíteni, amikor a modellt különböző validációs paraméterek segítségével minősítjük. A validáció részleteit majd később más cikkeimmel kapcsolatban tárgyalom, de alapvető, hogy a belső és külső validáció is szükséges, ahol az első a tanító halmazzal kapcsolatosan az illeszkedés jóságát és a modell robusztusságát vizsgálja, míg a modell predikciós (egyes kollégák szóhasználatával jósló, becslési, előre becslő, előre jelző) teljesítménye legjobban a modellalkotástól független külső teszt készleten végezhető el. A validációs paraméterek és a következő fejezetekben levő változók definíciója a függelékben található, az értekezés szövegében azok nagy száma miatt nem definiálom azokat.

A kutatás során szimulált és valós adatsorokon vizsgáltuk, hogy a különböző mutatók milyen módon kapcsolódnak egymáshoz. Eredményeinket legegyszerűbben a kétváltozós lineáris regresszióhoz szimulált adatsorainkon lehet bemutatni. A független változók síkjában szisztematikusan generáltunk tanító halmazt az 5.14. ábrának megfelelően a terület közepére zsúfolt adatoktól a sarokba zsúfolt esetig. Ezen kívül egy fix kompozíciós terv szerint elrendezést is alkalmaztunk [Box&Behnken 1960].

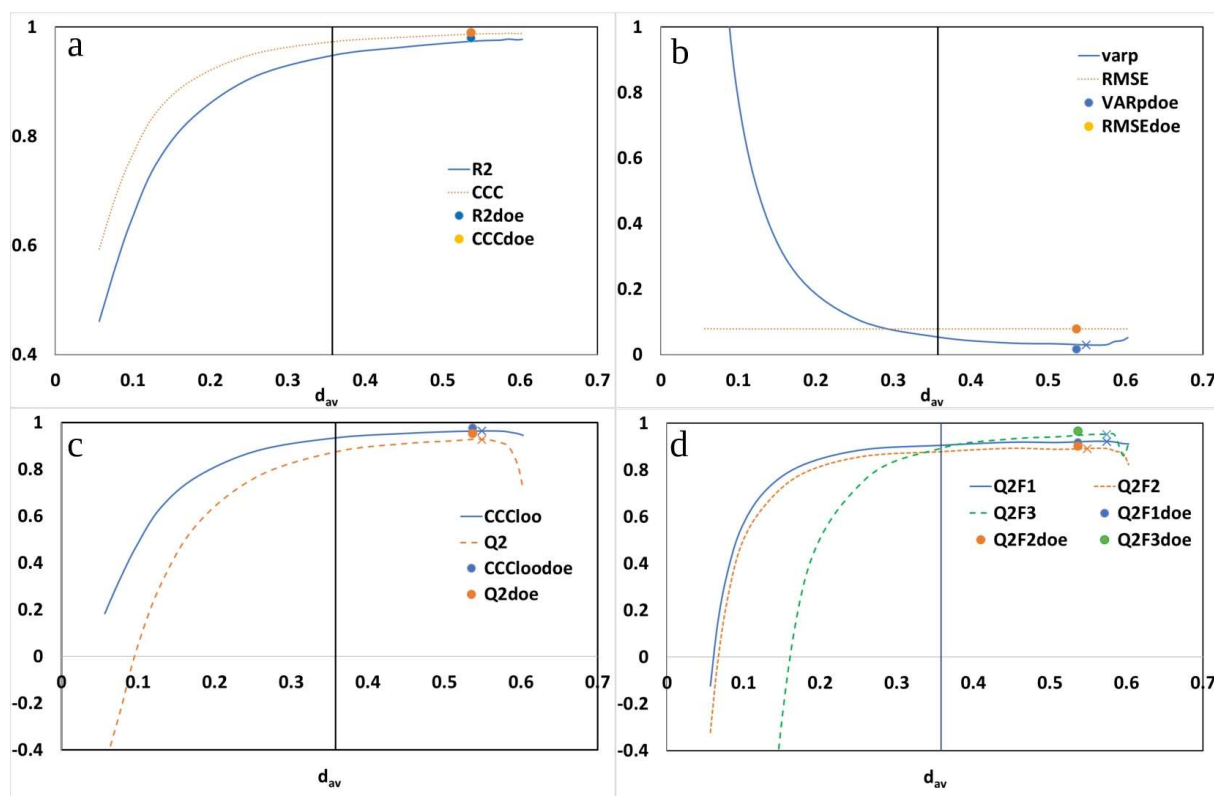


5.14. ábra. Példa a kiválasztott allokációs lehetőségekre két független változó esetén. Balról jobbra: a) allokáció a független változók terének közepére b) egyenletes eloszlású allokáció c) sarok közeli allokáció d) fix kompozíciós kísérleti terv 9 ponttal [s58]

Ezek után minden sorozathoz y értékeket generáltunk a következő képlet szerint:

$$y = x_1 + x_2 + \varepsilon \quad \text{5.17. egyenlet}$$

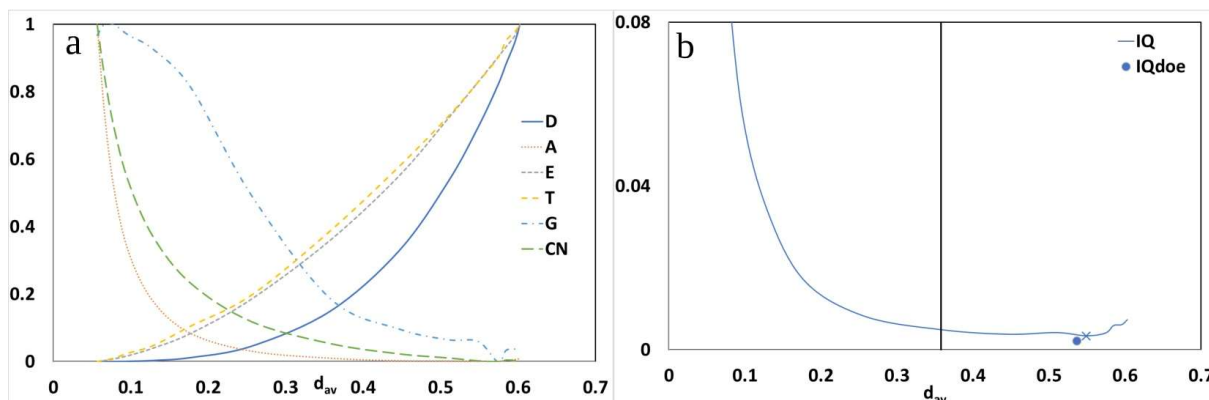
, ahol az utolsó tag $N(0, \sigma)$ normál eloszlású homoszkedasztikus zajt jelentett pl. $\sigma=0,1$ mellett, majd síkot illesztettünk y becslésére. A bemenő készletek mindegyikére kiszámoltunk számos optimális tervezési értékét, illetve általános változóként a független változópárok átlagos távolságát a változó tér közepétől (d_{av}). Az illesztett modellekre ezenkívül validációs paramétereket is számoltunk. Teszt halmazként egy egyenletes eloszlású pontokból álló készletet generáltunk. Az 5.15. ábrán a különböző validációs paraméterek függése látszik a d_{av} átlagos távolság függvényében:



5.15. ábra. Különböző validációs paraméterek átlagos értékei az adatsor allokációját mutató d_{av} függvényében, ami az adatpontok átlagos távolsága a minta középpontjától. A függőleges vonal az egyenletes mintavételezést mutatja, a jelölő a fix kísérleti tervet (*doe* végződés a jelmagyarázatban). A modell részletei: $\underline{x} \in \mathbf{R}^2$, $n=9$, $\sigma=0,01$. a) R^2 és CCC. b) RMSE és az illesztett paraméterek varianciáinak összege (var_p). c) Q^2 és CCC_{loo} . d) külső validációs metrikák: Q^2_{F1} , Q^2_{F2} and Q^2_{F3} . A validációs paraméterek definíciója a függelékben található. [s58]

Látható, hogy az R^2 család metrikái növekszenek a pontoknak a középről a sarkok fele való mozgásával, illetve kifejezetten jó a fix kísérleti terv. Az eltérésnégyzet-összeg jellegűek pedig csökkennek, beleértve a var_p -vel jelölt számot, amivel a paraméterek varianciáinak az összegét jelöltük.

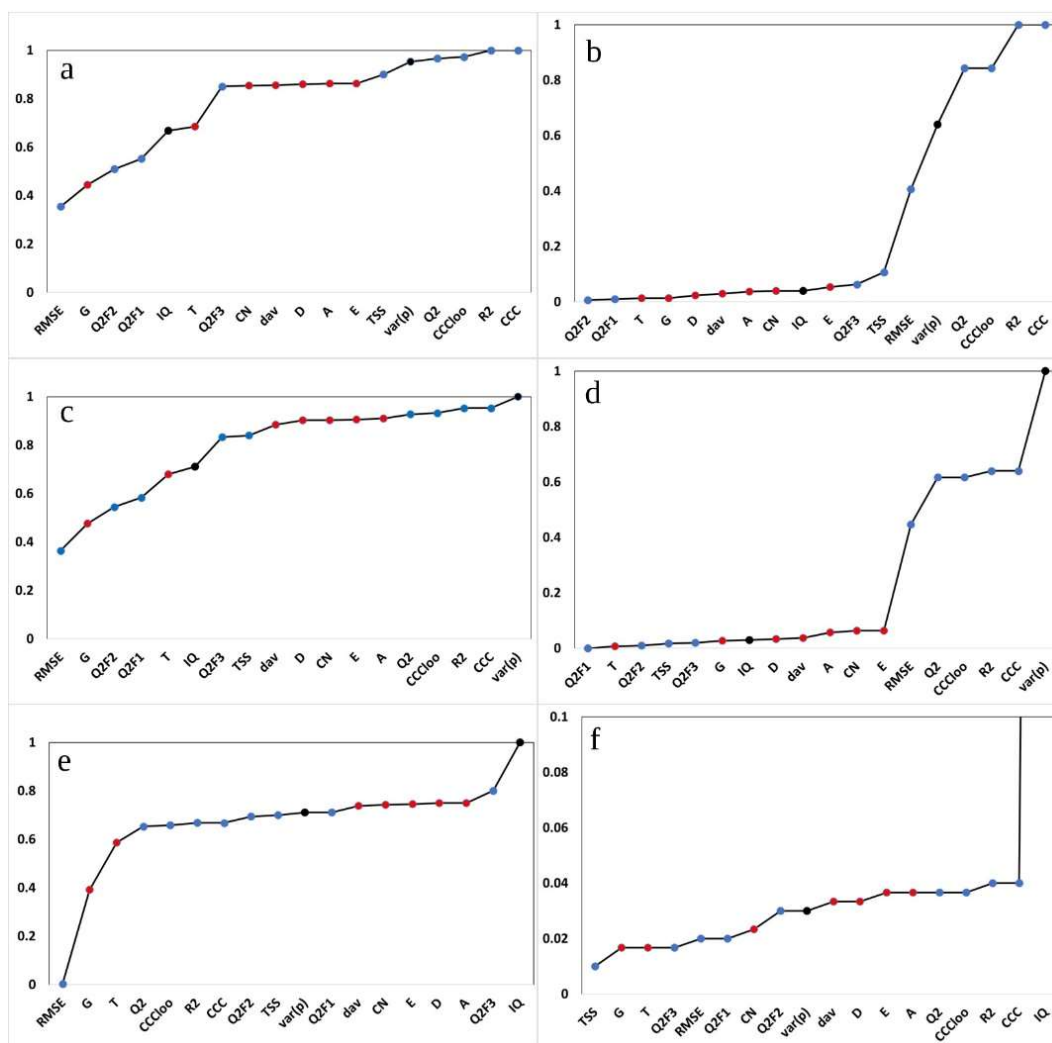
A kísérlettervezéshez kapcsolódó értékek lefutása a 5.16. ábrán bal oldalon szerepel, itt is monoton a függés, emelkedő ha az optimálisnak maximumát és csökkenő ha az optimálisnak minimumát kell célként kitűzni.



5.16. ábra. a) Egy betűs kóddal jelölt kísérlettervezési optimálisok függése az allokációtól [0,1] intervallumba skálázott értékekkel. A definíciók [s58]-ban és a függelékben találhatóak. b) Az integrált predikciós hiba, IQ , függése az allokációtól. A legjobb allokációt csillag, a fix kísérleti tervet pont jelöli. A függőleges vonal az egyenletes eloszlású tanító halmazt mutatja. További adatok: $\underline{x} \in \mathbb{R}^2$, $n=9$, $\sigma=0,01$. [s58]

Mivel szimulált adatsorokról volt szó, ezért levezettem, hogyan lehet geometriai módon kiszámolni az egzakt predikciós hibát [s58]. Ennek a hibának az allokációval való kapcsolata látszik a bal ábrán. A sarkok közelébe helyezett tanító pontokkal ígéretesen csökken a predikciós hiba, beleértve a fix allokációt is.

A két egyirányú változás alapján nagyon optimisták voltunk a további eredményekkel kapcsolatban. Az R^2 , a var_p és az IQ értékekkel rangkorreláltatva a kísérlettervezési optimálisokat és a validációs paramétereket egész jó korrelációkat értünk el egyes párokra (5.17. ábra bal oldal).



5.17. ábra Bal oldal) Két paraméterre vonatkozó rangkorrelációk abszolút értéke. Piros jelöli a modellezés előtt és kék a modellezés után számolható paramétereket. Jobb oldal) Két paraméterre vonatkozóan a legjobb 1% átfedési aránya. Felső sor) R^2 . Középső sor) var_p . Alsó sor) Integrált predikációs hiba (IQ). [s58]

Ha itt befejeztük volna a kutatást, akkor csak arról lett volna szó, hogy numerikusan vizsgáltuk a különböző allokációs lehetőségek és a validációs paraméterek kapcsolatát: minden jó és szép, a kísérlettervezés ellátja a feladatát. Ehelyett folytattuk a vizsgálatainkat, hogy mi van, ha megnézzük, pl. egy adott optimális által az adott helyen legjobb 1%-ként minősített adatkészlet vajon beletartozik-e a legjobb teljesítményű 1%-ba a validáció során (5.17. ábra jobb oldal). Vagy ugyanez vizsgálható volt pl. 5-5 %-nál is. Döbbenetes módon, a két készlet átfedése nagyon kicsi volt egy-egy kiválasztott százaléknál, az ábrán pirossal jelölt pontok átfedési értékei pár század alatt maradtak a legjobb 1%-ra vonatkozóan. Valamivel jobb volt a helyzet a validációs paraméterek legjobbjainak az átfedésével, de itt is főleg a belső validáció szerepelt jobb értékekkel.

A számolásokat megismételtük más szimulált adatsorokkal, QSAR adatokkal és egy hőerőművi adatsorral. Ugyanezeket a trendeket kaptuk. Sajnos, a valós adatoknál nem áll rendelkezésünkre az elvi predikciós hiba (IQ) számolásának a lehetősége. A kísérlettervezéssel kapcsolatos nem túl kedvező eredményeinket csak nagyon árnyaltan mertük a cikkben leírni, hiszen egy egész terület szakértőinek megcsontosodott véleményével kerültünk szembe. Szabadon fordítva a cikkből: „Az elméleti predikciós hiba erősen függött az adott tanító halmazban található egyedi hibáktól, így minimális keresztoptimalási lehetőséget találtunk a kísérleti tervek és az elvi predikciós hiba között. Arra a következtetésre jutottunk, hogy az optimális paraméterekkel való kísérlettervezés nem feltétlenül szolgáltatja azokat az allokációkat, amelyek a legjobbak a validációs paraméterek szerint.” [s58]

Validációs stratégiák az OECD 4. elv és a mintaméretfüggés figyelembevételével [s59] [s61] [s65]

A nemzetközi szakmai fórumokon több egymásnak ellentmondó előadást hallottam a belső és a külső validálás előnyeiről és hátrányairól. Számomra furcsa volt ennek a témának a megosztó jellege [pl. Esbensen&Geladi 2010, Hawkins et al. 2003, Gütlein et al. 2013, Rácz et al. 2015 és 2021, Gramatica 2007 és 2014, Gramatica et al. 2012], amikor közben az újonnan kialakuló „data science” és bioinformatika területeken inkább egyértelműen a külső validálás lett az elvárás [pl. Molnar 2022, Thampi 2022]. Az első alapötletem az volt, hogy vajon milyen hatása van a mintaméretnek a belső és a külső validációra, elképzelhető-e egy olyan mintaméret, ahol az addig minden információt megadó egyik típusú validációt fel kell cserélni egy másik típusúra. A munkát egy TDK munka keretében kezdtük Kovács Dániel hallgatóval Király Péter doktorandusz segítségével. Kovács Dániel irodalmazás közben talált egy olyan irányelvet [OECD 2007], amelynek a mentén többéves projekt lett a témából, sok elágazással, egészen addig, hogy idén már a Journal of Chemometrics egy validációs különszámának társszerkesztője lettem, illetve 2024. márciusában a Szegeden tartott „Occam borotvája vagy gépi tanulás? Természettudományos modellezés kevés vagy sok paraméterrel” workshop ötletadója és társszervezője lettem, ahol kicsit tágabban nézve a modellezés volt a téma.

A validáció értelmezéséhez a QSAR modellekre vonatkozó OECD ajánlás (principles-guidance) [OECD 2007] adta a mi munkáinkban a keretet. Eszerint 5 alapvető szempontot kell figyelembe venni egy QSAR modell készítésekor: 1) jól definiált válaszváltozót kell modellezni (néha végpontként fordítják) 2) egyértelmű algoritmussal 3) definiálni kell a modell alkalmazhatósági tartományát 4) megfelelő paraméterek használatával validálni kell az illesztés jóságát, a robusztusságát és a prediktív teljesítményét 5) szükséges a modell és az eredmények interpretálhatósága.

Az ebben az alfejezetben bemutatott kutatásaink a 4. ponthoz kapcsolódnak. Számos szempontból vizsgáltuk ezt a kérdést: néztük a validációs paraméterek függését a mintamérettől; vizsgáltuk a paraméterek korrelációit mind egy-egy célon belül illetve a különböző célok között; alaposan analizáltunk számos validációs paramétert, hogy vajon tényleg igazak-e a szakirodalmi állítások,

különösen a túlzóan pozitív állítások; próbáltunk egyszerű validálási sémát kidolgozni mind közönséges lineáris regresszióra, mind összetettebb modellezési módszerekre; vizsgáltuk a normálás szükségességét; vizsgáltuk egyes hiperparaméterek meghatározásának módszereit; vizsgáltuk az R^2 és a Q^2 eltéréseinek határait; a robusztusság esetében a két lehetséges utat, a keresztellenőrzés és a bootstrap megoldásokat; és elemeztünk néhány adatszivárgási (data leakage) lehetőséget, pl. a tanító/teszt felosztás módszereire, hiperparaméter optimalásra, változó szelekcióra. Meg kell említenem egy folyó kutatást is, ahol kémiai modellezésben releváns jellegű és méretű adatokra nézzük, hogy a modellek optimális komplexitása a bias-variance trade-off [Geman 1992] vagy a double-descent jelenség [Vallet 1989, Belkin 2019, Schaeffer 2023] alapján határozható-e meg. A két elméletnek nincs széleskörben elfogadott magyar elnevezése, tartalmi szempontból a torzítás-pontosság kompromisszum és kettős lecsengés viselkedés lehetne az elnevezésük.

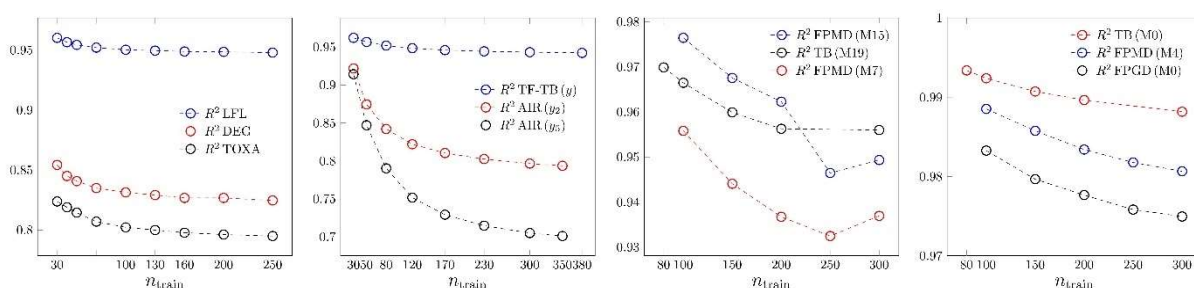
A többi QSAR elvet is érintik a kutatásaim, így a korábban tárgyalt szeriálás az 5. pontot, az interpretálhatóságot. Ma még kevésbé ismert, de az összetett új modellezési módszerek is interpretálhatóak, sőt egyes esetekben jogszabályokban rögzítetten is elvárt az interpretációjuk. A témában a 2024-es szegedi workshopunkon előadást tartottam két összefoglaló könyv alapján [Thampi 2022, Molnar 2022]. A kettős lecsengés (double descent) kérdése az algoritmusok egyértelműségét, a 2. OECD pontot érinti. Ma már kicsit szélesebben és kevésbé szigorúan követendőnek látom a 2. QSAR elvet, mint az [s59] és [s61] publikációk elkészítésekor, mivel a mai modellezések sztochasztikus elemeket is tartalmazó optimáló algoritmusai (pl. ADAM) és egyéb megközelítései (pl. „boosting” és „bagging”) nem teszik lehetővé egy-egy modell tökéletes reprodukálását.

A validációs paraméterek mintaméretfüggését az [s61] cikk ábráinak segítségével mutatom be, mert ebben a cikkben a korábbi többváltozós lineáris regressziós modelleket (MLR) [s59] kiegészítettük parciális legkisebb négyzetes (PLS), támogató vektor regressziós (SVR) és egy rejtett rétegű neurális hálós (ANN) modellekkel.

Kb. 30 különböző irodalmi adatsort modelleztünk, ezek 2-38 független változót, 100-10000 objektumot tartalmaztak széles területről: anyagtudományi, műszaki üzemeltetési, légszennyezési és különböző QSAR területekről. Adatsortól és módszertől függően 30-1000-es mintamérettel dolgoztunk, többnyire 80/20 arányú tanító/teszt halmaz felosztással. Nem mentünk bele részletesebben a tanító/teszt felosztás vitatott kérdéseibe [pl. Kraker et al. 2007, Mac Nally et al. 2018, Toher et al. 2007, Rácz et al. 2021], mint pl. a többszörös teszt halmazok alkalmazása vagy ismételt felosztások kérdésébe, hanem szándékosan az alapesetet, a teljesen véletlen egyszeres felosztást használtuk. A haladóbb módszerek hiperparamétereit külön optimaltunk az adatsorokra, de nem minden egyes mintaméretre külön-külön. Többnyire olyan kisebb komplexitású modelleket választottunk, hogy lehetőleg a szabad paraméterek száma kisebb legyen a legkisebb mintaméreteknél, vagyis kerültük az alulhatározott eseteket. A random mintavételezést 200-1000 alkalommal végeztük a jobb statisztika érdekében. Az [s61] cikkben

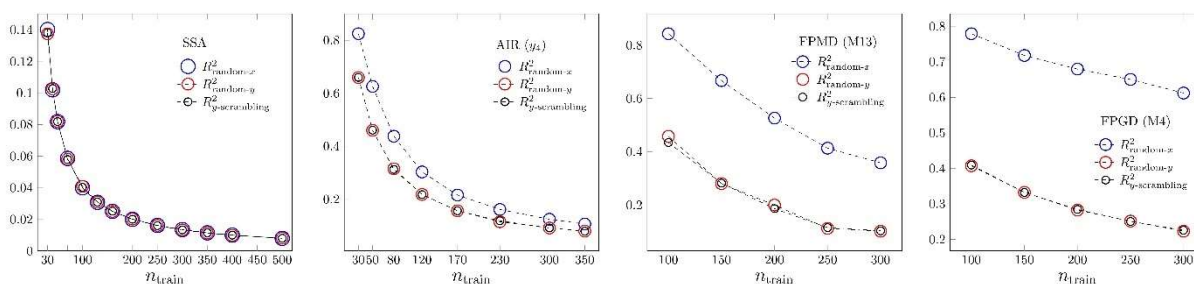
már csak R^2 és $RMSE$ jellegű validációs metrikákat használtunk, ebből is ide a dolgozatomba csak az R^2 családhoz kapcsolhatókat válogattam. Az ábrákon az adatsorok kódolva lettek jelölve, feloldásuk az eredeti publikációkban találhatóak.

Az 5.18. ábrán szereplő adatsorok és modellkódok feloldása az eredeti cikkekben találhatóak. Az ábrán jól látható, hogy az illeszkedés jóságát jelző paraméter értéke csökken a nagyobb mintaméretek felé, ez megfelel a túlillesztés sokszor leírt problémájának [pl. Rencher&Pun 1980]. Mindegyik ábrán 3-3 különböző adatsor vagy más hiperparaméterekkel készített modell szerepel. Érdekes megnézni az y tengelyen a beosztást, a haladó modellek esetében irreálisan jó az az R^2 értéke, ezek a modellek szinte mindent megillesztenek.



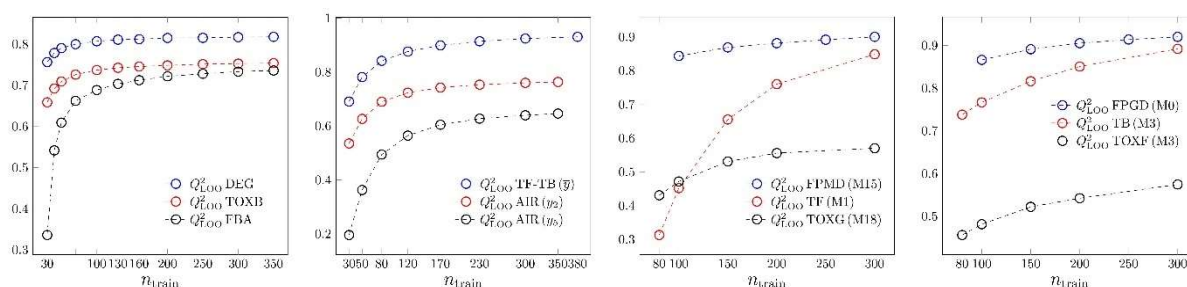
5.18. ábra. Az illeszkedés jóságát mutató R^2 paraméter mintaméret függése. n_{train} a tanító halmaz méretét jelöli. Balról jobbra: MLR, PLS2, ANN and SVR. [s61]

Az illeszkedés jóságának ez a trendje megjelenik a véletlen korreláció (chance correlation [pl. Topliss&Edwards 1979]) esetében is, ahol randomizációt követően az adatszerkezetben rejlő, de a modell valós összefüggései nélküli referencia validációját vizsgálhatjuk. Ábráinkon bemutattuk, hogy a hagyományos x és y randomizációs eljárások [pl. Rücker et al. 2007] nem azonos szintű véletlen korrelációt mutatnak az összetettebb modellek esetében (5.19. ábra).



5.19. ábra. A véletlen korreláció vizsgálata. Randomizált adatok R^2 értékei. Balról jobbra: MLR, PLS2, ANN and SVR. A random-x és random-y esetekben az eredeti x vagy y változók eloszlásai alapján generált változókkal dolgoztunk, y-scrambling esetben meg a prediktor változókhöz véletlenszerű sorrendben választottuk az eredeti válaszváltozókat. [s61]

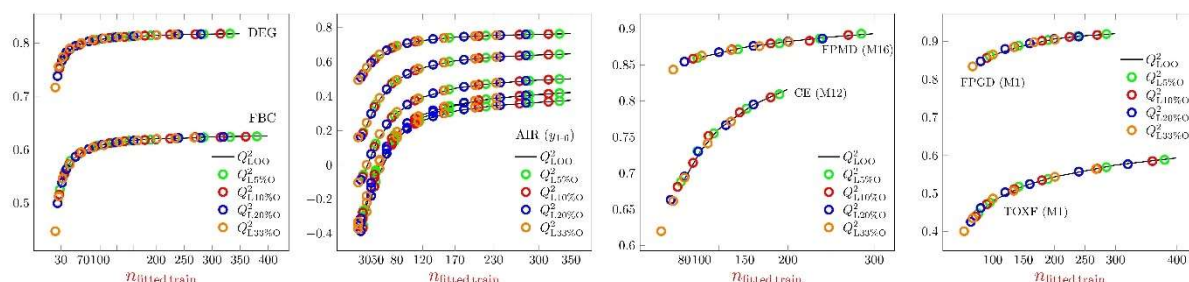
A belső validálás másik célja a modell robusztusságának a vizsgálata az OECD alapelvek szerint. Az [s61] cikkben a legegyszerűbb klasszikus keresztellenőrzési módszereket alkalmaztunk a robusztusság vizsgálatára. Nem folytunk bele a különböző ismétléses keresztellenőrzési módszerek vitatásába vagy a különböző mintázatokat követő eljárásokba (pl. réteges, „venetian blind”, ...) [lásd pl. Bro et al. 2008, Filzmoser et al. 2009, Hastie et al. 2009] Az egy elem kihagyásos R^2 csoportba tartozó Q^2_{LOO} értékeknél jól látszik az 5.20. ábrán, hogy a mintaméret növekedésével az átlagos validációs paraméter szigorúan monoton módon növekszik. Ezeknek a görbéknek az elvi határértéke az R^2 értéke MLR modelleknél, ugyanakkor PLS, SVR és ANN modelleknél, különösen a két utóbbinál szignifikánsan kisebb értékekhez tartanak (5.20. ábra). Az 5.18. és 5.20. ábrákat összehasonlítva látható, hogy a Q^2_{LOO} értékek a mintamérettel olyan értelemben változnak, hogy nagyobb mintán robusztusabb modellt alkotható. Itt nem jelenik meg a túlillesztés hatása, ami az 5.18. ábrán a mintamérettel csökkenő R^2 értékeket eredményezte.



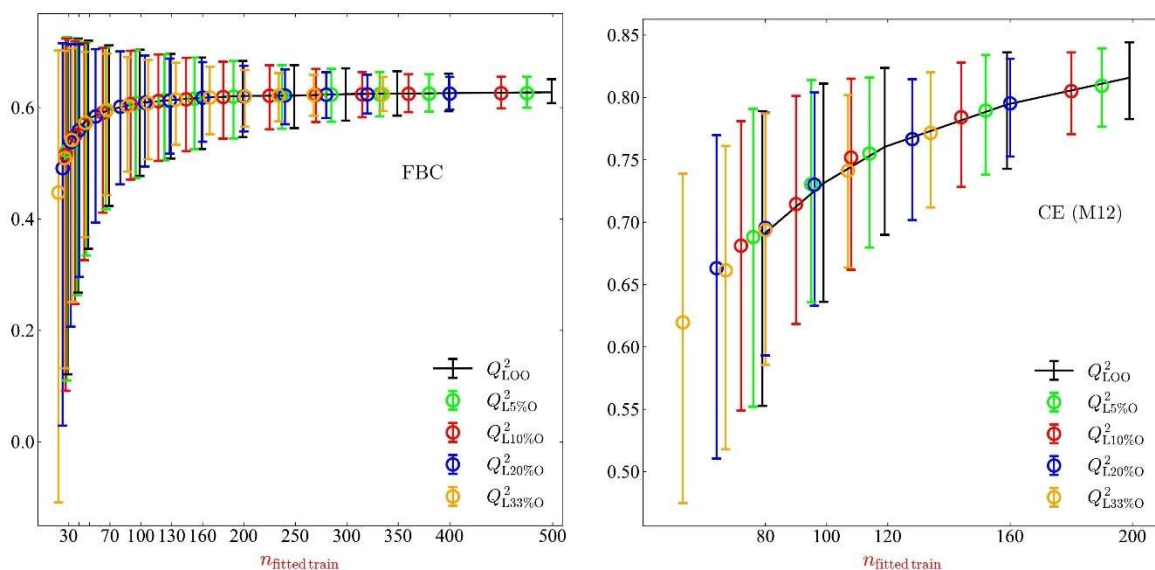
5.20. ábra. Q^2_{LOO} mintaméret függése a robusztusság vizsgálatára. Balról jobbra: MLR, PLS2, ANN és SVR modellek. [s61]

Míg az előző eredményeinkre az illesztés jóságát és robusztusságot érintve, ha nem is szisztematikusan, de töredékeiben több példát is találhattunk a szakirodalomban, a következő ábránkon az általunk felfedezett skálázási szabályt mutatom be, ami úgy gondolom, hogy akár a lezárása lehet az évtizedek óta vitatott preferálásának a több elem kihagyásos vagy az egy elem kihagyásos keresztellenőrzésnek. Az ábrán a különböző egy és több elem kihagyásos keresztellenőrzési módszerek átlagai szerepelnek a megismételt mintavételekre. Ha az x tengelyen nem a modell tanító halmazának a mérete szerint ábrázoljuk az eredményeket, hanem ebből kivonjuk az aktuális keresztellenőrzésnél kimaradt elemek számát, akkor megközelítőleg egy görbére esnek az értékek (5.21. ábra). Vagyis: mindegy, hogy hány elem kihagyásos módszert használunk, azok eredménye megfelel ennek a skálázott ábrázolásnak. Sőt, a következő 5.22. ábrán az egyedi számolások hibahatárait, pontosabban terjedelmét is feltettük, látszik, hogy mindegy melyik módszert választjuk. Ezt a jelenséget LMO-LOO skálázás néven vezettük be. A skálázást később vizsgáltuk az általunk LOO-val és LMO-val jelölt két keresztellenőrzési módszer mellett arra a változatra is, amikor csak

kettéosztjuk a tanító halmazt egy keresztellenőrző és egy ideiglenes tanító halmazra [s65]. Ilyenkor alapesetben a keresztellenőrzés csak egy kis mintán történik, nem érinti minden elemét a tanító halmaznak. A vizsgálatainkban az eljárást annyiszor ismételtük, hogy összesen n_{train} kihagyott elem tudjuk az eltérésnégyzeteket számolni, bár ez egy elemet akár többször is tartalmazhatott. Erre a keresztellenőrzési változatra is érvényesnek találtuk a skálázási szabályunkat [s65].

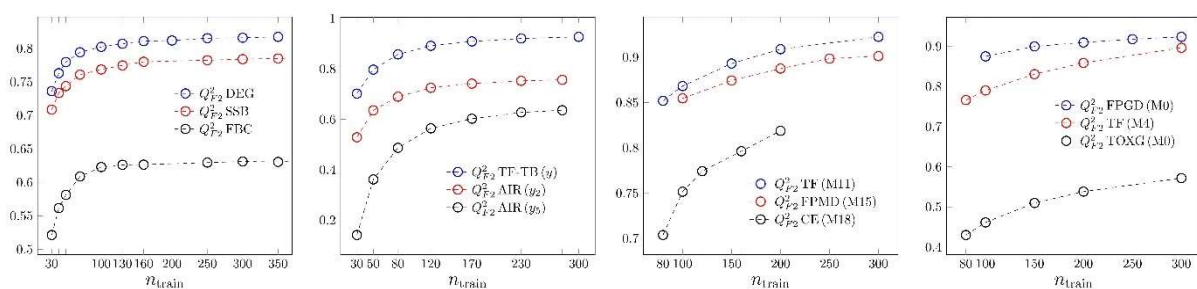


5.21. ábra A Q^2_{LOO} and Q^2_{LMO} értékek skálázhatósága, ha a paramétereket az azok számolásánál aktuálisan használt adatok számának függvényében ábrázoljuk ($n_{fitted\ train}$). Balról jobbra: MLR, PLS2, ANN and SVR modellek medián értékei. [s61]



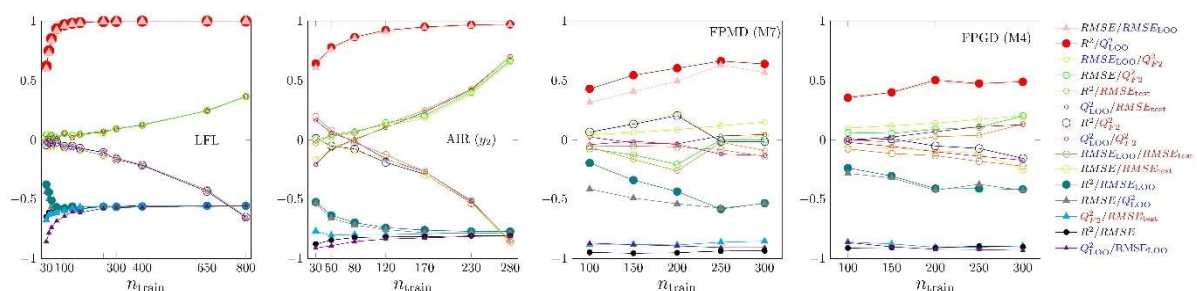
5.22. ábra. Az LOO/LMO skálázási szabály a mediánon felül a többszáz újramintavételezésekor kapott 10. és 90. percentilisek feltüntetésével. Bal) MLR Jobb) ANN [s61]

Az 5.23. ábrán a Q^2_{F2} külső teszt halmazon való validálás mintaméretfüggését mutatjuk be, amelynek lefutása nagyon hasonló keresztellenőrzéssel kapotthoz. Érdekes tapasztalatunk volt, hogy modellezéstől, adatsortól és átlag vagy medián mutatótól függően hol a Q^2_{LOO} , hol a Q^2_{F2} értékek jelöltek nagyobb nominális értékeket.



5.23. ábra Predikció vizsgálata. Q^2_{F2} mintaméretfüggése a tanítóhalmazra vonatkozóan. Balról jobbra: MLR, PLS2, ANN és SVR modellekre. [s61]

Az első kettő mintaméretfüggést vizsgáló cikkünkben bemutattunk olyan ábrákat, ahol a különböző validálási paraméterek rangkorrelációi szerepelnek. A rangkorrelációk számításának fő célja az volt, hogy egy modell minősítéskor vajon szükséges-e több validációs paramétert használni, illetve az OECD leírásban rögzített három cél mutatói között van-e statisztikus kapcsolat. Az 5.24. ábrán látszik, hogy az azonos célt szolgáló $RMSE$ és R^2 típusú metrikák általában szignifikánsan korrelálnak egymással, bár MLR esetében ez gyengébb, mint a komplexebb modellezési módszereknél. Az illesztés jóságát és a robusztusságot vizsgáló paraméterek MLR és PLS esetében főleg nagy mintáknál kifejezetten jól korrelálnak egymással, vagyis ebben az esetben az egyik mutató akár elhagyható is lenne. SVR és ANN esetében viszont a kétfajta paraméter szinte nem korrelál egymással, ahogy már említettem, az illeszkedés jósága ezeknél a modelleknél nem igazán informatív. A robusztusság keresztellenőrzéssel való vizsgálata viszont elvárt rutinjelzés minden komplexebb modellezésnél annak bemutatására, hogy a modell kellő gondossággal készült [Raschka 2020, Thampi 2022, Molnar 2022].



5.24. ábra Validációs paraméterek rangkorrelációja a mintaméret függvényében. Balról jobbra: MLR, PLS2, ANN and SVR modellek. A jobb oldali jelmagyarázatban kék jelzi a belső, piros a külső validációs paramétereket. [s61]

Érdekes összefüggést kapunk a belső és a külső validálási paraméterek korrelációira. Hosszasan elemeztük a jelenséget és sikerült kimutatnunk, hogy a kapott negatív vagy akár kis értékű pozitív

korrelációk az allokációs kérdés miatt jelennek meg. Kis mintaszámoknál nagy adathalmaz esetén az, hogy melyik adat kerül be a tanító vagy teszhalmazba kevésbé korrelál, hiszen az adatok többsége egyikbe se kerül be. Nagy mintaméretnél viszont egy könnyen reprodukálható tanító halmaz esetén a nehezen reprodukálható adatokból több van a teszt halmazban, illetve vice-versa. Ez fokozza a negatív korrelációt, ahogy ezt az [s61] cikk kiegészítő részében számolásokkal alátámasztott ábrákon is szemléltettük. Hasonló allokációs hatás az intenzív R^2 család esetében, hogy az osztóban levő TSS tag értéke az adott felosztásnál éppen milyen.

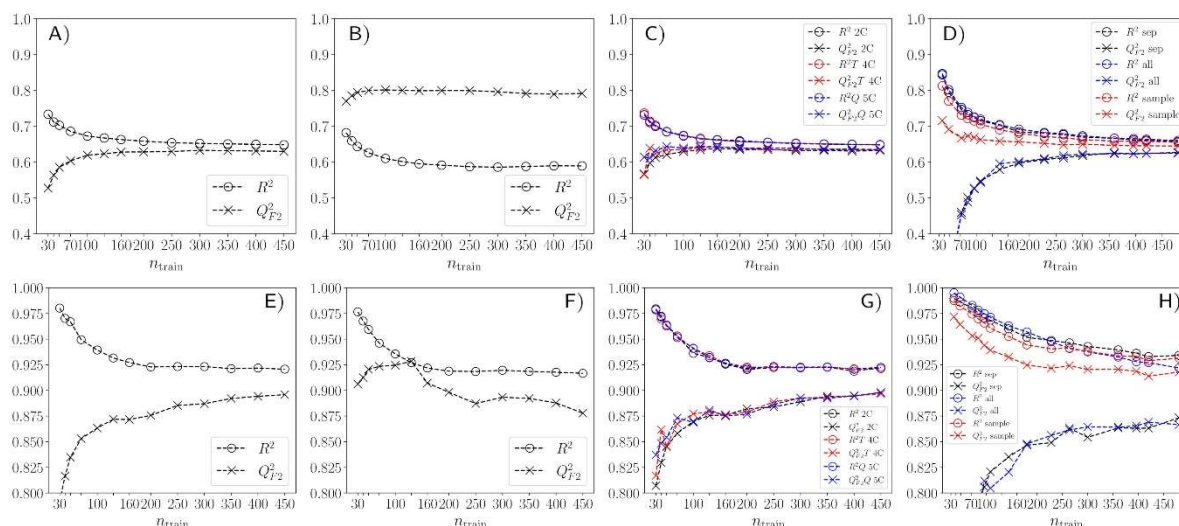
Az eddigiek a mintamérettel kapcsolatos fő következtetéseink voltak, de a több évi munka alatt számos kisebb, néha kritikai eredményt is elértünk. Ezek közül itt csak röviden említem meg:

A szabadsági fokokkal korrigált R^2 érték (R^2_{adj}) esetében az adatsortól függően véletlenszerű trendeket kaptunk a mintamérettől való függésre, ami kétségesse teszi, hogy praktikus esetben mennyire lehet ezeket az értékeket összehasonlítani, ha más mintamérettel készültek. Különösen, hogy a korrigált R^2 esetében az a szemléletes magyarázat már nem él, hogy értéke azt mutatja, hogy az adatok összehasonlítás-négyzetösszegének hányadrészét írja le a modell.

Az irodalomban részletesen pro és kontra véleményeket találtunk a CCC rövidítésű metrikáról [Lin 1989]. Bemutattuk, hogy semmilyen extra információt nem ad az R^2 jellegűekhez képest, azonban nem tartalmaz könnyen értelmezhető értékeket és ezzel a hibás modellek felismerésére is nehezebben használható.

Bemutattuk, hogy a Todeschini csoport által javasolt Q^2_{F3} számolásában [Consonni et al. 2009] konzisztensebb lenne más normáló tényezőt, a szabadsági fokkal azonosat használni. Bemutattuk, hogy PLS2 estében célszerű mindkét változó térben standardizálni az adatokat, hogy kevesebb látens komponenssel lehessen leírni a jelenségeket, valamint hogy PLS2 esetében a „repeated double cross validation” [Filzmoser et al. 2009] s értéke nem működik a komponens meghatározásnál [s61].

Bemutattuk, hogy ismételt mérés milyen hatással van a trendekre és mi a helyes kezelésük, hogy a trendeket ne sértsük. Bemutattuk, hogy a Kennard-Stone algoritmus [Kennard&Stone 1969] hogyan torzítja a metrikák értékeit és ez milyen veszélyt jelent a reális értékeléssel kapcsolatban. Az 5.25. ábrán 4 esetet mutatunk be, ahol adatszivárgás mentes referenciaként véletlen felosztás szerepel. A Kennard-Stone módszer és a rosszul értelmezett felosztása az ismételt méréseket is tartalmazó adatsoroknak drasztikusan, a trendeket is megfordító validációs paramétereket adnak. A réteges mintavételezés sokkal kevésbé befolyásolja az eredményeket, jó köztes megoldás azok számára, akik nem bíznak abban, hogy a véletlen mintavételezéssel kapott halmazok reprezentatívan kifestik a változók terét.



5.25. ábra. Az adatszivárgás hatása különböző tanító/teszt felosztási eljárásokra a tanító halmaz méretének a függvényében. A-D: egy MLR modellezés validációs paramétereinek mediánjai. E-H: egy ANN modellezés validációs paramétereinek a mediánjai. A,E) véletlen felosztás B,F) Kennard-Stone módszer C,G) réteges mintavételes felosztás 2, 4 and 5 réteggel mind a tanító és a teszt halmazokban D,H) ismételt mérések hatása. Részletek az eredeti publikációban találhatóak. [s65]

Eredményeinket egy táblázatban is összefoglaltuk, amely az általunk javasolt validálási metrikákat tartalmazza, kitérve, hogy mely esetekben mely értékek hagyhatóak el (5.3. táblázat).

Ezt a táblázatot valamennyire árnyalhatjuk, ha három további kutatásunkat is figyelembe vesszük. Ezek egyike talán az időben első validációs cikkünk, ahol azt vizsgáltuk, hogy mikor tekinthetjük szignifikánsnak a Q^2 és R^2 eltérését. A szakirodalomban az R^2 megítélése viszonylag egységes, legfeljebb a szabadsági fokokkal való korrekció kérdése szokott vitatott lenni. Az is újabban elfogadott, hogy a modern modellezési eljárások esetében többnyire nagyon jó az illeszkedés jósága, de ez nem jelenti feltétlenül a jó predikciós képességet. Ugyanakkor a Q^2 és azon belül is a Q^2_{LOO} megítélése szélsőséges. A keminformatikai vonalon rengeteg cikk támadta a Q^2 létjogosultságát [pl. Doweyko 2008a, 2008b], akár olyan sokkoló címmel is, hogy „Beware of Q^2 !” [Golbraikh&Tropsha 2002]. Nem szeretnék itt most a részletekbe belemenni, de több szempont alapján is vizsgálták a kérdést, és be kell vallanom, hogy szinte mindegyik esetben található olyan probléma az indoklásként szereplő számolásokban, amivel magyarázható az általunk tiszta trendektől való eltérése a Q^2 -nek, illetve az adott helyen a Q^2_{LOO} és Q^2_{LMO} közti eltérések furcsasága. A cikkekben szereplő adatoknál gyakran előfordul az adatszivárgás esete, vagyis amikor valamilyen módon a keresztellenőrzésnél használt eljárás során az adott keresztellenőrzési modell már kicsit függ a kihagyott elemtől. Ilyen eset például, amikor a modell hiperparamétereit előre leválasztott validációs szett helyett csak keresztellenőrzéssel optimalizálják [pl. Baumann 2003], vagy mesterségesen növelik az adatmennyiséget annak érdekében, hogy pl.

osztályozásnál kiegyenlített osztálylétszámokat kapjanak, vagy ismételt mérések is szerepelnek az adatsorban. Ez utóbbi két esetben a keresztellenőrzéskor a helyes eljárás az adott minta összes mérésének a kihagyása egyszerre, nem csak egy-egy mérés kihagyása [s61] [s65]. Külön kiemelném, hogy az LOO-LMO skálázásunk alapján várhatóan a skálázatlan LMO értékek kisebb Q^2_{LMO} számot mutatnak, mint a Q^2_{LOO} értékek, de ez korrigálható a skálázás alapján. Hagyományosan viszont inkább azt állították, hogy a Q^2_{LOO} „overoptimistic”, pusztán azért, mert nagyobb Q^2_{LOO} értéket ad, mint a Q^2_{LMO} . Ebben az utolsó témában, a Q^2_{LOO} reputációjának megmentésében készítettük az [s65] publikációt. Az [s65] publikációban bemutattuk, hogy számos cikkben kritika nélkül és pontatlanul vették át az egyik alapcikk [Shao 1993] megállapításait a több elem kihagyásos módszerek preferálásával kapcsolatban. Számos esetben csak a különböző keresztellenőrzési módszerek egységes elnevezésének hiány okozta a félreértést.

5.3. táblázat. Az általunk javasolt validálási mutatók regressziós modellekre. A vastagon szedett mutatók a javasoltak, a normál tipográfiájúak esetleg feleslegesek egy adott esetben vagy nem adnak új információt. A piros megoldás a szubjektív javaslatom az elsődlegesen predikciós célra készült regressziós modellekre. [s61]

	Belső validálás		Külső validálás	Ügyeljünk:
modell	illeszkedés jóságát	robusztusság (keresztellenőrzéssel)	prediktivitás	
MLR	R^2 és/vagy $RMSE_{\text{training}}$	elhagyható nagy mintaméretnél $Q^2_{\text{LOO}} (\approx Q^2_{\text{LMO}})$ és/vagy $RMSE_{\text{LOO}} (\approx RMSE_{\text{LMO}})$	Q^2_{F2} és/vagy $RMSE_{\text{test}}$	-
PLS2		x-y standardizálása a látens változók számának a kiválasztásakor		
ANN	esetleg R^2 és/vagy $RMSE_{\text{training}}$	$Q^2_{\text{LMO}} (\approx Q^2_{\text{LOO}})$ és/vagy $RMSE_{\text{LMO}} (\approx RMSE_{\text{LOO}})$		standardizálás hiperparaméterek optimalása ellenőrizzük a támogatvektorok számát
SVR				

Az 5.3-as táblázat nagyjából megfelel az [s61]-s publikációban levőnek annyi eltéréssel, hogy a piros kiemelés ott hiányzott. Az eredeti ábra megjelenése óta levő kutatások és az az azokkal kapcsolatos

irodalmazás után kicsit átalakult a véleményem, különösen [Esbensen&Geladi 2010, Molnar 2022, Thampi 2022] írásai alapján. A kemometriai modellek többféleképpen feloszthatóak. Egy lehetséges kettéosztás a modellezés elsődleges célja, pl. a modell alkalmazása későbbi predikcióra, vagy a modell fő célja az adott jelenség természettudományos értelmezése és megismerése. Az elsődlegesen predikcióra készített modellek valódi teljesítőképességét az mutatja, hogy a későbbi alkalmazások során hogyan teljesítenek. Egy ideális minősítő halmaznak ezeknek a későbbi alkalmazásoknak kell megfelelnie, nagyon nagyra kell lennie, hogy a bias-variance tradeoff variancia tagja kicsi legyen a minősítő halmazon, tartalmaznia kell a későbbi alkalmazások teljes varianciáját, adatszívárgásmentesnek kell lennie a modell fejlesztésénél használt halmazokkal szemben, hogy a validációs paraméterek ne legyenek torzítottak. Ezeknek a feltételeknek csak a külső validáció teszt halmaza felelhet meg az adatszívárgási kritérium miatt. Így arra következtetésre jutottam, hogy kicsit szembe menve a kemometriai kutatások elmúlt évtizedeivel, tökéletesnek csak az ilyen teszt halmazon történő validációt tekintem predikcióra készült modelleknél. A modell fejlesztésénél persze minden más validációs módszer használható, de a végső minősítés egy ilyen ideális teszt halmaz alapján kellene, hogy megtörténjen. Sőt, szubjektív módon a legegyszerűbben értelmezhető $RMSE_{test}$ és Q^2_{F2} értékeket erre elegendőnek tartom regressziós predikció esetén. Ez a külső validációt propagáló szemlélet azzal is alátámasztható, hogy így nem különböztetjük meg az illesztéssel kapott modelleket a teljesen elméleti megfontolások segítségével levezetett modellektől, mivel az utóbbiaknál a belső validáció nem lehetséges. Elsődlegesen természettudományos értelmezési célra készült modelleknél továbbra is a teljes sémát gondolom elfogadhatónak.

Térjünk vissza azonban a legelső validálási cikkünkre, az [s48]-ra!

Nagy hatású pontok detektálása keresztellenőrzési és illeszkedési validációs paraméterek alapján [s48]

Modellezéskor többféle kategóriákba oszthatóak az átlagosaktól eltérő adatpontok. A kiugró (outlier) megnevezés azokra a pontokra vonatkozik, amelyek akár a magyarázó, akár a válaszváltozók terében az átlagtól messze, többnyire 2-3 szóráson kívül esnek, sokszor kihagyásos módon és Mahalanobis térben számolva. Ezek egy speciális esete az ún. emelő pont (leverage), amelyek olyan, többnyire mindkét térben kiugró pontokat jelentenek, amelyek rendkívül nagy hozzájárulással rendelkeznek a modell paramétereinek meghatározásában. Ettől eltérően a nagy hatású (influential, másként hatásos vagy befolyásoló) pontok nem feltétlenül lógnak ki bármelyik térben, de általános leírás szerint „a többi megfigyeléshez hasonlítva viszonylag nagy hatással bírnak a becsült értékekre, pl. a válaszváltozó, a regressziós paraméterek és a hiba értékeire” [Frank&Todeschini 1994]. A téma az OECD pontjai alapján a robusztusság ellenőrzéséhez kapcsolódik, vagyis az adatsorra való érzékenységhöz, illetve a bias-variance trade-off elméletben a variancia taghoz.

A klasszikus regresszió két megközelítést alkalmaz ezek azonosítására, ezek egyike a Cook nevéhez is kapcsolható klasszikus reziduális analízis [Cook&Weisberg 1982], illetve a másik a robusztus regresszió [Rousseeuw&Hubert 1997]. Az első esetben a teljes adatsoron végzett modellillesztés után a maradékok és más mutatók alapján keressük meg azokat a pontokat, amelyek emelő és/vagy nagy hatású pontok. Ezek után különböző megfontolások alapján elhagyunk egyes pontokat és új modellt építhetünk ezek nélkül. A robusztus regresszió során a modell építését megelőzően, vagy abba integráltnak, kihagyjuk vagy kis súlyúra állítjuk be ezeknek a pontoknak a hatását, így a modell paraméterezésekor már egészében vagy részben kiiktattuk azokat.

[s48]-as publikációnkban abból indultunk ki, hogy pl. QSAR modellek esetében rutinszerűen közlik az R^2 és a Q^2_{LOO} (vagy Q^2_{LMO}) értékeket, míg viszonylag gyakori, hogy további részletei a modellek validálásának nem szerepelnek. Az így megadott értékek értelmezése azonban meglehetősen ad hoc, pl. közismertek a különböző ökölszabályok, pl. $Q^2_{LOO} > 0.5$, vagy R^2 - Q^2_{LOO} közti különbségre vonatkozó tapasztalati megjegyzések. Az [s48]-as publikációban számos ilyen irodalmi ökölszabály található a megfelelő hivatkozásokkal. Ha azonban megvizsgáljuk a két szám számítási egyenleteit, levezethető, hogy (lásd a jelölések definícióit a függelékben is)

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2, RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2, MSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad 5.18. \text{ egyenlet}$$

jelöléseket használva

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} \quad 5.19. \text{ egyenlet}$$

, ahol az MSS/TSS állítás csak teljes hagyományos többváltozós lineáris regresszióban igaz, ahol $TSS=MSS+RSS$.

A keresztellenőrzésnél az RSS-tagot felváltja a PRESS:

$$PRESS = \sum_{i=1}^n (\hat{y}_{i/i} - y_i)^2 \quad 5.20. \text{ egyenlet}$$

ahol $\hat{y}_{i/i}$ azt a modellezett értéket jelöli, ahol épp az i -dik elem nem szerepelt a modellalkotás keresztellenőrzési fázisában. Ezek alapján

$$Q^2 = R^2_{cv} = 1 - \frac{PRESS}{TSS} \quad 5.21. \text{ egyenlet}$$

$R^2 \in [0; 1]$ MLR modellezésre, illetve $R^2 \in (-\infty, 1]$ általánosan, míg $Q^2 \in (-\infty, 1]$ minden esetben.

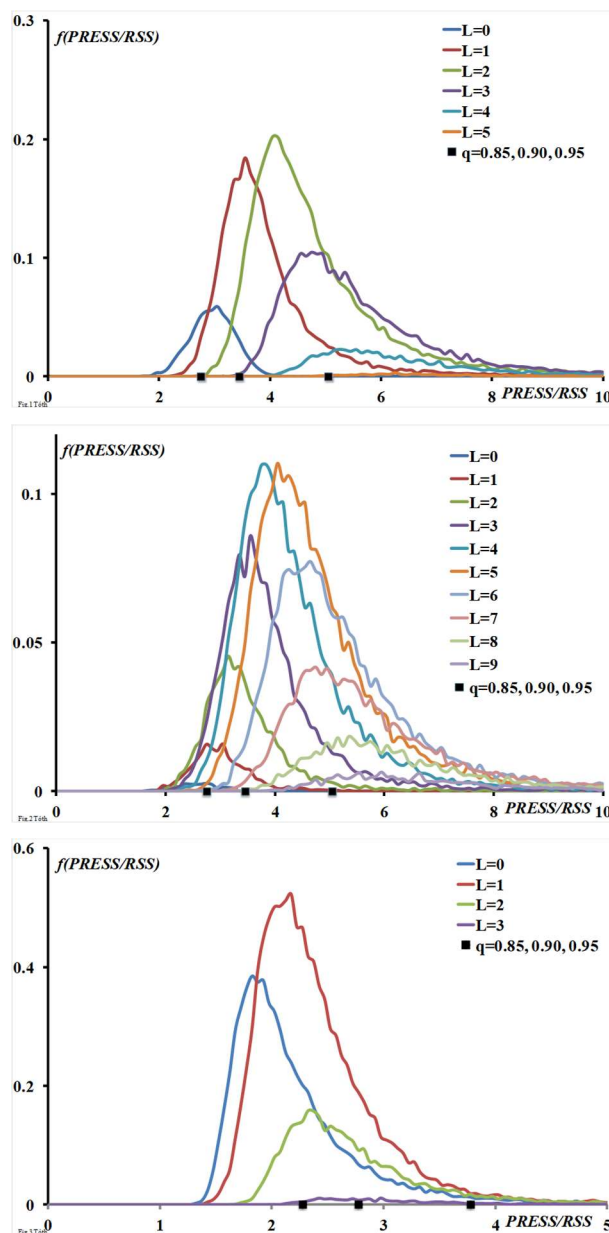
Feltételezve egy optimális adatsorra vonatkozó helyzetet azt várhatjuk, hogy RSS és $PRESS$ ugyanabból mintából származik, vagyis

$$\frac{PRESS/df_{PRESS}}{RSS/df_{RSS}} = \frac{(1-Q^2)}{(1-R^2)} \approx F - \text{eloszlást követ} \quad 5.22.$$

, ahol df a megfelelő szabadsági fokot jelenti.

Az 5.22. egyenletben levő hányados értékére ennek megfelelően F-tesztet végezhetünk azzal a H_0 és H_a hipotézissel, hogy a két szórásnégyzet azonosnak tekinthető-e, vagy ez statisztikai alapon elvethető. A szakirodalomban használt nagy hatású pontok azonosítására használható módszerek

közül három, a COOK-2, DFBETAS és a DFFITS-1 módszereknél [Frank&Todeschini 1994] jó korrelációt találtunk a teszt és ezen mutatók között. Az adatokat megnéztük különböző F kritériumok szerint is, illetve szimulált (5.26. ábra) és valós adatokon is.



5.26. ábra A COOK-2 (fent), DFBETAS (középen) és DFFITS-1 módszerekkel detektált módon különböző számú hatásos pontot (L) tartalmazó adatsorok relatív gyakorisága a $PRESS/RSS$ hányados függvényében. A fekete pontok három percentilisét jelentik az F-tesztnek a $PRESS/RSS$ hányadosra. A szimulált adatsorok generálása az eredeti cikkben szerepel. [s48].

Két másik nagy hatású pontokra kifejlesztett mutatónál nem kaptunk ilyen korrelációt, mivel azok igazából emelő pontok esetére lettek kifejlesztve, amelyek mind az RSS , mind a $PRESS$ értékeket

hasonlóan befolyásolják. Meg kell jegyeznem, hogy a *PRESS* esetében MLR regressziónál és a *PRESS* és az *RSS* estében a PLS regressziónál is belefutottunk a szabadsági fokok kérdésébe az 5.22. egyenletnél. Végül egy χ^2 eloszlásra visszavezetett módon megoldottuk a kérdést és az jött ki, hogy kb. azonosnak tekinthető az *RSS* és a *PRESS* szabadsági foka, vagyis elhagyható az 5.22. egyenletből.

Az általunk javasolt tesztet néhányan alkalmazták, akár úgy is, hogy Tóth-féle tesztként hivatkoztak rá [pl. Hohmann et al. 2015, Khazaei et al. 2015, Baghgoli et al. 2018]. A tesztrel gyorsan ellenőrizhetőek pl. QSAR modellek abból a szempontból, hogy megfelelően robusztus modell jött-e létre az adott adatsoron.

A robusztusság vizsgálatának két alapszere [s64]

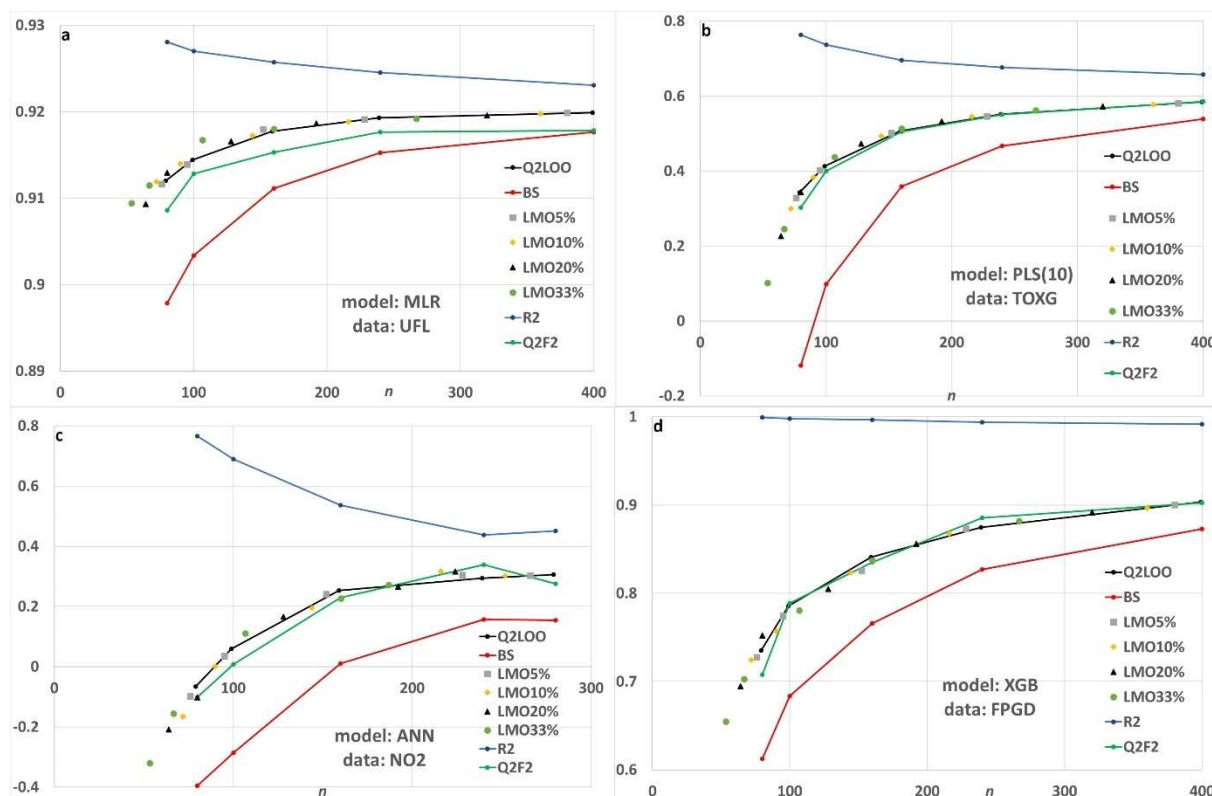
Az OECD útmutató 4. pontja leírja a három validálási szempontot, illetve módszereket is javasol ezekhez [OECD 2007]. A belső validáció robusztusságra vonatkozó része alapjában két eljárás segítségével végezhető el: keresztellenőrzéssel és bootstrap módszerrel.

Keresztellenőrzés során a korábban leírt módon egy vagy több elemet hagyunk ki a tanító készletből, azok nélkül modellt építünk és a kihagyott elem(ek)re számoljuk a mért és a számított válaszváltozók közti eltéréseket. Alapesetben a kihagyást addig ismételjük, amíg minden elem pontosan egyszer marad ki és a számolt *PRESS* érték alapján képezzük a megfelelő Q^2 vagy $RMSE_{CV}$ mutatókat.

Bootstrap esetén N_{bs} alkalommal N_{tr} elemet választunk a tanító halmazból, de megengedve az ismétléseket is. Ezeken az újra mintavételezett mintákon modelleket tanítunk be és ezekkel számoljuk a becslt és eredeti válaszváltozók közti különbséget [Hastie et al. 2009]. A bootstrap módszernek nem ez a leggyakoribb alkalmazása, inkább konfidencia intervallumok nem paraméteres (valójában torzított) meghatározására és modell paraméterek bizonytalanságának becslésére szokták használni, illetve újabban alapszere a bias-variance elmélet variance tagjának csökkentésére az ún. „bagging” eljárásen keresztül [Thampi 2022, Molnar 2022].

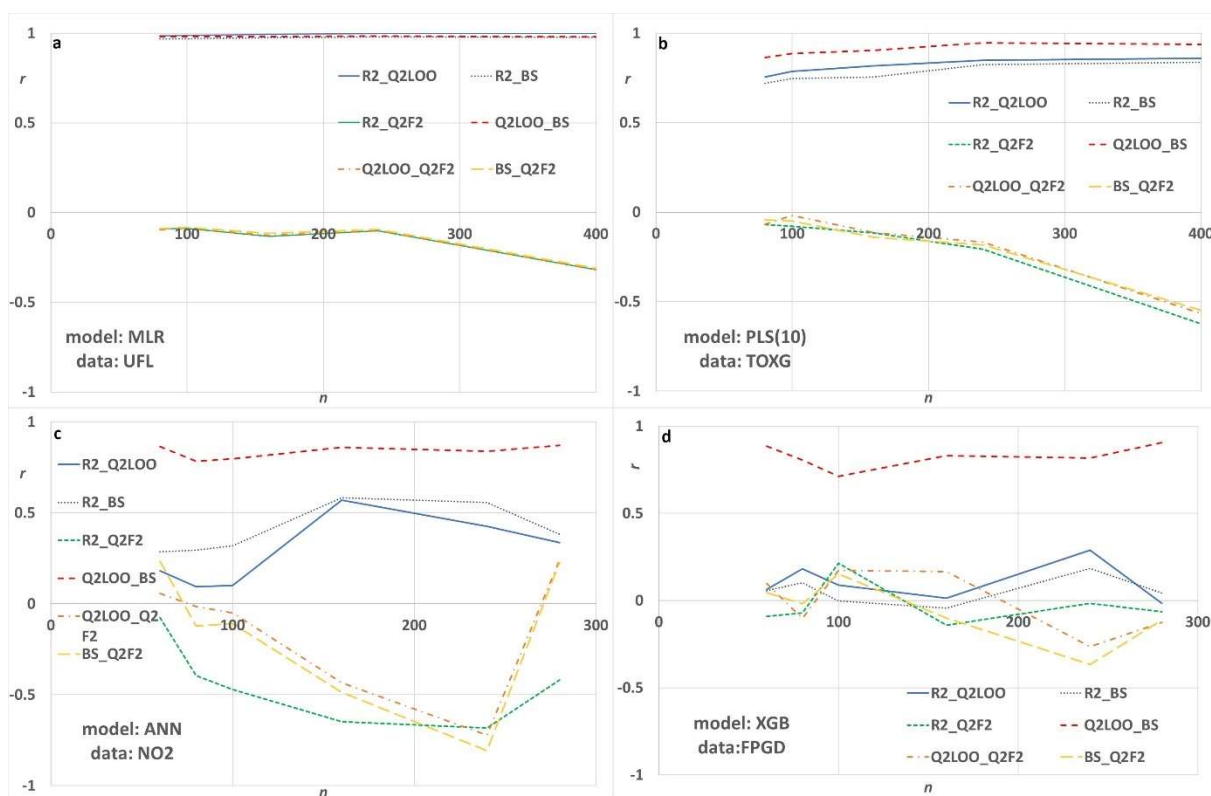
Ha megvizsgáljuk, hogy mely elemeken szokás egy-egy bootstrap minta esetén az eltérésnégyzeteket számolni, többfajta megoldást találunk. Az aktuális bootstrap mintán való számolás illeszkedés jósága jellegű paraméter, a külső teszt készleten való pedig inkább predikciós paraméter. A robusztusság esetében az OECD [OECD 2007] olvasatában az adott mintában éppen nem szereplő tanító halmazbéli elemek jöhetnek szóba, ezt „out-of-bag” elemeknek nevezik. Nagy minták esetében ez kb. 37 százalékát jelenti az N_{tr} elemeknek, tehát viszonylag kismintás számolást. A szakirodalomban létezik a különböző értékek keverése is [Hastie 2009], pl. a 0.632 mutató esetében ennyi részben R^2 érték és ahhoz hozzáadva 0.368 részben „out of bag” érték kombinációjából egy mutatót alakítanak ki a több mutatós OECD szemlélettel szemben. Esetünkben azonban egyedül az out-of-bag kompatibilis az OECD robusztusságra vonatkozó megközelítésével, így ebben a munkában ezeket az értékeket hasonlítottuk össze a keresztellenőrzés egy és több elem kihagyásos változataival.

14 adatsoron PLS és ANN modelleken vizsgáltuk a validációs paramétereket, ismételten a mintaelemszámtól való függésből kiindulva. Mint az 5.27. ábrán látszik, a bootstrap értékek hasonló méretfüggést mutatnak, mint a keresztellenőrzési és a predikciós mutatók, de az értékük numerikusan kisebb, mint a Q^2 paramétereké. Ennek egyik oka, hogy a valós statisztikai indok nélkül ismételt méréseket tartalmazó módszerek általában gyengébbek az azonos méretű ismétlést nem tartalmazóknál.



5.27. ábra. Validációs paraméterek mintaméretfüggése 4 modellezési módszernél 4 különböző adatsoron. BS jelöli a bootstrap out-of-bag paramétert. További részletek az [s64] publikációban találhatóak. [s64]

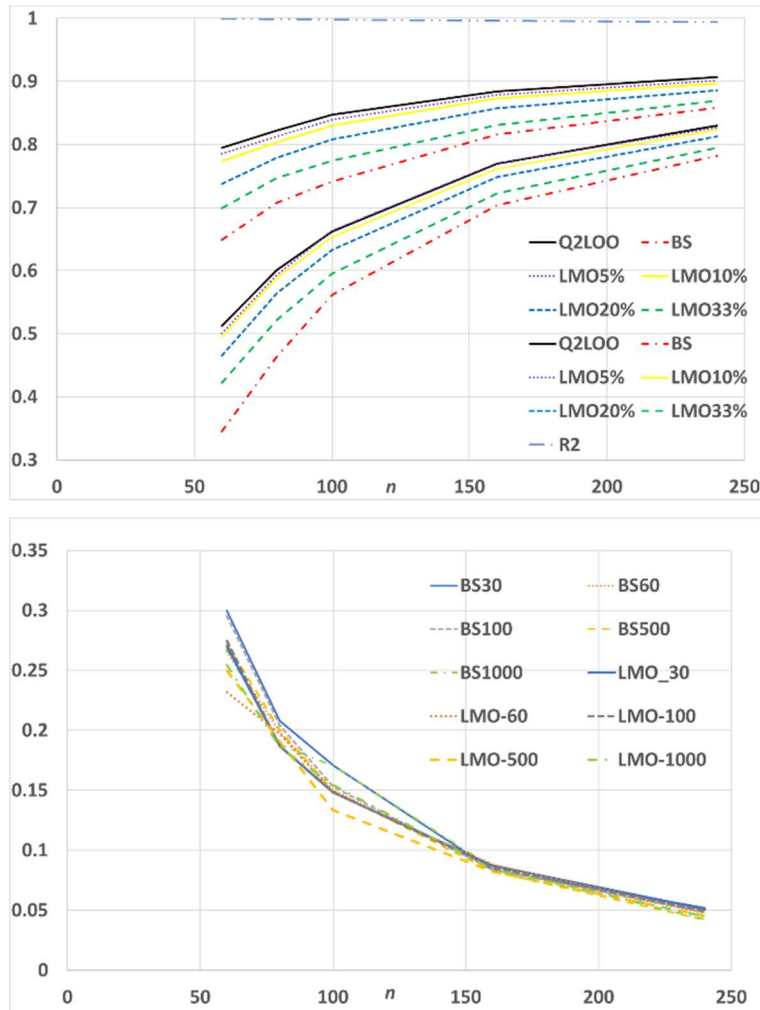
Az 5.28. ábrán látható, hogy a bootstrap és a keresztellenőrzési módszerek rangkorrelációja nagyon hasonló, vagyis hasonló információt hordoznak, egymással különösen jól korrelálva.



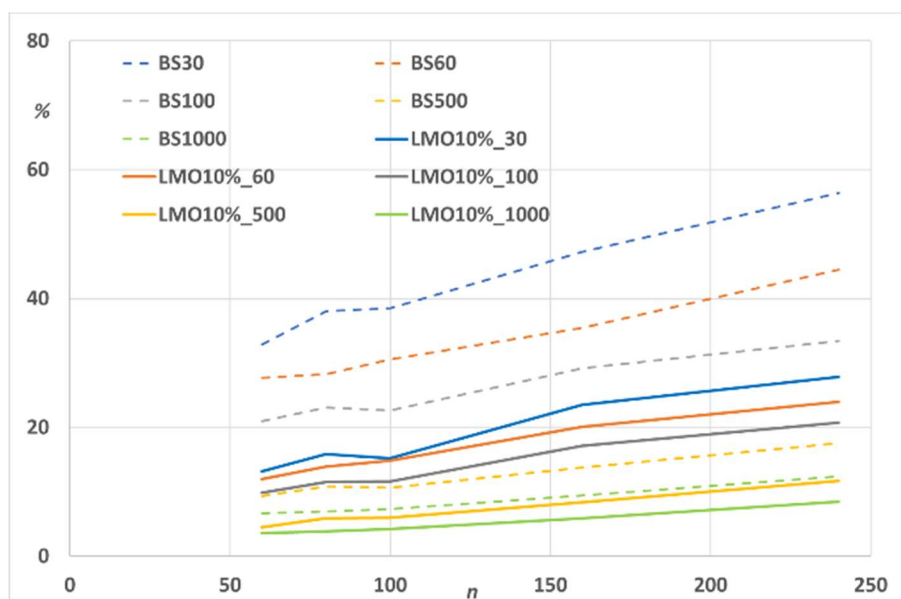
5.28. ábra. Validációs paraméterek párjainak rangkorrelációja mintaméretfüggéssel. 4 modellezési módszernél más-más adatsoron számolva. BS jelöli a bootstrap out-of-bag metrikát. További részletek az [s64] publikációban találhatóak. [s64]

Lényeges kérdés, hogy azonos számítási igény mellett melyik módszer ad kisebb szórású (pontosabban standard hibájú) eredményt. Számításaink szerint, a két módszer közel azonos szélességű intervallumon belüli értékeket ad azonos számítási igény esetén (5.29. ábra).

Azonban, ha alaposabban megnézzük az eredményeket, kiderül, hogy ez a variancia alapján a tanító halmaz választásából ered (kb. a variancia rész a bias-variance trade-off elméletből) és csak egy kis része köthető a bootstrap vagy a több elem kihagyásos keresztellenőrzés belső választásához. Az egy elem kihagyásos keresztellenőrzésnél nincs is ilyen variancia komponens (5.30. ábra).



5.29. ábra. A módszerek bizonytalansági intervallumai. Fent: A különböző validációs paraméterek 5-95 %-os percentilisei 100 bootstrap minta esetén az ismételt számolások során FPGD adatsor XGB modellezéssel. Lent: A fenti ábrához hasonlóan számolt eltérések az 5-95% percentilisek között bootstrap és LMO 10% számításokban, amennyiben a bootstrap-LMO 10% párok ugyanakkora számítási igényt jelentettek. TF adatsor PLS modellezéssel. További részletek és a az adatsorra és modellre vonatkozó kódok feloldása az [s64] publikációban találhatóak. [s64]



5.30. ábra. A standard hiba forrása azon százalékainak mintaméretfüggése, amelyek a bootstrap vagy a 10%-os LMO keresztellenőrzés technikai kivitelezéséből származtak. TF adatsor PLS modellezéssel. [s64]

Összességében megállapítottuk, hogy a robusztusság vizsgálatára alkalmazható két fő módszer egyikénél se kaptunk olyan eredményeket, amelyek alapján valamelyik ne lenne alkalmazható erre a célra. A két megközelítés eredményei erősen korrelálnak adott mintafelosztás esetén. Ugyanakkor a keresztellenőrzési értékek szám szerint közelibbek a külső validáció értékeihez és jól konvergálnak egyszerűbb (pl. MLR vagy PLS) modellek esetén az illeszkedés jóságát mutató metrikákhoz nagy mintaszámoknál. Vagyis ilyen szempontból könnyebb az interpretációjuk. Azonos számítási igény esetén a keresztellenőrzés valamivel kisebb varianciájú módszer a bootstraphez képest, bár ez az öröklött hiba viszonylag kicsi a mintafelosztásból származó varianciához képest. Ezek alapján inkább a keresztellenőrzési módszereket ajánljuk a robusztusság becslésére, főleg, mivel itt a skálázási törvény miatt egyenértékű választás az egy vagy a több elem kihagyásos megoldás.

6. Összefoglalás

Egy szokásos, maximum 2-3 témát magába foglaló értekezés esetén az összefoglalás rövid átismétlése a bemutatott kutatásoknak és a főbb eredményeknek, némi kitekintéssel. Az esetemben három nagy részre tagoltam az anyagot, de valójában jóval több inhomogén témáról van szó.

Ha a dolgozat címére gondolok, „Adatok értelmezése szimulációkkal és a többváltozós statisztika eszközeivel a kémiában”, egyből felmerül, milyen statisztikát lehetne magáról a dolgozatról készíteni. Lássunk néhány adatot az összes eddigi publikációmra (65 cikk, beleértve azokat a közleményeket is, amelyek valamilyen okból az értekezés leadásakor nem voltak beszámíthatóak): 47 cikkben vagy valamilyen szimulációs módszert használtunk, vagy szimulált adatsorokat készítettünk valamilyen mintázatnak megfelelően. 18 cikkben lett új módszer bevezetve, esetleg elméleti levezetéssel, valamint ide számoltam azt is, amikor módosítva lett egy korábbi módszer. 14 cikket nevezhetek témája szerint szinte kizárólagosan kemometriai cikknek, 11 cikkben történtek kvantumkémiái számítások is, 9 cikkben szerepel valamilyen rövidebb-hosszabb csak elméleti levezetés, 4 cikk összefoglaló jellegű vagy csak véleményt közöl, 3 cikk oktatási-képzési kérdésekről szól.

Ha a közös vonást keresem ezekben, látható, hogy a kutatói munka módja és a használt elméleti-gyakorlati eszköztár a közös, a témák vegyesek. Ez a kapcsolat a témák között hasonló egy műszeres kolléga munkásságához, aki egy műszercsaládot alkalmaz vegyes kutatási területeken.

Az összefoglalás egy önértékelés is. Ha a tudomány komoly-nehéz oldaláról kell közelítenem az eddigi eredményeimet, akkor három olyan eredményt/területet emelnék ki, amit jelentősebbnek tartok.

A párkorrelációs függvény fluktuációjának meghatározása igazi példája annak, amikor numerikus kísérlet vezetett el egy elméleti felfedezéshez. A párkorrelációs függvény a kondenzált anyagokkal kapcsolatos szimulációk és diffrakciós mérések egyik legfontosabb függvénye, sőt talán a legfontosabb. Mivel átlag jellegű függvényről van szó, a fluktuációjának ismeretét alapvető információnak tekintem. [s15,s18]

Klasszikus fizikokémikusi szemmel a statisztikus mechanikai potenciálmeghatározás inverz problémájára kidolgozott négy módszerre vagyok a legbüszkébb. A Born-Green-Yvon hierarchia RMC módszerrel való összekapcsolását Baranyai Andrással közösen végeztük, ő inkább az elméleti résszel foglalkozott, én a gyakorlati megvalósítással, ha ide értem a kockás füzetes több heti paraméteres integrálást is. A módszer megalkotása és kipróbálása már a levegőben volt az RMC módszer létrejötte óta, de mi vállaltuk be először a feladatot. Az inverz Monte Carlo módszer továbbfejlesztése egyedi ötletem volt, szinte teljes tudományos vákuumban végeztem, minimális tudományos visszajelzéssel.

Több mint 10 év kellett Hanke-nak és munkatársainak, hogy újra felfedezze a módszeremet, addig senkit se érdekelt. A neurális hálós megoldás tudományos játéknak számított akkor, a közélet csodálkozott rajta, felbolygattam vele egy konferenciát, de még a következő években is kevesen mertek feketedoboz megoldásokat keresni hasonló fizikai problémák megoldására. Az energia illesztéses megoldásom szintén 10 évig nem érdekelt senkit se, a szakdolgozóm alig csúszott át a továbbfejlesztési próbálkozással. Szerencsémre William Noid csoportja valahol belebotlott a cikkbe, és ebből kiindulva dolgozták ki a kombinált célfüggvényes módszereiket. Máig is úgy gondolom, hogy a klasszikus kutatói képhez, a nehéz egyenletek és bonyolult szimulációk toposzával, ezek az eredményeim illeszkednek a legjobban. [s26, s27, s29, s31, s36]

A harmadik jelentősebb eredményemnek a validáció szisztematikus vizsgálatát tartom. Ezen a területen fokozatosan jutottunk el arra a szintre, hogy egy szélesebb megértés után bátran ki mertünk állni néhány tákolmány jellegű validációs metrika ellen, sikerült bemutatnunk több eljárás veszélyét és végül sikerült kidolgozni egy jól átlátható egyszerű validációs alapsémát. Ugyanakkor, ha egy tudományos eredményt kell felsorolni itt, akkor az egy elem kihagyásos és több elem kihagyásos keresztellenőrzési módszerek skálázási szabálya az, amire a büszke lehetek. [s52, s58, s59, s61, s64, s65]

Ha hallgatók az elmúlt években kérdeztek, milyen hozzáállást javasolnék a tudományhoz, akkor sokszor a játékoságot említettem nekik. Saját tudományos játékként két olyan eredményemet szeretném kiemelni, amire büszke vagyok:

A karbamid kristályosnövekedés kinetikai Monte Carlo szimulációs alapkérdése önmagában csak egy tudományos játéknak volt felfogható: tartalmaz-e a makroszkopikus kristályforma annyi információt, hogy mikroszkopikusan definiálható sebességi állandókat meghatározhassunk belőle. A kidolgozott módszer és az eredmények alapján igen. Ki gondolta volna előre, hogy sikerül? [s39]

A másik tudományos játék a szeriálás. Háttér információ és irodalmi ismeretek nélkül egy bő félév alatt jutottunk el Szepesváry Pállal oda, ahová a tudomány kb. száz év alatt: vagyis érdemes a sor-oszlop permutációval matematikai célfüggvény szerint rendezni az adatmátrixokat. Amikor az első cikk közzlése után megtudtuk, hogy egy létező területen kutatunk, már egyszerűbb volt az út, csak olyan célfüggvényt kellett kigondolni, ami a mi eredeti adatstruktúránkra jobban működik. Ez lett a foltszeriálás a lokális hasonlósági mátrix alapján. [s41, s53, s63]

Egy összefoglalásban természetesen illik a jövőbe is tekinteni. Nyilvánvalóan vannak megkezdett vagy tervezett témák.

A klasszikus szimulációs területről említettem a memóriefüggvény kérdését. Itt Király Péterrel eljutottunk odáig, hogy milyen reprezentációban érdemes eltárolni egy coarse graining potenciál esetén a múltat. Klasszikus kemometriai alapon ezekből az adatokból nem sikerült megfelelő memória függvényt összeállítani, de tervünk, hogy egyszer pl. neurális hálókkal a gépi tanulást vetjük majd be. A dolgozat írásakor végiggondoltam az inverz problémára kidolgozott módszereimet is. A konfigurációs hőmérsékletnél említettem, a Clausius-féle ekvipartíció kérdését. Az itt számolható értékek is eloszlásokkal bírnak egyensúlyi rendszerekre. Ezen eloszlások ismerete talán lehetőséget adhat egy új coarse graining potenciál meghatározásra, talán egy iteratív szimulációk nélküli módszerre.

Az adatmodellezés oldalon a kettős lecsengés jelensége, aminek drasztikus hatása lehet a jövőben. Ennek a vizsgálatát a kemometriában megszokott méretű irodalmi adatokon folyamatban van. Egy kis elméleti játékot is tervezek itt, a neurális háló koefficiensek statisztikus mechanikai feldolgozását. Hasonlóan újabb témának gondolom a fekete doboz modellek interpretációjának a kérdését, itt a folyadékok integrálegyenletes témáinál szeretném folytatni a már elkezdett munkát. Természetesen rengeteg a kisebb feldobott labda, mint pl. a már említett bagging eljárás vizsgálata, hogy nem a keresztellenőrzés sémája lenne-e a hatékonyabb.

7. Tézispontok

A tézispontoknál a váltakozó egyes szám első személy és többes szám első személy minden esetben arra utal, hogy szeparálhatónak érzem-e az általam elért új tudományos eredményt a társszerzőkétől.

1. Vizsgálatokat végeztem a fordított Monte Carlo módszerrel kapcsolatban. Megállapítottuk, hogy a különböző szerkezeti, termodinamikai és dinamikai tulajdonságok átlagát és fluktuációját nem reprodukálja megfelelően a módszer. Az RMC a kanonikus sokaságtól elérően mintavételezi a konfigurációs teret, valamint az általunk kifejlesztett szerkezeti alapú molekuláris dinamikai illesztés dinamikája sem felel meg a kanonikus sokaság dinamikájának. [s18] [s23] [s24] [s32] [s28] [s32]

2. Felfedeztem, hogy a párkorrelációs függvény fluktuációját önmaga a függvény határozza meg. Levezettük egy bizonyítását ennek, illetve egy új bizonyítását a Henderson-féle tételnek. A szerkezeti függvény fluktuációjának számolására közelítő képletet dolgoztam ki. [s15] [s18] [s25]

3. Módszereket dolgoztam ki belső szerkezettel rendelkező rendszerek fordított Monte Carlo modellezésére. Bemutattam, hogy monodiszperz kolloid rendszerek estén célszerű a részecskeátmérő előzetes optimalása, de reális eredmények csak polidiszperz közelítéssel kaphatóak, ami megoldható részecske átmérők párhuzamos illesztésével. Új módszert dolgoztam ki belső szerkezettel rendelkező nanométeres tartományú részecskék fordított Monte Carlo modellezésére, amit sikerrel alkalmaztunk hengeres nikkal katalizátor szemcsék és nemionos felületaktív anyag micelláinak modellezésére. [s16] [s22] [s25] [s30] [s33] [s34]

4. Négy új módszert dolgoztam/dolgoztunk ki a statisztikus mechanika szerkezet $\rightarrow$ párkölsönhatás inverz problémájának numerikus megoldására. a) Bemutattuk elsőként a Born-Green-Yvon hierarchia teljesülését numerikusan, majd kombináltan alkalmaztuk a fordított Monte Carlo módszerrel párpotenciál meghatározására. b) Továbbfejlesztettem az inverz Monte Carlo módszert, hogy alkalmas legyen szerkezeti függvénnyel, mint bemenő adattal dolgozni, illetve növeltem az iteratív módszer numerikus stabilitását. c) Mesterséges ideghálós közelítő megoldást dolgoztam ki szerkezeti függvényből kiinduló párpotenciál meghatározásra. d) Szimulációk konfigurációiból származó párkorrelációs függvényei és összenergiái ismeretén alapuló potenciál meghatározó módszert dolgoztam ki, amely hatékonyan használható coarse graining potenciál meghatározására. [s26] [s27] [s29] [s31] [s36] [s37]

5. Coarse graining potenciálok esetére tisztáztam, hogy párhuzamosan mind a szerkezet, az energetika és a dinamika reprodukciója szükséges egy új redukált modellben. [s24] [s32] [s37]

6. Alkáli ion/halogenid-higany és szisztematikus alkáli ion/halogenid-víz kölcsönhatásokat határoztam meg kvantumkémiai számolások segítségével határfelületi és tömbfázisbeli molekuláris dinamikai szimulációkhoz. [s10] [s14]

7. QM/QM szemiempirikus kvantumkémiai Monte Carlo szimulációs módszert fejlesztettünk ki és sikerrel alkalmaztuk amorf Si-re, amorf C-re és kloroszilánokra. [s7] [s19] [s5] [s6]

8. A Car-Parrinello szimulációs módszer segítségével vizsgáltam a sztratoszférikus ózon lebontásában résztvevő HCl, ClONO₂ és H₂O molekulák adszorpcióját salétromsav-monohidrát kristály felületén. A szimulációk során elsőként észleltem protonátmenetet a salétromsav-monohidrát tömbfázisában. [s17]

9. Karbamid kristálynövekedésére vonatkozó kinetikai Monte Carlo számolásaimmal bemutattam, hogy reális kristályosodási mikroszkopikus sebességi állandók kaphatóak pusztán a kristályok makroszkopikus morfológiájából kiinduló paraméteroptimalással. Ezzel alternatív útját mutattam be a megszokott modellezési irányoknak, amely a mikroszkopikus modellezésből indulva jut el a mezoszkopikus modellekig. [s39]

10. Bemutattuk, hogy kinetikai Monte Carlo szimulációk sebességi állandói megkaphatóak szabadenergia számoló módszerek szisztematikus alkalmazásával. A kutatásban metadinamika segítségével számoltuk ki argon kristályosodásának kinetikai Monte Carlo szimulációjához szükséges sebességi állandókat. [s46]

11. Definiáltuk a feltételes valószínűség szabályait alkalmazva a többminimumú rendszerek kommittor függvényeit és az ahhoz kapcsolódó további függvényeket, amelyek szükségesek a kinetikai Monte Carlo módszerre való áttéréshez. Informatív ábrázolási módokat vezettünk be a számolt függvényekre. [s54]

12. Matematikai módszereket dolgoztam ki adatok sor-oszlop rendezésére annak érdekében, hogy könnyebb legyen felismerni a két és háromdimenziós adattömbökben található mintázatokat. Olyan adattömbök kiértékelését céloztam meg, ahol az objektumok hasonlóságát csoportonként eltérő néhány változóban levő hasonlóságuk okozzák, míg a többi változó értéke nem jellemző az adott

csoportra. A módszerek alapja egy lokális távolság vagy hasonlóság mátrix/tömb, amely értékeire egy globális diagonális vagy folt jellegű célfüggvény szélsőértékei kereshetőek meg az eredeti adattömb sorainak illetve oszlopainak csatolt permutációjával. Számos példán bemutattuk a diagonális és a folt jellegű szeriálás hatékonyságát és gyakran egyedülként elfogadhatóan jó teljesítőképességét a bevezetőben vázolt adatstruktúrára. [s41] [s53] [s63]

13. Kifejlesztettem a Neumann-féle trend teszt és a Durbin-Watson teszt egytényezős varianciaanalízisre visszavezethető változatát. Továbbfejlesztettem a tesztek időeltolásokra és keresztkorrelációk vizsgálatára, valamint számos példán demonstráltuk a gyakorlati alkalmazhatóságát. [s42] [s47].

14. Szisztematikusan áttekintettem a szórás alapú skálázás és átlag alapú centrálás hat lehetséges esetét normál eloszlású adatokra, bemutatva a mintaszámtól függően kialakuló esetek néha meglepő eloszlásait. Az egyik alesetre elsőként határoztam meg a kialakuló eloszlás matematikai képletét. Bemutattuk, hogy a $c_4(N)$ korrekcióval számolt szórás alkalmazása megváltoztatja a t -eloszlás alakját, valamint kurtózisra vonatkozó tesztek esetében, pl. a Geary-féle tesztnél, a $c_4(N)$ korrekció kiterjeszti azok alkalmazhatóságát kis elemszámú mintákra. [s44] [s51]

15. A bemenő adatok allokációjának szisztematikus vizsgálatával megállapítottuk, hogy a modellezést megelőző kísérlettervezés különböző optimálisainak és az elkészült modellek validálásakor használt paramétereknek gyenge a keresztoptimalizációs lehetősége. Szisztematikus numerikus eredményekkel megerősítettük, hogy a validálási paraméterek alapján a szélső allokációknál kapunk átlagosan jobb modelleket. [58]

16. Bemutattuk, hogy periodikus változókat tartalmazó adatok vizsgálata során azt az adatablakot érdemes választani a periodikus változóra, amelyikben az adott változó varianciája minimális. [52]

17. Szisztematikus vizsgálatainkkal megerősítettük, hogy az illeszkedés jóságára vonatkozó validációs paraméterek a túlillesztés miatt kis mintaméreteknél túlzottan optimisták, nagy teljesítményű modellek esetén pedig nem relevánsak. A robusztusság és a prediktivitás metrikái megfelelnek az elvárásnak, hogy nagyobb minta jobb modellt eredményez. A validációs paraméterek rangkorrelációinak segítségével értelmeztünk számos jelenséget. Egyes validációs paramétereket és mintavételezési módszereket erős kritikával illettünk, használatukat legfeljebb feltételesen javasoljuk. Számolásaink összegzéseképpen megalkottunk egy átlátható validálási protokollt. [s59] [s61] [s65]

18. Felfedeztem, hogy az egy és a több elem kihagyásos keresztellenőrzési validációs paraméterek egymásba skálázhatóak, így azok közül mindig csak a modell típusától függően számítási szempontból kedvezőbbet/egyszerűbbet érdemes használni. [s59] [s61] [s65]

19. Bemutattuk, hogy a robusztusság vizsgálatára mind a keresztellenőrzés, mind az „out of bag” bootstrap eljárás alkalmas, de az előbbi kisebb varianciájú és könnyebben értelmezhető értékeket ad. [s64]

20. Kidolgoztam egy gyors eljárást annak a detektálására, hogy ún. nagy hatású pontokat tartalmaz-e a mintánk. Számos példán bemutattuk, hogy a teszt könnyen elvégezhető pusztán az R^2 és Q^2 validációs paraméterek ismeretében. [s48]

8. Irodalomjegyzék

(A saját munkák számozása a valós kutatási sorrendnek felel meg, nem a közlemény hivatalos megjelenésének. Ez a sorrendiséget mutató szám „s” jelet kapott, ha szerepel a dolgozatban és „e”-t, ha nincs részletezve.)

Az értekezésben szereplő saját munkák

s05 Náray-Szabó, Gábor; Tóth, Gergely; Ferenczy, György; Csonka, Gábor, The neglect of diatomic differential overlap (NDDO) fragment self-consistent field method for the treatment of very large covalent systems, Int. J. Quantum Chem. 52 (S21), 227-236. (1994) DOI:[10.1002/qua.560520719](https://doi.org/10.1002/qua.560520719)

s06 Tóth, Gergely; Gereben, Orsolya; Náray-Szabó, Gábor, Semiempirical fragment self-consistent field Monte Carlo simulations for liquid chlorosilanes, J. Mol. Struct. THEOCHEM 313, 165-172. (1994) DOI:[10.1016/0166-1280\(94\)85039-9](https://doi.org/10.1016/0166-1280(94)85039-9)

s07 Tóth, Gergely; Náray-Szabó, Gábor, Novel semiempirical method for quantum Monte Carlo simulation - application to amorphous silicon, J. Chem. Phys. 100, 3742-3746. (1994) DOI:[10.1063/1.466361](https://doi.org/10.1063/1.466361)

s10. Tóth, Gergely; Spohr, Eckhard; Heinzinger, Karl, SCF calculations of the interactions of alkali and halide ions with the mercury surface, Chem. Phys. 200, 347-355. (1995) DOI:[10.1016/0301-0104\(95\)00289-8](https://doi.org/10.1016/0301-0104(95)00289-8)

s14. Tóth, Gergely, Ab initio pair potential parameter set for the interaction of a rigid and a flexible water model and the complete series of the halides and alkali cations, J. Chem. Phys. 105, 5518-5524. (1996) DOI:[10.1063/1.472392](https://doi.org/10.1063/1.472392)

s15. Baranyai, András; Tóth, Gergely, Fluctuation of the pair-correlation function, J. Chem. Phys. 107, 8575-8576. (1997) DOI:[10.1063/1.475009](https://doi.org/10.1063/1.475009)

s16. Tóth, Gergely, The structure of dilute polymer colloid suspensions as seen by the reverse Monte Carlo method, Chem. Phys. Lett. 269, 413-418. (1997) DOI:[10.1016/S0009-2614\(97\)00304-7](https://doi.org/10.1016/S0009-2614(97)00304-7)

s17 Tóth, Gergely, Quantum chemical study of the different forms of nitric acid monohydrate, J. Phys. Chem. A 101, 8871-8876. (1997) DOI:[10.1021/jp971492d](https://doi.org/10.1021/jp971492d)

s18. Tóth, Gergely; Baranyai, András, Conceptual and technical improvement of the reverse Monte Carlo algorithm, J. Chem. Phys. 107, 7402-7408. (1997) DOI:[10.1063/1.474978](https://doi.org/10.1063/1.474978)

s19 Tóth, Gergely; Náray-Szabó, Gábor; Ferenczy, György; Csonka, Gábor, Monte Carlo simulation of amorphous systems with the fragment self-consistent field method, J. Mol. Struct. THEOCHEM 398-399, 129-133. (1997) DOI:[10.1016/S0166-1280\(96\)04969-X](https://doi.org/10.1016/S0166-1280(96)04969-X)

s21. Tóth, Gergely; Pusztai, László; Baranyai, András, Comment on “A new algorithm for reverse Monte Carlo simulations” [J. Chem. Phys. 109, 2624 (1998)], J. Chem. Phys. 111, 5620-5621. (1999) DOI:[10.1063/1.479826](https://doi.org/10.1063/1.479826)

- s22. Tóth, Gergely, Simultaneous Monte Carlo determination of particle size distribution and pair-correlation function of spherical colloids from a diffraction experiment, *Langmuir* 15, 6718-6723. (1999) DOI:[10.1021/la9901534](https://doi.org/10.1021/la9901534)
- s23. Tóth, Gergely; Baranyai, András, Comparison of reverse Monte Carlo algorithm variants via the accuracy of their three-particle correlation functions and configurational temperatures, *Mol. Phys.* 97, 339-346. (1999) DOI:[10.1080/002689799163730](https://doi.org/10.1080/002689799163730)
- s24 Tóth, Gergely; Baranyai, András, Structural models of diffraction experiments from the aspects of statistical mechanics, *Trends in Statistical Physics* 3, 165-174. (2000)
- s25. Tóth, Gergely, Monte Carlo determination of the radii and the pair-correlation function of spherical colloids, *Physica B Condens. Matter* 276, 404-405. (2000) DOI:[10.1016/S0921-4526\(99\)01599-9](https://doi.org/10.1016/S0921-4526(99)01599-9)
- s26. Tóth, Gergely; Baranyai, András, Direct determination of two-body potentials from measured pair structures, *J. Mol. Liq.* 85, 3-9. (2000) DOI:[10.1016/S0167-7322\(99\)00159-2](https://doi.org/10.1016/S0167-7322(99)00159-2)
- s27. Tóth, Gergely, Determination of pair-potential parameters from experimental structure factors, *J. Chem. Phys.* 115, 4770-4775. (2001) DOI:[10.1063/1.1391453](https://doi.org/10.1063/1.1391453)
- s28. Tóth, Gergely; Baranyai, András, Molecular dynamics analog of the reverse Monte Carlo method, *J. Chem. Phys.* 114, 2027-2035. (2001) DOI:[10.1063/1.1337044](https://doi.org/10.1063/1.1337044)
- s29. Tóth, Gergely, An iterative scheme to derive pair potentials from structure factors and its application to liquid mercury, *J. Chem. Phys.* 118, 3949-3955 (2003) DOI:[10.1063/1.1543142](https://doi.org/10.1063/1.1543142)
- s30. Tóth, Gergely; Körmendi, Krisztina; Vrabecz, Attila; Bóta, Attila, Evaluation of small-angle x-ray scattering data of a Raney-type Ni catalyst with computer simulation, *J. Chem. Phys.* 121 10634-10640 (2004) DOI:[10.1063/1.1812271](https://doi.org/10.1063/1.1812271)
- s31. Tóth, Gergely; Király, Norbert; Vrabecz, Attila, Pair potentials from diffraction data on liquids: A neural network solution, *J. Chem. Phys.* 123, 174109 (8 pages) (2005) DOI:[10.1063/1.2102887](https://doi.org/10.1063/1.2102887)
- s32. Tóth, Gergely; Baranyai, András, On the theoretical aspects of inverse methods: reverse Monte Carlo simulation or potential determination? *J. Condens. Matter Phys.* 17, S159-S166 (2005) DOI:[10.1088/0953-8984/17/5/016](https://doi.org/10.1088/0953-8984/17/5/016)
- s33. Tóth, Gergely; Madarász, Ádám, Structure of BRIJ-35 nonionic surfactant in water: A reverse Monte Carlo study, *Langmuir* 22, 590-597 (2006) DOI:[10.1021/la051380a](https://doi.org/10.1021/la051380a)
- s34. Tóth, Gergely, Reverse Monte Carlo analysis of small-angle scattering data on colloids and nanoparticles, *J. Mol. Liq.* 129, 108-114 (2006) DOI:[10.1016/j.molliq.2006.08.016](https://doi.org/10.1016/j.molliq.2006.08.016)
- s36. Tóth, Gergely, Effective potentials from complex simulations: a potential-matching algorithm and remarks on coarse-grained potentials, *J. Condens. Matter Phys.* 19, 335222 (12 pages) (2007) DOI:[10.1088/0953-8984/19/33/335222](https://doi.org/10.1088/0953-8984/19/33/335222)

- s37. Tóth, Gergely, Interactions from diffraction data: historical and comprehensive overview of simulation assisted methods, *J. Condens. Matter Phys.* 19, 335220 (18 pages) (2007)
DOI:[10.1088/0953-8984/19/33/335220](https://doi.org/10.1088/0953-8984/19/33/335220)
- s39. Tóth, Gergely, Microscopic kinetic data on crystal growth from macroscopic morphology, *Cryst. Growth Des.* 8, 3959-3964 (2008) DOI:[10.1021/cg800005g](https://doi.org/10.1021/cg800005g)
- s41. Tóth, Gergely; Szepesváry, Pál, A diagonal measure and a local distance matrix to display relations between objects and variables, *J. Chemom.* 24, 14-21 (2010) DOI:[10.1002/cem.1267](https://doi.org/10.1002/cem.1267)
- s42. Tóth, Gergely, The replacement of the Neumann trend test and the Durbin-Watson test on residuals by one-way ANOVA with resampling and an extension of the tests to different time lags, *J. Chemom.* 24, 140-148 (2010) DOI:[10.1002/cem.1293](https://doi.org/10.1002/cem.1293)
- s43. Tóth, Gyula Imre; Tegze, György; Pusztai, Tamás; Tóth, Gergely; Gránásky, László, Polymorphism, crystal nucleation and growth in the phase-field crystal model in 2D and 3D, *J. Condens. Matter Phys.* 22, 364101 (17 pages) (2010) DOI:[10.1088/0953-8984/22/36/364101](https://doi.org/10.1088/0953-8984/22/36/364101)
- s44. Tóth, Gergely, Destruction of normal distribution in small samples by centering and scaling, *J. Chemom.* 25, 247-253 (2011) DOI:[10.1002/cem.1382](https://doi.org/10.1002/cem.1382)
- s45. Tóth, Gyula Imre; Pusztai, Tamás; Tegze, György; Tóth, Gergely; Gránásky, László, Amorphous nucleation precursor in highly nonequilibrium fluids, *Phys. Rev. Lett.* 107, 175702 (4 pages) (2011)
DOI:[10.1103/PhysRevLett.107.175702](https://doi.org/10.1103/PhysRevLett.107.175702)
- s46. Kozma, Dániel; Tóth, Gergely, Microscopic rate constants of crystal growth from molecular dynamic simulations combined with metadynamics, *Advances Physical Chem.* 2012, 135172 (10 pages) (2012) DOI:[10.1155/2012/135172](https://doi.org/10.1155/2012/135172)
- s47. Tóth, Gergely; Balogh, Bertalan, How to identify cross correlations: a statistical test with time lag and its application on air-pollutant time series, *J. Chemom.* 26, 125-133 (2012)
DOI:[10.1002/cem.2414](https://doi.org/10.1002/cem.2414)
- s48. Tóth, Gergely; Bodai, Zsolt; Héberger, Károly, Estimation of influential points in any data set from coefficient of determination and its leave-one-out cross-validated counterpart, *J. Comput. Mol. Des.* 27, 837-844 (2013) DOI:[10.1007/s10822-013-9680-4](https://doi.org/10.1007/s10822-013-9680-4)
- s51. Tóth, Gergely; Szepesváry, Pál, Replacement of biased estimators with unbiased ones in the case of student's t-distribution and Geary's kurtosis, *Hung. J. Ind. Chem.* 45, 23-27 (2017)
DOI:[10.1515/hjic-2017-0005](https://doi.org/10.1515/hjic-2017-0005)
- s52. Tóth, Gergely; Király, Péter; Kiss, Dóra Judit, On the definition of mean, variance and covariance for periodic variables to avoid ambiguity in chemometric and bioinformatic data evaluation, *Chemom. Intell. Lab. Syst.* 168, 10-14 (2017) DOI:[10.1016/j.chemolab.2017.07.001](https://doi.org/10.1016/j.chemolab.2017.07.001)
- s53. Tóth, Gergely; Amariamir, Sasan, Seriation, the method out of a chemist's mind, *J. Chemometrics* 32, e2995 (14 pages) (2018) DOI:[10.1002/cem.2995](https://doi.org/10.1002/cem.2995)

- s54. Király, Péter; Kiss, Dóra Judit; Tóth, Gergely, Commitor of elementary reactions on multistate systems, J. Chem. Phys. 148, 134107 (12 pages) (2018) DOI:[10.1063/1.5007032](https://doi.org/10.1063/1.5007032)
- s58. Tóth, Gergely; Király, Péter; Kovács, Dániel, Effect of variable allocation on validation and optimality parameters and on cross-optimization perspectives, Chemom. Intell. Lab. Syst. 204, 104106 (10 pages) (2020) DOI:[10.1016/j.chemolab.2020.104106](https://doi.org/10.1016/j.chemolab.2020.104106)
- s59. Kovács, Dániel; Király, Péter; Tóth, Gergely Sample-size dependence of validation parameters in linear regression models and in QSAR, SAR QSAR Environ. Res. 32, 247-268 (2021) DOI:[10.1080/1062936X.2021.1890208](https://doi.org/10.1080/1062936X.2021.1890208)
- s61. Király, Péter; Kiss, Ramóna; Kovács, Dániel; Ballaj, Amine; Tóth, Gergely, The relevance of goodness-of-fit, robustness and prediction validation categories of OECD-QSAR principles with respect to sample size and model type, Mol. Inform. 41, 37-46 (2022) DOI:[10.1002/minf.202200072](https://doi.org/10.1002/minf.202200072)
- s63. Lasfar, Rita; Tóth, Gergely, Patch seriation to visualize data and model parameters, J. Cheminform. 15, 78 (17 pages) (2023) DOI:[10.1186/s13321-023-00757-1](https://doi.org/10.1186/s13321-023-00757-1)
- s64 Lasfar, Rita; Tóth, Gergely, The difference of model robustness assessment using cross-validation and bootstrap methods, J. Chemometrics 38, e3530 (13 pages) (2024) DOI:[10.1002/cem.3530](https://doi.org/10.1002/cem.3530)
- s65 Király, Péter; Tóth, Gergely, Being aware of data leakage and cross-validation scaling in chemometric model validation, J. Chemometrics 39, e70026 (12 pages) (2025) DOI:[10.1002/cem.70026](https://doi.org/10.1002/cem.70026)

További saját publikációk

- e01 Pusztai, L; Tóth, G, On the uniqueness of the Reverse Monte Carlo simulation: 1. Simple liquids, partial radial distribution functions, J. Chem. Phys. 94, 3042-3049 (1991) DOI:[10.1063/1.459827](https://doi.org/10.1063/1.459827)
- e02 Tóth, G; Pusztai, L, Determination of effective hard core diameters of liquids from their static vacancy distributions, Z. Phys. Chem. 178, 55-65 (1992) DOI:[10.1524/zpch.1992.178.Part_1.055](https://doi.org/10.1524/zpch.1992.178.Part_1.055)
- e03 Tóth, G; Pusztai, L, Determination of the radial distribution function of small particle polymer lattices using Reverse Monte Carlo Simulation, J. Phys. Chem. 96, 7150-7153 (1992) DOI:[10.1021/j100197a002](https://doi.org/10.1021/j100197a002)
- e04 Tóth, G; Pusztai, L, Comparative studies of the underlying local order corresponding to different radial distribution functions of disordered materials, Chem. Phys. 160, 405-413 (1992) DOI:[10.1016/0301-0104\(92\)80007-I](https://doi.org/10.1016/0301-0104(92)80007-I)
- e08 Baranyai, A; Tóth, G, Solvation dynamics from nonequilibrium molecular dynamics simulation, Mol. Phys. 14, 403-407 (1995) DOI:[10.1080/08927029508022033](https://doi.org/10.1080/08927029508022033)
- e09 Tóth, G; Heinzinger, K, Molecular dynamics study of an iodide and a lithium ion at the water-liquid mercury interface, Chem. Phys. Lett. 245, 48-53. (1995) DOI:[10.1016/0009-2614\(95\)00985-D](https://doi.org/10.1016/0009-2614(95)00985-D)

- e11 Bocker, J; Tóth, G; Heinzinger, K, Structure and dynamics of water and hydrated ions near a liquid mercury surface as studied by molecular dynamics simulations, Cond. Mater Phys. 8, 17-36 (1996)
- e12 Lu, T; Tóth, G; Heinzinger, K, Systematic study of the spectroscopic properties of isotopically substituted water by MD simulations, J.Phys. Chem. 100, 1336-1339 (1996) DOI:[10.1021/jp951546o](https://doi.org/10.1021/jp951546o)
- e13 Spohr, E; Tóth, G; Heinzinger, K, Structure and dynamics of water and hydrated ions near platinum and, mercury surfaces as studied by MD simulations, Electrochim. Acta 41, 2131-2144 (1996) DOI:[10.1016/0013-4686\(96\)00045-X](https://doi.org/10.1016/0013-4686(96)00045-X)
- e20 Tóth, G, Molecular dynamics study of alkali cations at the water/liquid mercury interface, J. Electroanal. Chem. 443, 73-80 (1998) DOI:[10.1016/S0022-0728\(97\)00471-3](https://doi.org/10.1016/S0022-0728(97)00471-3)
- e35. Vrabecz, A; Tóth, G, Simulation of binary hard-sphere systems with 1 : 5 and 1 : 10 size ratios, Mol. Phys. 104, 1843-1853 (2006) DOI:[10.1080/00268970600602313](https://doi.org/10.1080/00268970600602313)
- e38. Tóth, G; Vrabecz, A, Parameterization of Coulomb interaction in three-dimensional periodic systems, Mol. Simul. 33, 1033-1044 (2007) DOI:[10.1080/08927020701579345](https://doi.org/10.1080/08927020701579345)
- e40. Mile, V; Demeter, A; Tóth, G, Quantum chemical study of the ground-state alcoholic complexation of selected dual luminescent compounds, Mol. Phys. 107, 1987-1996. (2009) DOI:[10.1080/00268970903084953](https://doi.org/10.1080/00268970903084953)
- e49. Bakó, I; Pálinkás, G; Grósz, T; Bálint, S; Tóth, G; Radnai, T, A new approach to the determination of the uncertainty in neutron diffraction experiments with isotopic substitution method, J. Mol. Liq. 198, 5-10 (2014) DOI:[10.1016/j.molliq.2014.07.001](https://doi.org/10.1016/j.molliq.2014.07.001)
- e50 Tóth, G; Furtenbacher, T; Turányi, T, 2013 és 2015 között végzett kémia BSc-, vegyész MSc- és kémia PhD-hallgatóink továbbtanulási és elhelyezkedési szokásai, Magyar kémikusok lapja 71, 277-278 (2016)
- e55 Kiss, DJ; Oláh, J; Tóth, G; Menyhárd, DK; Ferenczy, GG, Quantum chemical calculations support pseudouridine synthase reaction through a glycal intermediate and provide details of the mechanism, Theor. Chem. Acc. 137, Paper: 162 (12 pages) (2018) DOI:[10.1007/s00214-018-2361-6](https://doi.org/10.1007/s00214-018-2361-6)
- e56. Tóth, G, Echemtest as an examination, control and tuning tool in chemistry education, Virt&I-comm 14 Paper: 5 (16 pages) (2018)
- e57. Manuali, C; Laganà, A; Tóth, G, Report on the prosumer implementation of EChemTest, Virt&I-comm 14, Paper: 7 (7 pages) (2018)
- e60 Kiss, DJ; Oláh, J; Tóth, G; Varga, M; Stirling, A; Menyhárd, D; Ferenczy, G, The structure-derived mechanism of box H/ACA pseudouridine synthase offers a plausible paradigm for programmable RNA editing ACS Catal. 12, 2756-2769 (2022) DOI:[10.1021/acscatal.1c04870](https://doi.org/10.1021/acscatal.1c04870)
- e62. Fatemeh, K; Garmarudi, B; Hadi, P; Tóth, G, Evaluation of FT-IR spectroscopy combined with SIMCA and PLS-DA for detection of adulterants in pistachio butter Infrared Phys. Tech. 127, Paper: 104369 (7 pages) (2022) DOI:[10.1016/j.infrared.2022.104369](https://doi.org/10.1016/j.infrared.2022.104369)

Külső irodalomjegyzék

- Allen, MP; Tildesley, DJ, Computer Simulation of Liquids, Oxford University Press, Oxford, (1987)
- Anscombe, FJ; Glynn, WJ, Biometrika 70, 227 (1983) <https://doi.org/10.1093/biomet/70.1.227>
- Arai, T; McGreevy, RL, J. Phys. Condens. Matter 10, 9221 (1998) <https://doi.org/10.1088/0953-8984/10/41/005>
- Arthur, DE; Uzairu, A; Mamza, P; Stephen, AE, Shallangva, GA, Chem. Data Coll. 7-8, 8 (2017) <https://doi.org/10.1016/j.cdc.2016.12.002>
- Baghcoli, T; Mousavi, M; Bababdani, BM, Chemom. Intell. Lab. Syst. 182, 31 (2018) [doi:10.1016/j.chemolab.2018.08.007](https://doi.org/10.1016/j.chemolab.2018.08.007)
- Baumann, K, Trends Anal. Chem. 22, 395 (2003) [https://doi.org/10.1016/S0165-9936\(03\)00607-1](https://doi.org/10.1016/S0165-9936(03)00607-1)
- Belkin, M; Hsu, D; Ma, S; Mandal, S, PNAS 116, 15849 (2019) [doi:10.1073/pnas.1903070116](https://doi.org/10.1073/pnas.1903070116)
- Bonett, DG; Seier, E, Comput. Stat. Data An. 40, 435 (2002) [DOI:10.1016/S0167-9473\(02\)00074-9](https://doi.org/10.1016/S0167-9473(02)00074-9)
- Bopp, P; Jancsó, G; Heinzinger, K, Chem. Phys. Lett. 98, 129 (1983) [https://doi.org/10.1016/0009-2614\(83\)87112-7](https://doi.org/10.1016/0009-2614(83)87112-7)
- Borbély, S, Langmuir 16, 5540 (2000) <https://doi.org/10.1021/la991265y>
- Bóta A; Goerigk, G; Drucker, T; Haubold, HG; Petró, J, J. Catal. 205, 354 (2002) <https://doi.org/10.1006/jcat.2001.3428>
- Box, GEP; Behnken, DW, Technometrics. 2, 455 (1960) [doi:10.1080/00401706.1960.10489912](https://doi.org/10.1080/00401706.1960.10489912)
- Böcker, J; Nazmutdinov, RR; Spohr, E; Heinzinger, K, Surf. Sci. 335, 372 (1995) [https://doi.org/10.1016/0039-6028\(95\)00408-4](https://doi.org/10.1016/0039-6028(95)00408-4)
- Bras, W; Xu, R; Wicks, J D; Van der Horst, F; Oversluizen, M; McGreevy, RL; van der Lugt, W, Nucl. Instrum. Methods Phys. Res. A 346, 394 (1994) [10.1016/0168-9002\(94\)90728-5](https://doi.org/10.1016/0168-9002(94)90728-5)
- Bro, R; Kjeldahl, K; Smilde, AK; Kiers, HAL, Anal. Bioanal. Chem. 390, 1241 (2008) <https://doi.org/10.1007/s00216-007-1790-1>
- Car, R; Parrinello, M, Phys. Rev. Lett. 55, 2471 (1985) [DOI:10.1103/PhysRevLett.55.2471](https://doi.org/10.1103/PhysRevLett.55.2471)
- Chaudhuri, BB, Pattern Recognit. Lett. 15, 893 (1994) [doi:10.1016/0167-8655\(94\)90151-1](https://doi.org/10.1016/0167-8655(94)90151-1)
- Chaudhuri, D; Murthy, CA; Chaudhuri, BB, IEEE Trans. Syst., Man, Cybern. Syst. 24, 1416 (1994) [doi:10.1109/21.310520](https://doi.org/10.1109/21.310520)
- Cheng, Y; Church, GM, Proceedings. International Conference on Intelligent Systems for Molecular Biology 8, 93 (2000)
- Christie, OHJ; Rácz, A; Elek, J; Héberger, K, J. Chemometr. 28, 287 (2014) [doi: 10.1002/cem.2601](https://doi.org/10.1002/cem.2601)
- Consonni, V; Ballabio, D; Todeschini, R, J. Chem. Inf. Model. 49, 1669 (2009) <https://doi.org/10.1021/ci900115y>
- Cook, DR; Weisberg, S, Residuals and influence regression, Chapman & Hall, New York, 1982.

- CPMD: Hutter, J; Ballone, P; Bernasconi, M; Focher, P; Fois, E; Godecker, S; Parrinello, M; Tuckerman, M, *CPMD*, Version 3.0, MPI für Festkörperforschung; Stuttgart and IBM Research Laboratory: Zürich, 1990-1996.
- Daru János, Theoretical Investigation of reaction mechanisms – from methodological development to applications, PhD értekezés, Hevesy György Doktori Iskola, ELTE, Budapest, (2015)
- Daszykowski, M; Walczak, B; Massart, DL, *Anal. Chim. Acta.* 468, 91 (2002) doi:10.1016/S0003-1670(02)00651-7
- Dellago, C; Bolhuis, PG, in *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, pp 167–233, edited by Holm, C and Kremer, K, Springer, Heidelberg, (2009)
- Dewar, MJS; Thiel, W, *JACS* 99, 4899 (1977) <https://doi.org/10.1021/ja00457a004>
- Dowdy, S; Wearden, S; Chilko, D, *Statistics for research* (3. ed.), Wiley, Hoboken (2004)
- Doweyko, AM, *IDrugs* 11, 894 (2008) <https://pubmed.ncbi.nlm.nih.gov/19051151>
- Doweyko, AM, *J. Comput. Aided. Mol. Des.* 22, 81 (2008) <https://doi.org/10.1007/s10822-007-9162-7>
- Duine, PA; Sietsma, J; Thijsse, BJ; Pusztai, L, *Phys. Rev. B Condens. Matter* 50, 13240 (1994) DOI: 10.1103/physrevb.50.13240
- Durbin, J; Watson, GS, *Biometrika* 37, 409 (1950)
- Durbin, J; Watson, GS, *Biometrika* 38, 159 (1951)
- Eriksson, L; Johansson, E; Kettaneh-Wold, N; Wikström, C; Wold, S, *Design of experiments, principles and applications*, Umetrics, Umea (2008)
- Esbensen, KH; Geladi, P, *J. Chemom.* 24, 168 (2010) <https://doi.org/10.1002/cem.1310>
- Evans, R, *Molec. Simul.* 4, 409 (1990)
- Ferenczy, GG; Rivail, JL; Surján, PR; Náray-Szabó, G, *Comput. Chem.* 13, 830 (1992)
- Filzmoser, P; Liebmam, B; Varmuza, K, *J. Chemomet.* 23, 160 (2009) <https://doi.org/10.1002/cem.1225>
- Frank, IE; Todeschini, R, *The data analysis handbook*, first ed., Elsevier, Amsterdam, (1994)
- Franz, JR, *Phys. Rev. Lett.* 57, 889 (1986)
- Frenkel, D; Smit, B, *Understanding Molecular Simulation*, Academic Press. (2001)
- Gaussian: M.J. Frisch et al., *Gaussian 92*, Gaussian Inc., Pittsburgh, PA, 1992
- Geary, RC, *Biometrika* 28, 295 (1936) DOI: 10.2307/2333953
- Geman, S; Bienenstock, E; Doursat, R; "Neural Networks and the Bias/Variance Dilemma," in *Neural Computation*, 4, 1-58 (1992) <https://doi.org/10.1162/neco.1992.4.1.1>
- Gereben, O; Pusztai, L; Baranyai, A, *J. Phys.: Condens. Matter* 6, 10939 (1994) <https://doi.org/10.1088/0953-8984/6/50/006>

- Gereben, O; Pusztai, L; Baranyai, A, Phys. Rev. B Condens. Matter 49, 13251 (1994) doi: 10.1103/physrevb.49.13251. PMID: 10010252.
- Gereben, O; Pusztai, L, Phys. Chem. Liq. 31, 159 (1996) <https://doi.org/10.1080/00319109608029569>
- German B, Glass Identification dataset, Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire RG7 4PN, (1987)
- Golbraikh, A; Tropsha, A; J. Mol. Graph. Model. 20, 269 (2002) [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
- Goodall, EA; Lee, A, Soft Matter, 17, 5393 (2021) <https://doi.org/10.1039/D1SM00402F>
- Graedel TE, Crutzen P J, Atmospheric Change: An Earth System Perspective, 1993, W.H. Freeman, San Francisco
- Gramatica, P, Mol. Inform. 33,311 (2014) <https://doi.org/10.1002/minf.201400030>
- Gramatica, P, QSAR and Comb. Sci. 26, 694 (2007) <https://doi.org/10.1002/qsar.200610151>
- Gramatica, P; Cassani, S; Roy, PP; Kovarich, S; Yap, CW; Papa, E, Mol. Inform. 31, 817 (2012) <https://doi.org/10.1002/minf.201200075>
- Gütlein, M; Helma, C; Karwath, A; Kramer, S, Mol. Inf. 32, 516 (2013) <https://doi.org/10.1002/minf.201200134>
- Hanke, M, Oberwolfach Reports 14, report No. 08/2017 doi: 10.4171/OWR/2017/8
- Hansen, JP; McDonald, IR, Theory of simple liquids, Academic, London (1986)
- Hastie, T; Tibshirani, R; Friedman, J, The Elements of Statistical Learning, 2nd ed., Springer, New York, (2009)
- Hawkins, DM; Basak, SC; Mills, D, J. Chem. Inf. Comput. Sci. 43, 579 (2003) <https://doi.org/10.1021/ci025626i>
- Henderson, RL, Phys. Lett. 49A, 197 (1974)
- Hill, TL, Statistical Mechanics, McGraw-Hill, New York, (1956)
- Hohmann, M; Monakhova, Y; Holzgrabe, U, J. Agric. Food. Chem. 63, 9666 (2015) <https://doi.org/10.1021/acs.jafc.5b03853>
- Hungarian Air Quality Network, <http://www.kvvm.hu/olm> , 2007-es adatok (Accessed at 2009-2011) later it has been transported to <http://legszenyezettség.met.hu/>
- Hungarian Air Quality Network, <http://www.levegominoseg.hu> (Accessed at June 2017) later it has been transported to <http://legszenyezettség.met.hu/>
- Irsik Nikoletta, Kalcitkristály néhány releváns makroszkopikus formája kinetikai állandók becsléséhez, Kémia BSc szakdolgozat, ELTE Kémiai Intézet, Budapest, (2010)
- IUPAC-IUB Commission on Biochemical Nomenclature, J. Biol. Chem. 245, 6489 (1970)
- Izvekov, S; Parrinello, M; Bunham, CJ; Voth, GA, J. Chem. Phys. 120, 10896 (2004) <https://doi.org/10.1063/1.1739396>

- Jain, S; Garge, S; Kumar, SK, Ind. Eng. Chem. Res. 45, 5614 (2006) <https://doi.org/10.1021/ie060042h>
- Jákli, I; Svend, J; Knak Jensen, SJ; Csizmadia, IG; Perczel, A, Chem. Phys. Lett. 547, 82 (2012)
<https://doi.org/10.1016/j.cplett.2012.08.002>
- Jarque, CM; Bera, AK, Econ. Lett. 6, 255 (1980) DOI:10.1016/0165-1765(80)90024-5
- Jorgensen, WL; Chandrasekhar, J; Madura, JD; Impey, RW; Klein, ML, J. Chem. Phys. 79, 926 (1983)
<https://doi.org/10.1063/1.445869>
- Juhász G, Reduction of biodiesel combustion reaction mechanism, BSc thesis, Eötvös Loránd University, Institute of Chemistry, Budapest, (2015)
- Kampa SM; Sammüller F; Schmidt M; Evans, R; Phys. Rev. Lett. 134, 107301 (2025)
<https://doi.org/10.1103/PhysRevLett.134.107301>
- Kawka, L; Juhász, G; Papp, M; Nagy. T; Zsély, IG; Turányi, T, Z. Phys. Chemie 243, 1329 (2020)
<https://doi.org/10.1515/zpch-2020-1649>
- Kennard, RW; Stone, LA, Technometrics 11, 137 (1969)
- Khazaei, A; Sarmasti, N; Seyf, JY, J. Mol. Struct. 1108, 168 (2016) doi:10.1016/j.molstruc.2015.11.072
- Kiss Dóra Judit, Párolgás vizsgálata számítógépes szimulációval, Kémia BSc szakdolgozat, ELTE Kémiai Intézet, Budapest, (2013)
- Kohut Gergely, Párolgás szabadenergiájának meghatározása súlyozott mintavételezéssel számítógépes szimulációban, Kémia BSc szakdolgozat, ELTE Kémiai Intézet, Budapest, (2014)
- Koop, T; Lao, B; Biermann, UM; Crutzen, PJ; Peter, T, J. Phys. Chem. A 101, 1117 (1997)
<https://doi.org/10.1021/jp9626531>
- Kraker, JJ; Hawkins, DM; Basak, SC; Natarajan, R; Mills, D, Chemom. Intell. Lab. Syst. 87,33 (2007)
<https://doi.org/10.1016/j.chemolab.2006.03.001>
- Laio, A; Parrinello, M, PNAS 99, 12562 (2002) <https://doi.org/10.1073/pnas.202427399>
- Lebold, KM; Noid, WG, J. Chem. Phys. 150, 234107 (2019) <https://doi.org/10.1063/1.5094330>
- Lebold, KM; Noid, WG, J. Chem. Phys. 151, 164113 (2019) <https://doi.org/10.1063/1.5125246>
- Liiv, I, Stat. Anal. Data Min. 3, 70 (2010) <https://doi.org/10.1002/sam.10071>
- Lin, ILK, Biometrics 45, 255 (1989) <https://doi.org/10.2307/2532051>
- Lyubartsev, AP; Laaksonen, A, Phys. Rev. E 52, 3730 (1995)
<https://doi.org/10.1103/PhysRevE.52.3730>
- Mac Nally, R; Duncan, RP; Thomson, JR; Yen, JDL, J. Appl. Ecol. 55, 1441 (2018)
<https://doi.org/10.1111/1365-2664.13060>
- Mahalanobis, PC, Proc. Nat. Inst. Sci. India, 2, 49 (1936)
- McGreevy, RL, J. Phys. Condens. Matter 13, R877 (2001) <https://doi.org/10.1088/0953-8984/13/46/201>
- McGreevy, RL; Pusztai, L, Mol. Simul. 1, 359 (1988) <https://doi.org/10.1080/08927028808080958>

- McKenzie, DR; Davis, CA; Cockayne, DJH; Muller, DA; Cassallo, AM, Nature 355, 622 (1992)
<https://doi.org/10.1038/355622a0>
- Meng, Y; Shukla, Y; Pande, VS; Roux, B, PNAS 113, 9193 (2016)
<https://doi.org/10.1073/pnas.1602790113>
- Metzner, P; Schütte, C; Vanden-Eijnden, E, J. Chem. Phys. 125, 084110 (2006)
<https://doi.org/10.1063/1.2335447>
- Metzner, P; Schütte, C; Vanden-Eijnden, E, Multiscale Model. Simul. 7, 1192 (2009)
<https://doi.org/10.1137/070699500>
- Moller, C; Plesset, M, Phys. Rev. 46, 618. (1934) <https://doi.org/10.1103/PhysRev.46.618>
- Molnar, C, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/ (2022)
- Náray-Szabó G; Surján, PR, Chem. Phys. Lett. 96, 499 (1983) [https://doi.org/10.1016/0009-2614\(83\)80739-8](https://doi.org/10.1016/0009-2614(83)80739-8)
- Nazmutdinov, RR; Probst, M; Heinzinger, K, J. Electroanal.Chem. 369, 227 (1994)
[https://doi.org/10.1016/0022-0728\(94\)87103-5](https://doi.org/10.1016/0022-0728(94)87103-5)
- Neumann, J; Kent, RH; Bellison, HB; Hart, BI, Ann. Math. Statist. 12, 153 (1941)
- Noid, WG; Szukalo, RJ; Kidder, KM; Lesniewski, MC, Annu. Rev. Phys. Chem. 75, 21 (2024)
<https://doi.org/10.1146/annurev-physchem-062123-010821>
- OECD “OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models” 2004, “Guidance document on the validation of (Quantitative) Structure-Activity relationships[(Q)SAR models” 2007 Organisation for Economic Cooperation and Development (OECD), can be found under <https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm>. Last accessed June 2022.
- Peters, B, Reaction Rate Theory and Rare Events, Elsevier, Amsterdam, (2017)
- Petrie, WM; JRAI 29, 295 (1899)
- Piana, S; Gale, JD, J. Am. Chem. Soc. 127, 1975 (2005) <https://doi.org/10.1021/ja043395l>
- Piana, S; Gale, JD, J. Cryst. Growth 294, 46 (2006) <https://doi.org/10.1016/j.jcrysgro.2006.05.020>
- Piana, S; Reyhani, M; Gale, JD, Nature 438, 70 (2005) <https://doi.org/10.1038/nature04173>
- Pusztai, L; McGreevy, RL, Phys. B: Condens. Matter 234–236, 357 (1997)
[https://doi.org/10.1016/S0921-4526\(96\)00986-6](https://doi.org/10.1016/S0921-4526(96)00986-6)
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Rácz, A; Bajusz, D; Héberger, K, Molecules 26, art.no. 1111 (2021)
<https://doi.org/10.3390/molecules26041111>

- Rácz, A; Bajusz, D; Héberger, K, SAR and QSAR in Env. Res. 26, 683 (2015)
<https://doi.org/10.1080/1062936X.2015.1084647>
- Rácz, A; Héberger, K; Rajkó, R; Elek, J, Herit. Sci. 1, 2 (2013) doi:10.1186/2050-7445-1-2.
- Raschka, S, arXiv:1811.12808 (2020)
- Rencher, AC; Pun, FC, Technometrics, 22, 49-53 (1980) <https://doi.org/10.2307/1268382>
- Rodionova, OY; Pomerantsev. AL, J. Chemom. 22, 674 (2008) doi:10.1002/cem.1103.
- Rosenberger, D; Hanke, M; van der Vegt, NFA, Eur. Phys. J. Special Topics 225, 1323 (2016)
<https://doi.org/10.1140/epjst/e2016-60120-1>
- Rousseeuw, PJ; Hubert, M, Recent developments in PROGRESS, in L1 -Statistical Procedures and Related Topics, (ed. Dodge Y) Institute of Mathematical Statistics Lecture Notes-Monograph Series, Volume 31, Hayward, California, 201–214, (1997)
- Rücker, C; Rücker, G; Meringer, M, J. Chem. Inf. Model., 47, 2345 (2007)
<https://doi.org/10.1021/ci700157b>
- Sachs, L, Angewandte Statistik, 7th ed., Springer, Berlin (1992)
- Salma, I, Daily air pollution and meteorological data Budapest, 2007. Mendeley Data, V1, (2023)
doi:10.17632/2mmwv3j4ms.1
- Schaeffer, R; Khona, M; Robertson, Z; Boopathy, A; Pistunova, K; Rocks, JW; Fiete, IR; Koyejo, O, arXiv:2303.14151 (2023)
- Schommers, W, Phys. Lett. A 43, 157 (1973).
- Shao, J; J. Am. Stat. Assoc. 88, 486 (1993)
- Spigler, S; Geiger, M; d'Ascoli, S; Sagun, L; Biroli, G; Wyart, M, J. Phys. A-Math. 52, 474001, (2019)
doi:10.1088/1751-8121/ab4c8b
- Sullivan, DM; Bagchi, K; Tuckerman, ME; Klein, ML, J. Phys. Chem. A 103, 8678 (1999) DOI: 10.1021/jp991067f
- Sullivan, DM; Bagchi, K; Tuckerman, ME; Klein, ML, J. Phys. Chem. A 103, 8678 (1999), DOI:10.1021/jp991067f
- Tetteh, J; Suzuki, T; Metcalfe, E; Howells, S, J. Chem. Inf. Comput. Sci. 39, 491 (1999)
<https://doi.org/10.1021/ci980026y>.
- Thampi, A, Interpretable AI: Building Explainable Machine Learning Systems, Manning, Shelter Island, (2022)
- Toher, D; Downey, G; Murphy, TB, Chemom. Intell. Lab. Syst. 89, 102 (2007)
<https://doi.org/10.1016/j.chemolab.2007.06.005>
- Topliss, JG; Edwards, RP, J. Med. Chem. 22, 1238 (1979) <https://doi.org/10.1021/jm00196a017>
- Vallet, F; Cailton, JG; Refregier, P, Europhysics Letters. 9, 315 (1989) doi:10.1209/0295-5075/9/4/003

Vanden-Eijnden, E; in Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology, 453–493, edited by M. Ferrario, G. Cicotti, and K. Binder, Springer, Berlin, Heidelberg, (2006)

Verdes, M, J. Mol. Graph. 86, 256 (2019) <https://doi.org/10.1016/j.jmgm.2018.10.025>

Wooten, F; Wearie, D, Solid State Phys. 40, 1, (1987) [https://doi.org/10.1016/S0081-1947\(08\)60689-X](https://doi.org/10.1016/S0081-1947(08)60689-X)

Zwicker J., Lovett R, J. Chem. Phys. 93, 6752 (1990) <https://doi.org/10.1063/1.458944>

9. Függelék

9.1. Rövidítések, angol szakkifejezések és a magyar szakzsargon feloldása/értelmezése

A (W 'szócikk') megjegyzés arra utal, hogy egy eredeti hivatkozásnál könnyebben érthető angol nyelvű Wikipédia lap érhető el az adott szócikkhez, sokszor szemléletes ábrákkal.

ANN (Artificial neural network, mesterséges idegháló) Mintázatfelismerési, osztályozási és regressziós célokra kifejlesztett olyan gépi tanulási módszerek összefoglaló neve, ahol a modellben többnyire rétegekbe rendezetten matematikai feldolgozó függvények találhatók az idegsejtekkel analóg be és kimeneti kapcsolatokkal. A legegyszerűbb klasszikus szerkezetű modell egy bemeneti réteget, egy rejtett réteget és egy kimeneti réteget használ és osztályozási vagy regressziós feladatokra tanítható be. A tanítás ebben az esetben a csomópontokat összekötő szakaszokhoz rendelt súlyok optimalizálását jelenti a tanító halmaz alapján. Ennél jelentősen összetettebb szerkezetű hálózatok is léteznek, újabban különösen sok helyen alkalmazzák a konvolúciós és a rekurrens hálókat. (W 'Neural network', 'Machine learning')

ANOVA (Analysis of variance, varianciaanalízis) Adatmátrix varianciájának, pontosabban eltérésnégyzetösszegeinek felbontásán alapuló módszerek összesítő neve, amelyek segítségével értelmezhető különböző hatások szerepe. Ún. egytényezős változata pl. statisztikai tesztként alkalmazható annak eldöntésére, hogy számít-e, hogy melyik megfigyelés melyik csoporthoz köthető [Dowdy et al. 2004] (W 'Analysis of variance').

bagging (Bootstrap aggregating, bootstrap alapú modellösszesítés) Gépi tanulás során sokaság alapú modell létrehozására javasolt módszer. Az eredeti mintából bootstrap eljárással készített számos új minta alapján betanított modellekből készül egy egyesített (sokaság alapú, néha konszenzusnak is nevezett) modell (W 'Bootstrap aggregating').

BGY (Born-Green-Yvon equation / egyenlet) Az egyenlet egzakt módon köti össze egy párkölcsönhatásokkal leírható rendszer párkölcsönhatási függvényét, párkorrelációs függvényét és a triplétt korrelációs függvényét. Részletek az 4.4. fejezetben és az [s26] publikációban találhatóak. [Hill 1956, Hansen&McDonald 1986].

bias-variance tradeoff (közelítő fordítás: torzítás-szórás kompromisszum) A modellezések predikciós képességének és komplexitásának összefüggésére kidolgozott elmélet. Egy sematikusan csökkenő és egy növekedő görbe által létrehozott minimum alapján egy optimális modellkomplexitást jelöl ki. A csökkenő görbe azt mutatja, hogy az egyre komplexebb modellek jobban képesek elméletileg torzításmentesen leírni a modellezendő jelenséget. A növekvő görbe azt mutatja, hogy adott mennyiségű rendelkezésre álló adat alapján egyre kevésbé lehetséges a komplexebb modellek paramétereit meghatározni, egyre nagyobb lesz az adatok

esetlegességéből adódó varianciája a modellnek. Az elmélet segítségével a modellezés hibája három részre bontható, a torzítási (bias), a variancia (hiányzó adatmennyiség) és a természetes ingadozásból vagy mérési hibából eredő tagra. [Geman et al. 1992] (W 'Bias-variance tradeoff')

BJH vízmodell, Philipp Bopp, Jancsó Gábor és Karl Heinzinger által kifejlesztett klasszikus mechanikai vízmodell. Flexibilis, de nem polarizálható modell. Főleg Karl Heinzinger és csoportja használta, határfelületi szimulációkban. [Bopp et al. 1983]

bootstrap (szó szerint csizmaszj) Statisztikai módszer, ahol a minta visszatevéses újramintavételezésével számos az eredeti mintával azonos elemszámú mintát hozunk létre és ezeket értékeljük ki vagy modellezük a további adatfeldolgozás során. Alapfeltevése, hogy az eredeti mintát tekinti a populációnak, és abból veszi a bootstrap mintákat (W 'Bootstrapping').

coarse graining (szó szerint durva szemcsézés, vagy durva szemcsés modellezés) Modellezésnél azokat a megoldásokat nevezik így, amikor a rendszert alkotó részecskék egy részét összevontan kezeljük. Alapvető célja, hogy lehetőséget adjon a rendszerméret növelésére vagy hosszabb modellezési idő megvalósítására, pl. szimulációkban. Néhány példa: kvantummechanikai leírás helyettesítése klasszikus mechanikai kölcsönhatásokkal az atomok/molekulák leírásánál. Oldószer elhanyagolása modellezéskor pl. kolloid rendszerek szimulációjakor. Molekularészletek vagy egész molekulák teljes kölcsönhatási helyként való kezelése. Coarse grained integrálásnak nevezik a statisztikus mechanikában, amikor a fázistér felett való integrálást nem folytonosan végezzük, hanem figyelembe vesszük a kvantummechanikából ismert legkisebb megkülönböztethető térfogatelemeket. (W 'Coarse grained modelling')

CPMD (Car-Parrinello molecular dynamics, Car-Parrinello molekuláris dinamika) Az első széleskörűen használt menetközben számolt kvantumkémiai erőteret használó molekuláris dinamikai módszer. A klasszikus dinamikai változók (atomi hely és sebesség koordináták) Lagrange formalizmusban felírt mozgásegyenletébe bekerültek a kvantumkémiai számolásokban szereplő bázisfüggvény koefficiensek és azok időbeli változásai is. A koefficiensek dinamikájához szükséges egy fiktív tömeg feltételezése. Amennyiben a valós és a fiktív dinamika adiabatikusan szétcsatolt, a szimulációs módszer sikeresen alkalmazható. A módszer gyakorlati implementációjában DFT alapú kvantumkémiai módszereket használnak többnyire, pszeudopotenciálokkal és síkhullámbázissal, periodikus rendszerekre. Helyes paraméterezés esetén a fiktív dinamika jól követi a Born-Oppenheimer potenciális energiafelületet. A 90-es évek végétől az itt használt metodikai fejlesztéseket hagyományos Born-Oppenheimer dinamikai programokba olyan sikerrel implementálták, hogy azok számítási igénye közel lecsökkent a Car-Parrinello dinamika szintjére [Car&Parrinello 1985]. (W 'Car-Parrinello molecular dynamics')

DFT (Density functional theory, sűrűségfukcionál-elmélet) A dolgozatban a kvantummechanikai sűrűségfukcionál-elméleten (Kohn Sham egyenleteken) alapuló kvantumkémiai módszerekre használok ezt a rövidítést. Általában elmondható, hogy a DFT módszerek alapesetben a Hartree-Fock módszernél nem sokkal nagyobb számítási igény mellett valamivel jobb eredményeket adnak. (W 'Density functional theory')

DOE (Design of experiment, kísérlettervezés) A matematikai statisztika eszközeit használó módszerek neve, ahol a cél a kísérletek olyan előre való eltervezése, hogy pl. a termék minősége, a selejt mennyisége, a költség, a kísérletek száma, a kockázatbecslés, a kísérleti eredmények, a létrehozandó modellek szempontjából optimális beállítások és paraméterek mellett végezzük el a kísérleteket. Egyes célokhoz néha speciális elnevezések is tartoznak. Modellezési cél esetén elsődleges, hogy megfelelően robusztus és jó predikciós képességű modelleket hozzunk létre, lehetőleg kis varianciájú modellparaméterekkel. Modellezést megelőző tervezéshez gyakran használunk ún. fix kísérleti terveket vagy az ún. kísérlettervezési optimálisok segítségével választjuk ki a lehetséges kísérleti paramétereket. [Eriksson et al. 2008] (W 'Design of experiment')

double descent (kettős lecsengés jelensége) A 2010-es évek végén felfedezett, a bias-variance tradeoff elmélettől eltérő viselkedése egy modell teljesítményének, ahol nagy modell komplexitásoknál a bias-variance tradeoff minimumánál jobb modellt kapunk. Többnyire sok rejtett réteget tartalmazó, ún. mélytanulási mesterséges ideghálónál észlelhető a jelenség, valamint a túlillesztés veszélye miatt jobban észlelhető, ha az optimáló algoritmus leállási kritériuma nagy. Bizonyos feltételek mellett lineáris regresszió is előfordul. A jelenség alapján ellentétes azzal a képpel, hogy az illesztendő paraméterek száma kisebb legyen, mint a felhasznált független esetek száma a tanító halmazban. A jelenség egyik lehetséges forrása, pl. hogy az optimáló algoritmusok miatt a paraméterek valójában nem teljesen függetlenek, az ott alkalmazott regularizáló és véletlen hatások miatt a modellparaméterek egyedi hatása korlátozott. [Vallet et al. 1989, Spigler et al. 2019, Belkin et al. 2019]. (W 'double descent')

FSCF (fragment self consistent field) Szemiempirikus kvantumkémiai módszeren belüli közelítés alapján a rendszer kettéosztására, ahol a külső részen szigorúan lokalizált közelítést alkalmaznak az elektronszerkezet számolásakor [Náray&Surján 1983].

$g(r)$ (pair correlation function, radial distribution function, párkorrelációs függvény) Kondenzált fázisok szerkezetének jellemzésére használt leggyakoribb dimenziómentes függvény, ahol két részecske távolságának függvényében az átlagos sűrűséghez képest vett relatív sűrűség értékét ábrázoljuk. Elméleti levezetése a fázistér sűrűségfüggvényéből lehetséges, ahol parciálisan kiintegráljuk az összes változót két helyváltozó kivételével, majd izotrópiát és homogenitást feltételezve áttérünk a maradék két változó különbségére. Ezzel ekvivalens az a szemléletes definíció, hogy

$g(r) = \langle \rho_{loc}(r) \rangle / \rho_0$, ahol r a központra tekintett részecskétől mért távolságot jelöli, ρ_0 az átlagos részecskeszámsűrűséget, $\langle \rho_{loc}(r) \rangle$ pedig az összes részecskére mint központra kiátlagolt távolság függő részecskeszámsűrűséget. Ideális gáz esetében a függvény minden r értéknél egységnyi értékű. Atomi-molekuláris szimulációkban a gyakorlatban diszkrét távolságintervallumokra számoljuk, kb. 0.01-0.2 Å felosztással. A sűrűséggel szorzott és térfogatelemre vett integrálja az adott távolságig levő átlagos szomszédszámot adja, az $n(r)$ -t. Több komponensű rendszereknél és molekuláris rendszereknél a rendszer leírásához párhuzamosan több párkorrelációs függvényt használunk, ezeket parciális párkorrelációs függvényeknek nevezzük. Az ún. teljes vagy totál párkorrelációs függvény a parciális párkorrelációs függvények lineáris kombinációja, ahol a koefficiensek értéke a móltörtektől és az adott párra vonatkozó diffrakciós állandóktól függ. [Hill 1956, Hansen&McDonald 1986] (W 'Radial distribution function')

KMC (Kinetic Monte Carlo simulation, kinetikai Monte Carlo szimuláció) Véletlenszámokat használó módszer rendszerek időbeli fejlődésének modellezésére. Az időfejlődés során a rendszer diszkrét állapotai között átmeneteket generálunk, az állapotok adott sorával modellezve a jelenséget. Előfeltétele, hogy ismerjük az adott átmenetekhez tartozó reakció sebességet vagy átmeneti valószínűséget. A kémiában többek között alkalmazható klasszikus reakciórendszerek modellezésére, kristálynövekedés modellezésére, fehérjemozgások állapot alapú vizsgálatára. Algoritmus szempontjából többfajta megközelítés létezik, ezek egy része kapcsolható a folyamatok valós időbeli evolúciójához. [4.6. fejezet, Peters 2017] (W 'Kinetic Monte Carlo')

kommittor (committor, kb.: elkötelezettség függvény) Reakciók vizsgálata során annak a valószínűségnek az értéke a reakciókoordináta függvényében, hogy a reakció az adott reakciókoordináta érték mellett a termék állapotba jut. Két minimumú rendszernél igaz, hogy egyből kivonva a kommittor értékét a kiindulási állapotba való jutás valószínűségét kapjuk. Jó reakció koordináta esetén a kommittor értéke az átmeneti állapotban 0,5. Elvileg a legjobb reakciókoordináta maga a kommittor. [4.6. fejezet, s54, Dellago&Bolhuis 2009, Peters 2017]

MC (Monte Carlo simulation, Monte Carlo szimuláció) Statisztikus mechanikai alapokon álló szimulációs módszer. Alapesetben a kanonikus sokaság konfigurációs terének a felderítésére alkalmazzák. Egy többnyire kocka alakú periodikus szimulációs cellában helyezik el az N darab kölcsönhatási pontot, amelyek a megfelelő, többnyire párkölcsönhatási potenciállal reprezentálják a célzott atomot vagy molekulát. A rendszer potenciális energiáját a pozíciók és a kölcsönhatási potenciál ismeretében számolják ki minden egyes konfigurációra, ahol a konfiguráció az rendszert alkotó N részecske egy adott elhelyezkedését jelenti. A szimuláció során véletlen elmozdításokkal újabb konfigurációkat generálnak. Egy adott konfiguráció elfogadásra kerül, ha potenciális energiája kisebb mint az azt megelőző konfigurációé, illetve ΔE

pozitív változás esetén is elfogadásra kerül $p = \exp(-\Delta E/k_b T)$ valószínűséggel. Elutasított próbálkozás esetén a korábbi konfigurációból történik egy újabb próbálkozás, valamint ez a korábbi próbálkozás még egyszer beszámításra kerül a sokaságból vett mintába. Általában egy véletlenül kijelölt részecske adott elemi kockán belüli véletlen elmozdítása jelenti az új konfiguráció generálását atomi rendszereknél. Egy kezdeti tranziens lépésszám után a konfigurációk reprezentatívan mintavételezik a kanonikus sokaságot és a Gibbs-féle sokaságtágnak megfelelően számolhatóak a konfigurációs tértől függő szerkezeti és termodinamikai tulajdonságai a rendszernek. A módszer dinamikai tulajdonságok számolására nem alkalmas. [Hansen&McDonald 1986, Allen&Tildesley 1987, Frenkel&Smit 2001] (W 'Monte Carlo method in statistical mechanics')

MCGR (Monte Carlo determination of $g(r)$, Monte Carlo meghatározása a $g(r)$ -nek) Monte Carlo alapú Fourier transzformációs eljárás, ahol a bemenő $S(q)$ függvényt irányítottan, de véletlen módosításokkal generált $g(r)$ függvényből lehet előállítani. A Fourier transzformáció a valós tér $\rightarrow$ inverz tér irányában történik olyan valós térbeli függvényből kiindulva, amelynek a valós térbeli hossza miatt elhanyagolható az ún. levágási hiba [Pusztai&McGreevy 1997].

MD (molecular dynamics, molekuláris dinamika) Klasszikus mechanikai szimulációs módszer, ahol a Monte Carlo szimulációhoz hasonlóan reprezentált N darab részecskét tartalmazó rendszert dinamikai módon követünk. A szimuláció során diszkrét időlépéseken keresztül numerikusan oldjuk meg a mozgásegyenleteket. Többnyire Lagrange vagy Hamilton formalizmust használunk a dinamika leírására. A Verlet-Störmer algoritmus valamilyen változatát, vagy prediktor-korrektor módszereket használunk az időbeli integrálás során. Atomi/molekuláris rendszereknél az időlépés kb. $0.1-2 \times 10^{-15}$ s, míg a teljes követett idő elérheti a 10^{-6} s-t. Hagyományosan pár száz vagy pár ezer részecske szimulációja is elegendő tömbfázisú rendszerek modellezésére, de összetett esetekben több százszazer kölcsönhatási pont is előfordul. A statisztikus mechanika Boltzmann-féle szemléletének módszere, ahol időbeli átlag segítségével számítjuk a kívánt mennyiségeket. Alapértelmezésben mikrokanonikus sokaságok modellezése felel meg a két mechanikai leírásnak, de pl. termosztátok és barosztátok hozzákapcsolásával más sokaságok is modellezhetőek. [Allen&Tildesley 1987, Frenkel&Smit 2001] (W 'molecular dynamics')

MLR (Multiple linear regression, többváltozós lineáris regresszió) Az egyváltozós lineáris regresszió kiterjesztése többváltozós adatokra. Többnyire a legkisebb négyzetek módszerét használjuk becslőfüggvényként. Ebben az esetben a $\mathbf{p}$ paramétervektor meghatározható a $\mathbf{p} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ egyenlet megoldásával, ahol $\mathbf{X}^{n \times m}$ a prediktor mátrix, $\mathbf{y}$ a válaszváltozó vektora, n a mérések száma. m többnyire a változók száma +1, mivel a prediktor mátrixban egy 1-es értékeket tartalmazó oszloppal a regressziós modellbe beépíthetjük a konstans tagot is. Erősen korrelált

változók esetén az invertálandó mátrix szinguláris vagy kvázisinguláris lesz, ilyenkor célszerű más regressziós módszert használni. (W 'Linear regression')

MNDO (modified neglect of diatomic overlap) Szemiempirikus kvantumkémiai módszer az elektronszerkezet számolására [Dewar&Thiel 1977](W 'MNDO')

MP2, MP4 A Møller–Plesset féle perturbációs elméleten alapuló ún. poszt Hartree-Fock módszerek, amelyek az elektronkorreláció különböző perturbációs közelítésekkel való figyelembevételével javítják a Hartree-Fock módszer megoldását. (W 'Møller–Plesset perturbation theory')

multiscale (multiscale modelling, több skálájú modellezés) Időben és/vagy térben olyan modellezési módszerek, ahol a vizsgált rendszer részekre van felosztva vagy a modellezés részletessége, vagy az időbeli fejlődés leírása alapján. A klasszikus térbeli felosztás során egy középső rendszert részletesebb leírással modellezünk, míg a környezetét elnagyoltabban. Akár többrétegű héjszerkezet is lehetséges a belső magas szintű kvantumkémiai modellezéstől a külső kontinuum jellegű szintig. Hasonló héjszerkezet lehetséges időbeli felosztásra is, de lehetséges az is, hogy a különböző sebességgel változó módusokat különböző időbeli felosztással léptetjük. Tágabb értelemben több skálás szimulációként hivatkoznak azokra az esetekre is, amikor egymás után vizsgálnak egy esetleg változtatott méretű rendszert, egyre részletesebb, vagy elnagyoltabb szimulációs módszerekkel. (W 'multiscale modelling')

on fly (szó szerint: menet közben) Szimulációk során a kölcsönhatási potenciál meghatározása menet közben. Tipikus példája amikor a Born-Oppenheimer közelítés mellett az atomokra/molekulákra ható erőt kvantummechanikai módon számoljuk minden időlépésben és nem előre meghatározott klasszikus mechanikai kölcsönhatásokat használunk.

PLS (Partial least squares regression, részleges legkisebb négyzete módszere) Többváltozós lineáris regressziós módszer korrelált változók esetére. A válaszváltozók számától függően egy változó esetén PLS-sel szokás rövidíteni a módszert, több válaszváltozó párhuzamos modellezése esetén inkább PLS2-vel. A módszer iteratív. Először megkeresi a magyarázó változóknak azt a lineáris kombinációját, amelyik a legjobban korrelál a válaszváltozóval, vagy több válaszváltozó esetén azoknak egy szintén optimált lineáris kombinációjával. A talált látens változó és a válaszváltozó, vagy azok látens változója között egyváltozós lineáris regresszióval meghatározza az együtthatókat, majd ezt a megmagyarázott részt kivonja mind a prediktor mátrixból, mind a válasz értékekből. Ezt követően a maradék mátrixokon folytatja az iterációt, maximum annyi alkalommal, ahány magyarázó változó volt az eredeti mátrixban. A látens változók optimális száma többféleképpen határozható meg, pl. keresztellenőrzéssel, validációs halmaz használatával, megmagyarázott összvariancia aránya alapján, vagy pl. a „repeated double cross-validation” módszerrel. (W 'Partial least squares regression')

PM (potential matching) Az egyik általam kifejlesztett megoldás a statisztikus mechanikai inverz problémára. A módszer elsődlegesen coarse graining potenciál levezetésére használható. Részletek a 4.4 fejezetben és az [s38]-as publikációban találhatók.

RMC (Reverse Monte Carlo simulation, fordított Monte Carlo szimuláció) A klasszikus Monte Carlo szimuláció analógjaként kidolgozott módszer diffrakciós mérési eredmények kiértékelésére, pontosabban azokkal konzisztens konfigurációk generálására. Az eljárás teljesen analóg a klasszikus Monte Carlo szimulációval annyi eltéréssel, hogy modellhez nem tartozik klasszikus mechanikai párpotenciál és annak alapján számolható potenciális energia. Ehelyett a konfigurációra számolt párkorrelációs függvény (vagy az abból Fourier transzformált szerkezeti függvény) és egy mért párkorrelációs függvény (vagy szerkezeti függvény) közti eltérésnégyzetösszeg alapján számolt χ^2 értéket használjuk.

$$\chi^2 = \frac{\sum_i^n \left(S^{\text{mért}}(q_i) - S^{\text{számolt}}(q_i) \right)^2}{\sigma^2}$$

Az egyenletben n mérési pontra számoltuk a szerkezeti függvények négyzetes eltérésösszegét. Az osztó σ elvileg a mérési hibával lenne összekapcsolható, de a gyakorlatban szimulációs paraméter, amit a megfelelő illesztéshez a valósnál kisebbre kell állítani. Az új konfiguráció elfogadása analóg a klasszikus Monte Carlo szimulációban alkalmazott eljárással, annyi különbséggel, hogy növekvő χ^2 esetén az elfogadás valószínűsége $\exp(-\Delta\chi^2/2)$ -vel arányos. További részletek a 3. fejezetben találhatók. [McGreevy&Pusztai 1988] (W 'Reverse Monte Carlo')

$S(q)$ (structure factor, szerkezeti függvény) A megfelelő normálások és korrekciók után diffrakciós mérésekből kapható inverz hossz dimenziós változójú függvény. Egykomponensű rendszer esetén Fourier transzformáltja a $g(r)$, többkomponensű rendszer esetén a totál szerkezeti függvény Fourier transzformáltja a teljes párkorrelációs függvény. Parciális változata a parciális $g(r)$ -ek transzformáltja. Inverz Fourier transzformációval a párkorrelációs függvényből is megkapható. [Hansen&McDonald 1986] (W 'Structure factor')

site-site (site-site interaction method, molekulák kölcsönhatási pont jellegű leírása) Modellezési módszer, ahol a molekuláris rendszerekben nem a teljes molekulára értelmezett szögváltozókat is tartalmazó részecskékre számoljuk az intermolekuláris kölcsönhatásokat, hanem a molekulán belül kijelölt kölcsönhatási pontok és a másik molekulán levő kölcsönhatási pontok között, többnyire csak azok távolságától függően. Ezzel a leírással sokszor elkerülhető a merev testekre vonatkozó mozgásegyenletek alkalmazása. Az intramolekuláris szerkezet lehet merev, ilyenkor numerikus módszerekkel tartjuk meg a kívánt szerkezetet, vagy leírható saját belső potenciállal is, megengedve a molekula flexibilitását. [Hansen&McDonald 1986, Allen&Tildesley 1987]

SVR (support vector regression, támogató vagy támasztó vektorok módszerével történő regresszió).

Eredetileg osztályozási feladatokra kifejlesztett módszer. Szemléletesen szokás úgy magyarázni a jó teljesítőképességét, hogy ellentétben a többi módszerrel nem az adott osztályok átlagos tulajdonságai alapján keresi meg az osztályok határait a változók terében, hanem a határok közelében levő, könnyen félreosztályozható pontok távolságát próbálja meg maximálni az elválasztó felülettől. Kernel módszerek közé tartozik, ezáltal alkalmas az eredeti térben nem csak lineáris elválasztófelületek létrehozására is. A támogató vektor regresszió egy továbbfejlesztése az osztályozásra alkalmas módszernek, azon belül is az úgynevezett szoft (puha) elválasztó felületeket használó módszereknek. Regressziónál egy ún. „slack” optimált változókkal kiegészített lineáris függvény meghatározása a cél. (W 'Support vector machine')

TIP4P vízmodell, Jorgensen csoportja által kifejlesztett 4 kölcsönhatási ponttal rendelkező merev, nem polarizálható klasszikus mechanikai vízmodell. A későbbiek során többször újraoptimalták a paramétereit. [pl. Jorgensen et al. 1983] (W 'water model')

$u(r)$ (pair potential, párpotenciál, párkölcsönhatási függvény) Két kölcsönható részecskének a távolságfüggő kölcsönhatási függvénye klasszikus mechanikai közelítésben. Gömbszimmetrikus részecskék esetén jól közelíthető a rendszer leírása többnyire csak a párkölcsönhatás szintjén, a több részecske hozzájárulások elhanyagolhatóak. Pontosabban, többnyire effektív párkölcsönhatási potenciálokat használnak, amelyek a több részecske kölcsönhatásokat átlagolt formában tartalmazzák. [Hill 1956, Hansen&McDonald 1986] (W 'Pair potential')

9.2 Validációs és optimalizációs paraméterek táblázata

jelölés	jelentés, definíció
n, n_{test}, n_{train}	minta elemszáma, elemszám a tanító és a teszt halmazban
$n_{fitted\ train}$	tanító minta elemszáma a keresztellenőrzéses ideiglenes modellekben
p	modell paramétereinek száma (tengelymetszet nélkül)
$y_i, \bar{y}$	i -dik válaszváltozó, átlaguk
$\hat{y}_i, \hat{y}$	modellezett i -dik válaszváltozó, átlaguk
$\hat{y}_{i/i}, \hat{y}_{i/j}$	kereszt ellenőrzéssel kapott i -dik válaszváltozó, egy elem kihagyásos és több elem kihagyásos esetre
RSS	reziduális négyzetösszeg $\sum_{i=1}^n (y_i - \hat{y}_i)^2$
MSS	modellezett négyzetösszeg $\sum_{i=1}^n (\hat{y}_i - \hat{y})^2$
TSS	teljes négyzetösszeg $\sum_{i=1}^n (y_i - \bar{y})^2$
$PRESS$	keresztellenőrzéssel kapott reziduális négyzetösszeg $\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2$
Illeszkedés jósága, belső validáció	
R^2	determinációs együttható $1 - \frac{RSS}{TSS}$
R_{adj}^2	szabadsági fokkal korrigált determinációs együttható $1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$
CCC	egyezési korrelációs együttható $\frac{2 \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - \hat{y})^2 + n(\bar{y} - \hat{y})^2}$
$RMSE$	átlagos négyzetes hiba gyöke (reziduálisok négyzetes közepe) $\sqrt{\frac{RSS}{n}}$
s	becsült szórás $\sqrt{\frac{RSS}{n-1}}$
MAE	átlagos abszolút eltérés $\frac{\sum_{i=1}^n y_i - \hat{y}_i }{n}$
F	lineáris regressziós analízis F-értéke $\frac{MSS/p}{RSS/(n-p-1)}$
Robusztusság, belső validáció	
$Q_{LOO}^2, Q_{LMO}^2 (R_{CV}^2, Q^2)$	egy és több elem kihagyásos keresztellenőrzés determinációs együtthatója $1 - \frac{PRESS}{TSS}$
CCC_{LOO}, CCC_{LMO}	CCC keresztellenőrzéssel számolva
$RMSE_{LOO}, RMSE_{LMO}$	$RMSE$ keresztellenőrzéssel számolva
$Q_{BS}^2, RMSE_{BS}$	bootstrap módszerrel számolt validációs paraméterek, "out of bag" értelmezésben

Prediktivitás, külső validálás	
Q_{F1}^2	$1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y}_{train})^2}$
Q_{F2}^2	$1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y}_{test})^2}$
Q_{F3}^2	$1 - \frac{\frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{n_{test}}}{\frac{\sum_{i=1}^{n_{train}} (y_i - \bar{y}_{TR})^2}{(n_{train} - 1)}} \quad (\text{korrigált normálással})$
CCC_{test}	CCC a teszt halmazra
$RMSE_{test}$	$RMSE$ a teszt halmazra
Kísérlettervezési rész, kísérlettervezési optimálisok	
X	prediktor mátrix (csak 1-eseket tartalmazó oszloppal kiegészítve a tengelymetszethez)
IQ	integrált hiba (2 változós esetre): $IQ = \iint (f_{modell} - f_{exact})^2 dx_1 dx_2$
D	$D = X^t X $ tervezéskor maximálva
A	$A = trace(X^t X)^{-1}$ tervezéskor minimalizálva
E	$E = \min_{\lambda} (X^t X)$ tervezéskor maximálva
T	$T = trace(X^t X)$ tervezéskor maximálva
G	$G = \max(\underline{x}_i^t (X^t X)^{-1} \underline{x}_i)$ tervezéskor minimalizálva
CN	$CN = \frac{\max.\text{szinguláris érték}(X)}{\min.\text{szinguláris érték}(X)}$ kondíció szám, tervezéskor minimalizálva

A táblázat képleteinek forrásai: áttekintő munkák [Frank&Todeschini 1994, Hastie 2009, Eriksson 2008], speciális változatok esetén [Consonni 2009, Filzmoser 2009, Lin 1989] és a saját publikációk [s48, s58, s59, s61, s64,s65]