

Az értekezés témaválasztása

Az értekezésben bemutatott kutatásaimban olyan módszereket fejlesztettem és alkalmaztam, amelyek különböző forrásból származó kémiával és különösen a fizikai kémiával kapcsolatos adatokból a klasszikus statisztikai módszereken túlmutató információk kinyerését tették lehetővé. Olyan rejtett információk keresésével próbálkoztam, amelyekhez néha speciális szimulációs és adatkiértékelési módszereket kellett kitalálnom, számítógépes kód formájában megvalósítanom majd alkalmaznom. Az értekezés látszólag eltérő három fő területét (vagy 10-15 részterületét) ez a kutatási módszer köti össze: mind számítógépes kutatást jelentenek, adatok felkutatását, beolvasását, generálását, esetleges szimulációk elvégzését, a hozzá kapcsolódó programozást, programok módosítását és alkalmazását, majd az eredmények értelmezését. Az MTA doktori értekezések leggyakrabban a jelölt kutatásainak tárgyára vonatkozóan egy-egy szűkebb területen való eredményeket mutatnak be. Műszeres vizsgálatokkal kapcsolatos értekezéseknél azonban előfordul, hogy kutatások tárgya meglehetősen diverz, a szakterületi specializálódás a kutatás eszközeiben és módszerében, pl. adott műszercsoport alkalmazásában vagy fejlesztésében jelenik meg. A számítógépes kémiát újabban szokás mint önálló kutatási eszköztárat definiálni. Emiatt gondoltam, hogy lehetséges egy olyan dolgozat elkészítése, ahol a bemutatott eredményeket a kutatás eszköze, jelen esetben a használt számítógépes szimulációs és egyéb adatkiértékelési módszerek kötik össze. A dolgozat gerincét azok a cikkek adják, amelyek belefértek a választott címbe, azokat a munkákat hagytam ki, amelyek témája az adatfeldolgozási címtől távoli volt. A tematikai szelekción felül kihagytam azokat a munkákat, ahol úgy lettem társszerző, hogy se a ráfordított idő, se a megvalósítással kapcsolatos ötleteknél nem volt jelentős a hozzájárulásom. A tézisek 42 cikkre vonatkoznak, a dolgozat 45 munkára.

A dolgozat kémiai bizottsághoz sorolása nem egyértelmű, annak ellenére, hogy a kutatások eszköze és módszere egységes volt. A szimulációkkal és az inverz problémával kapcsolatos részek hagyományosan a fizikai kémiai területhez, míg a kemometria témák az analitikai kémiához tartoznak. A két terület megértéséhez az értekezés függelékben rövid leírása található azoknak a módszereknek és elnevezéseknek, amelyek az egyik területen járatos kutatóknak a másik területről ismeretlen alapismeretnek tekinthetőek.

Adatok értelmezése szimulációkkal és a többváltozós statisztika eszközeivel a kémiában

Tóth Gergely
MTA doktori értekezés
téziszülete

Budapest
2025.

A kutatások főbb területei

Vizsgálatok a fordított Monte Carlo módszerrel kapcsolatban

A 3.1., 3.2 és részben a 4.3. fejezetek az egyetemi doktori (PhD) dolgozatom folytatásának tekinthetők. A kutatásokban a fordított Monte Carlo módszer (RMC) [McGreevy, RL; Pusztai, L, Mol. Simul. 1, 359 (1988)] egyértelműségével és teljesítményével foglalkoztunk Baranyai András társszerzőmmel. A munka során különböző statisztikus mechanikai függvényeken és azok fluktuációin keresztül vizsgáltuk az RMC módszert. Az első kérdés az volt, közelebb tudjuk-e vinni az RMC módszer fázistérbeli mintavételezését a kanonikus sokaságnak megfelelő mintavételezéshez? Bár kisebb eredményeket értünk el, de itt nem a kidolgozott módosításokat, vagy az RMC dinamikai változatának megalkotását gondolom fő eredménynek, hanem azt, hogy viszonylag egzakt módon sikerült definiálnunk a módszer korlátait, azok közül is a legfontosabbat, hogy az RMC nem képes a kanonikus sokaságnak megfelelően mintavételezni a konfigurációs teret. A fluktuációk elemzésekor elméleti jelentőségű eredményt értünk el. Bemutattuk, hogy kanonikus sokaságban a párkorrelációs $g(r)$ függvény meghatározza a saját fluktuációját:

$$\sigma_g^2(r) = \frac{\langle g(r) \rangle}{4\pi r^2 dr \rho}$$

, ahol ρ a rendszer sűrűségét jelöli.

A másik a fordított Monte Carlo szimulációhoz kapcsolódó téma a módszer átdolgozása volt kolloid mérettartományú rendszerek diffrakciós adatainak értelmezésére. Itt a fő kérdés a belső szerkezet paramétereinek a párhuzamos meghatározása volt, hiszen a mért diffrakciós függvények függenek ezektől a sokszor más forrásból nem megkapható adatoktól. Az általam végzett szimulációkban a klasszikus részecskepozíciók mellett párhuzamosan a részecskék belső szerkezetére vonatkozó paraméterek becslésére is módszereket dolgoztam ki, pl. alkalmassá tettem a módszert polidiszperz gömbszerű kolloidok, hengeres rendszerek és speciális szerkezetű felületaktív anyagokból álló micellák modellezésére.

Klasszikus szimulációk: kölcsönhatások meghatározása a priori és a szimuláció közben

A 4.1. és a 4.2. fejezetekben a kondenzált fázisok szimulációban való leírásának két módját mutatom be. A 4.1. fejezetben a hagyományos módszert, amikor a szimuláció előtt pl. kvantumkémiai számítások (QM) segítségével meghatározzuk a részecskék közötti kölcsönhatási függvényeket. A bemutatott példák elektrolit határfelület molekuláris dinamikai szimulációihoz számolt kölcsönhatási függvényeimre vonatkoznak: ion-higany felület és ion víz kölcsönhatásokra. A 4.2. fejezetben egy olyan fejlesztést és egy alkalmazást mutatok be, ahol a kölcsönhatási függvényt nem előre határozzuk meg, hanem a szimuláció közben számoljuk. Az első eset a sajátfejlesztésű szemiempirikus kvantumkémiai Monte Carlo szimuláció, ahol maga a rendszer két kvantummechanikai szinten kerül leírásra. Ez a módszer volt az első

QM/QM felosztást tartalmazó Monte Carlo szimulációs eljárás. A másik kutatás a Car-Parrinello molekuláris dinamika [Car, R; Parrinello, M, Phys. Rev. Lett. 55, 2471 (1985)] első alkalmazása a sztratoszférikus ózombomlás kinetikájának vizsgálatára, amely valószínűleg az első magyar alkalmazása is volt a módszernek.

Henderson tétele és az inverz probléma

Henderson 1974-ben közölte a tételét [Henderson, RL, Phys. Lett. 49A, 197 (1974)], amely szerint egy párkölcsönhatással $u(r)$ leírható klasszikus mechanikai rendszer esetén létezik egy elméleti funkcionál pár, $F(u(r))$ és $F^{-1}(g(r))$, amelyek kölcsönösen egyértelműen összekötik az $u(r)$ -t a $g(r)$ -rel:

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

A párkölcsönhatás $\rightarrow$ szerkezet irányban a funkcionál keresése helyett az elmúlt évtizedekben klasszikus és kvantummechanikai szimulációkat használunk. A fordított irányt nevezi az irodalom inverz problémának. Itt tisztán elméleti és szimulációval kombinált megoldások is léteznek, de egyelőre nincs olyan általános megoldás, mint amilyenek a fordított irányban a szimulációk.

A fordított Monte Carlo módszer hiányosságai miatt merült fel, hogy nem célszerűbb-e, ha a diffrakciós adatokból inkább a kölcsönhatást próbáljuk meghatározni, hogy azzal utána teljes értékű szimulációkat végezhessünk a kérdéses rendszerekre? Felmerült az is, hogy amennyiben egy rendszert különböző egyszerűsítésekkel és módszerekkel szimulálunk, mikor állíthatjuk, hogy a két szimuláció a vizsgált (rész)rendszerekre egyenértékű? Összességében négy oldalról közelítettem meg a kérdést, különböző bemenő adatok, modell egyszerűsítési célok és numerikus matematikai módszerek mentén. Mindegyik módszer az inverz probléma egy-egy új megoldására vezetett. A második kérdésre is választ kaptam, hogy a két szimuláció akkor tekinthető azonosnak a kérdéses (rész)rendszerekre, ha a szerkezeti, dinamikai és energetikai jellemzők mind azonosak.

A kinetikai Monte Carlo szimuláció néhány alkalmazása

A számítógépes kémiában az elvileg tökéletes modell egy a kölcsönhatásokra a kémiai átalakulások leírására is alkalmas kvantumkémiai módszerrel leírt olyan nagy rendszer, ahol a vizsgált jelenséget már nem befolyásolja annak a rendszer mérete, valamint a szimuláció időskálája is jelentősen nagyobb, mint a követendő folyamaté. A valóságban a három kritérium optimális kompromisszumát keressük minden feladatnál. Ennek az idő/tér/modell komplexitás hármasként adott feladatokra egy lehetséges optimális megoldása a kinetikai Monte Carlo szimuláció. Alapfeltétele, hogy a rendszerünk reális számú diszkrét állapotra bontható és ismert az állapotok közötti reakciók sebessége. Három kérdést vizsgáltam kinetikai Monte Carlo szimulációk segítségével: Meghatározhatóak-e a mikroszkopikus kristálynövekedés kinetikai állandói pusztán az elemi cella és egy ismert makroszkopikus

kristályforma alapján? Alkalmas-e a metadinamika [Laio, A; Parrinello, M, PNAS 99, 12562 (2002)] egy teljes mikroszkopikus kristályosodási állandó készlet meghatározására, amennyiben a hagyományos molekuláris dinamikai direkt út a ritka események kis száma miatt nem járható? Hogyan definiáljuk az ún. kommittor függvényt [Peters, B, Reaction Rate Theory and Rare Events, Elsevier, Amsterdam, (2017)] több minimummal rendelkező rendszerek szimulációjakor?

Szeriálás

Az 5. fejezetben a kemometriához kapcsolódó kutatásaimat mutatom be. Az 5.1. fejezet témája a szeriálás. A név a kémiai szakirodalomban alig ismert, pedig a tudomány minden ágában jelenlevő módszert jelent. Fő célja, hogy a szeriálást követően számos rejtett tulajdonsága/mintázata elérhetővé válik egy adattömbnek, mindemellett kisebb méret esetében a táblázatok számszerűen is jobban áttekinthetőek lesznek, míg pl. színekódos vizualizáció esetén értelmezhető csoportok-trendek válnak láthatóvá. Egy kétdimenziós kémiai adattömb (táblázat) hagyományosan a soraiban objektumokat, az oszlopaiban változókat tartalmaz. Ezek sorrendje valamilyen matematikai célfüggvény vagy algoritmikus előírás alapján átrendezhető, ezt nevezzük szeriálásnak [Liiv, I, Stat. Anal. Data Min. 3, 70 (2010)].

Az általunk kifejlesztett szeriálási módszerek egyediek mind a célzott adatszerkezetek, mind az objektum és a változó tengelyek összekapcsolása szempontjából. Az 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.

Trendvizsgálat és normáeloszlás kérdései

Az 5.2 fejezetben két kisebb témát tárgyalok. Az első rész a trendvizsgálatokkal foglalkozik, ahol a cél egy a korábbi módszereknél [pl. Neumann, J; Kent, RH; Bellison, HB; Hart, BI, Ann. Math. Statist. 12, 153 (1941)] ismertebb statisztikai alapokon álló új módszer kifejlesztése volt. Mindemellett sikeresen kiterjesztettem időfüggő és keresztkorrelációs esetekre is a régi és a saját új módszereket. A másik téma a normáeloszlású adatok két részkérdését tárgyalja, az egyik a felhasználók számára többnyire ismeretlen furcsaságait a centrálás és skálázás lehetőségeinek, a másik a normál eloszlású minták becsült szórásainak számolásakor használható korrekciós faktor hatását a t-eloszlásra és különböző kurtózis tesztekre [pl. Geary, RC, Biometrika 28, 295 (1936)].

Az allokáció és a modellvalidálás kérdései

Az 5.3. fejezetben szerepelnek a legújabb kemometriai eredményeim. 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. A fejezetben regressziós modelleket illetve modellezést vizsgáltunk több szempontból, tagolásunkban és a nomenklatúrában többnyire követve az OECD QSAR javaslatait [“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), <https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm>]. Az 5.3. fejezetben szereplő publikációk fő kérdései: Hogyan válasszuk meg az adatbázist periodikus változók esetén, ha nincs lehetőség a változók tóruszra helyezésének? Milyen kapcsolat található a modellezést megelőző kísérlettervezés stratégiai és a létrehozott modellek validált teljesítménye között? Igaz-e, amit a kísérlettervezésnél beígérnek? Kinek van igaza, a belső validáció, vagy a külső validáció ad az híveinek? Hogyan viszonyul ez a régi kemometrikusi vita a gépi tanulás új trendjeihez? Mit árul el a validációs paraméterek mintaméretfüggése és a korrelációik? Szükség van-e az összes irodalomban található validációs paraméterre? A robusztusságot keresztellenőrzéssel vagy bootstrap módszerrel vizsgáljuk? Van-e különbség az egy és a több elem kihagyásos keresztellenőrzési módszerek ítéletében? Mondhatunk-e valamit csak az ún. R^2 és Q^2 ismeretében az adathalmazunk alkalmasságáról? Mit jelent az adatszivárgás a kemometriában?

Tézispontok

A tézispontoknál a változó egy 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 elölően mintavételezi a konfigurációs teret, valamint az általunk kifejlesztett szerkezeti alapú molekuláris dinamikai illesztés dinamikája sem felel meg a kanonikus sokaság dinamikájának. [1] [2] [3] [4] [5]

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. [1] [3] [6]

3. Módszereket dolgoztam ki belső szerkezettel rendelkező rendszerek fordított Monte Carlo modellezésére. Bemutattam, hogy monodiszperz koloid rendszerek esetén célszerű a részecskeátmérő előzetes optimalálása, 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. [7] [8] [9] [10] [11] [12]

4. Négy új módszert dolgoztam/dolgoztunk ki a statisztikus mechanika szerkezet → párkölcsonhatás 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. [13] [14] [15] [16] [17] [18]

5. Coarse graining potenciálok esetére tisztáztam, hogy párhuzamosan mind a szerkezet, az energetika és a dinamika reprodukciója szükséges egy új redukált modellben. [3] [5] [17]

6. Alkáli ion/halogenid-higany és szisztematikus alkáli ion/halogenid-víz kölcsönhatásokat határoztam meg kvantumkémiail számolások segítségével határfelületi és tömbfázisbeli molekuláris dinamikai szimulációkhoz. [19] [20]

7. QM/QM szemiempirikus kvantumkémiail Monte Carlo szimulációs módszert fejlesztettünk ki és sikerrel alkalmaztuk amorf Si-re, amorf C-re és kloroszilánokra. [21] [22] [23] [24]

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. [25]

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 iránynak, amely a mikroszkopikus modellezésből indulva jut el a mezoszkopikus modellekig. [26]

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. [27]

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. [28]

12. Matematikail 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. [29] [30] [31]

13. Kifejtettem a Neumann-féle trend teszt és a Durbin-Watson teszt egytényezős varianciaanalízisre visszavezethető változatát. Továbbfejlesztettem a teszteket időeltolások esetre és keresztkorrelációk vizsgálatára, valamint számos példán demonstráltuk a gyakorlati alkalmazhatóságát. [32] [33].

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 $c_d(N)$ korrekció kiterjeszti azok alkalmazhatóságát kis elemszámú mintákra. [34] [35]

15. A bemenő adatok allokációjának szisztematikus vizsgálatával megállapítottuk, hogy a modellezést megelőző kísérlettervezés különböző optimálisainak és az elkészült modellek validálásakor használt paramétereknek gyenge a keresztoptimalizálási lehetősége. Szisztematikus numerikus eredményekkel megerősítettük, hogy a validálási paraméterek alapján a szélső allokációknál kapunk átlagosan jobb modelleket. [36]

16. Bemutattuk, hogy periodikus változókat tartalmazó adatok vizsgálata során azt az adatablakot érdemes választani a periodikus változóra, amelyekben az adott változó varianciája minimális. [37]

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 illetünk, használatukat legfeljebb feltételesen javasoljuk. Számolásaink összegzéseképpen megalkottunk egy átlátható validálási protokollt. [38] [39] [40]

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. [38] [39] [40]

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. [41]

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. [42]

A tézispontokat igazoló publikációk

[1] 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)

[2] 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)

[3] Tóth, Gergely; Baranyai, András, Structural models of diffraction experiments from the aspects of statistical mechanics, Trends in Statistical Physics 3, 165-174. (2000)

[4] 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)

[5] 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)

[6] 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)

[7] 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)

[8] 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)

[9] 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)

[10] Tóth, Gergely; Körmendi, Krisztina; Vrabec, 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)

[11] 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)

[12] 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)

[13] 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)

[14] 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)

[15] 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)

[16] Tóth, Gergely; Király, Norbert; Vrabec, 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)

- [17] 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)
- [18] 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)
- [19] 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)
- [20] 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)
- [21] 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)
- [22] 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)
- [23] 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)
- [24] 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)
- [25] 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)
- [26] 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)
- [27] 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)
- [28] 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)
- [29] 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)
- [30] 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)
- [31] 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)
- [32] 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)
- [33] 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)
- [34] 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)
- [35] 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)
- [36] 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)
- [37] 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)
- [38] 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)
- [39] 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)
- [40] 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)
- [41] 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)
- [42] Tóth, Gergely; Bodai, Zolt; 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)