

Gépi tanulás a számításos kvantumkémiaiában

BARCZA Gergely^{a,b,*}

^aWigner Fizikai Kutatóközpont, Konkoly-Thege Miklós út 29-33, 1121 Budapest, Magyarország

^bKomplex Rendszerek Fizikája Tanszék, ELTE, Pázmány Péter sétány 1/A, 1117 Budapest, Magyarország

1. Történeti bevezető

Alan Turing már a vákuumcsöves számítástechnika idején foglalkozott a mesterséges intelligencia megvalósíthatóságával. Az azóta eltelt hetven évben az informatika gyökereiben átalakult és átalakította a világot. Érzékeltetésekké: az 1945-ben félmillió dolláros ENIAC másodpercenként 5000 műveletre volt képes, míg egy mai asztali processzor ennek százmilliószorosára. A technológiával együtt az alkalmazások, illetve a numerikus módszerek is ugrásszerűen fejlődtek. Nem kivételek ez alól a mesterséges intelligenciához kapcsolódó kutatások sem. Kezdetben, szimbolikus megközelítést alkalmazva szabálygyűjtemények implementálásával egyre hatékonyabb és rugalmasabb modellek építésére törekedtek. Ezzel párhuzamosan már évtizedek óta zajlik a „utasításmentes” metodika fejlesztése is, mely nagy mennyiségű adat elemzéséből építi fel modelljét. Az áttöréshez szükséges számítási kapacitás és elméleti finomhangolás a kilencvenes években érte el a kritikus szintet és napjainkban robbant be a köztudatba személyi asszisztensek és önvezető autók formájában.

Megfigyelhető, hogy az utóbbi időben a számításos kvantumkémiai kutatói közösség is egyre bátrabban támaszkodik a gépi tanulás alapú technikákra és egyre több új módszer, illetve alkalmazás születik a határterületen [1, 2]. E gondolatébresztőnek szánt cikkben vázoljuk a legfontosabb tanítási módszerek alapgondolatát, illetve rövid ízelítőt adunk a legfrissebb és legérdekesebb kémiai vonatkozású eredményekből.

2. Gépi tanulási paradigmák

A mesterséges intelligencia azon kutatási ágát, melyben a modelt mint adatok felhasználásával, szabályok explicit programozása nélkül állítja elő gépi tanulásként (ML) nevezik. A rendszer önállóan, a tanító adatban rejlő minták alapján próbál általánosítani, szabályszerűségeket keresni. Sikeres tanítás esetén a modell nem „bemagolja” a választ, hanem általánosan képes helyes megoldást adni a problémára. A tanításhoz felhasznált adatok jellegük szerint két típusra bonthatók: 1) tipikus bemeneti minta adatok és azok modellezendő, helyes „megoldása” (címké) is az adathalmaz része 2) címkézetlen bemeneti paraméterek halmaza. Fontos megjegyezés, hogy a címke előállítását gépi vagy emberi erőforrást igényelhet. A tanítási adatok típusa és a tanítási feladat jellege szerint három fő kategóriába sorol-

hatók az ML módszerek, melyeket számtalan különböző numerikus eljárás megvalósít, köztük statisztikából, illetve fizikából ismert alapvetések.

Felügyelt tanítás előre címkézett adatok segítségével történik, azaz nemcsak a bemeneti minta adat, hanem a hozzá tartozó címke is a számítógép rendelkezésére áll. A tanítás célja a bemenő objektum lehető legpontosabb leképezése a kimeneti információra. Ilyen jellegű problémák egyik leg-egyszerűbb megoldási módszere a lineáris regresszió.

A felügyelet nélküli tanítás során a tanító halmaz nem tartalmaz címkéket, a módszer célja éppen a rendszer belső struktúrájának feltérképezése bármilyen preconcepció nélkül. A probléma felfogható mint dimenzióredukációs feladat. Tipikusan használt megoldási módszerek közé tartozik a főkomponens-elemzés, illetve a szinguláris érték felbontás.

Nem közvetlenül a címkementes adatokból, hanem a tapasztalatok alapján zajlik a megerősítési tanulás, melynek feladata egy célfüggvény maximalizálása. Ennek érdekében a program dinamikusan, „környezetéből” nyert visszacsatolás révén hangolja modelljét. Többek között a sztochasztikus Monte-Carlo módszerek is ebbe a kategóriába sorolhatók.

Az említett módszerek természetesen nem tekinthetők az ML teljes arzenáljának. Az elmúlt évtizedekben számtalan új, eltérő alkalmazási potenciállal rendelkező technikát vezettek be. Napjainkban a gépi tanulás svájci bicskájának a területet forradalmasító mesterséges neurális háló (ANN) tekinthető, mely a biológiai neuronok alapvető viselkedését próbálja a lehető legegyszerűbb matematikai eszközökkel mimálni. Gyakorlatban egy sor különböző, adott típusú problémákra optimalizált mesterséges neurális modellt alkotnak. Ezt illusztrálja az a tény is, hogy mind a lineáris regresszió, mind szinguláris érték felbontás, mind a variációs Monte-Carlo módszer tekinthető egy-egy mesterséges neurális háló egyszerűsített határesetének (azaz sorrendben: zérus rejtett rétegű neurális hálónak, lineáris autoencodernek, megszorított Boltzmann-gépnek).

3. Mesterséges neurális háló

Az alábbiakban vázaltszerűen összefoglaljuk a mesterséges neurális hálók alapjait. Számos különböző típus van, az 1. ábrán a felügyelt tanulás során jellemző, úgynevezett előre-

* Tel.: +36 1 392 2222; e-mail: barcza.gergely@wigner.hu

csatolt ANN felépítése látható mely a bemeneti adatokból (x vektor jelöli az ábrán) azonos struktúrájú matematikai műveletek sorozata segítségével határozza meg a kimeneti értékeket ($y=y(x)$ vektor). A két adat réteg közt lévő egy oszloponyi műveleti struktúrát rejtett rétegnek nevezzük. A szemléltető ábrával szemben jellemzően nem csak egyetlen rejtett réteggel definiálják a hálót, mely esetben mély neurális hálóról beszélhetünk. Fontos jellemző, hogy közvetlen műveleti összeköttetés csak a szomszédos rétegek között valósul meg. A rétegek egy-egy elemét neuronnak hívják, mely egy nemlineáris függvény és egy lineáris művelet kompozícióját hajtja végre a bemenő, előző $l-1$. rétegből származó $a_{i,l-1}^l$ adatokon. Azaz az l rétegbeli j neuron művelete az alábbi alakban írható

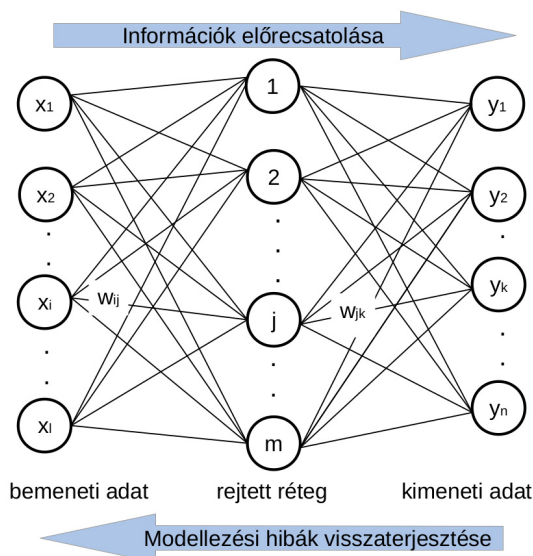
$$a_j^l = f(\sum_i w_{ij}^l a_i^{l-1} + b_j^l). \quad (1)$$

Az $l=1$ réteg bemeneti adata az x vektor, $l=L$ réteg kimeneti a^L vektor pedig maga az ANN $y(x)$ -re adott becslése. Az aktivációs függvény jellemzően használt alakjai: $f(z)=\tanh(z)$, $f(z)=\max\{0,z\}$, illetve a sigmoid függvény, $f(z)=1/(1+e^{-z})$. Az alkalmazott aktivációs függvény non-linearitása révén tetszőleges folytonos leképezés reprezentálható megfelelően mély háló alkalmazása esetén.

A tanítás feladata az optimális w_{ij}^l és b_i^l paraméterek meghatározása. A keresett paramétereket iteratív módon variáljuk az optimum eléréséig. Minden iterációt két lépésre bonthatunk:

1) az 1. képletnek megfelelően meghatározzuk az aktuális paraméter készletre vonatkozó $a^L(x)$ értékeket az n elemű tanuló halmazon.

2) Egy költségfüggvény, pl.: $C=1/(2n)\sum_x ||a^L(x)-y(x)||$, modell paraméterek szerinti numerikus deriváltja segítségével adunk pontosabb becslést a w_{ij}^l és b_i^l értékekre (gradient descent).



1. ábra. Egyetlen rejtett rétegű mesterséges neurális háló.

A súlyok optimalizálásához a rendelkezésre álló adatok jelentős részét (tréning adatok) felhasználják. Az iteráció során a tanuló adatok költsége monotonan csökkenve a betanított modell eljuthat egy olyan szintre, ahol már a tanító adatok zajára is érzékeny. Ezt a túltanulást elkerülendő a tréningre nem használt tesztelési adatok hibája révén kontrollálhatjuk az illesztett ANN modell tényleges pontosságát.

Összefoglalva, az ANN kiértékelése (az adatok előrecsatolása) során a modell paraméterek rögzített értékűek és a bemeneti tanító tanított sorozatára határozzuk meg az illesztett értékeket. Ezzel szemben, az algoritmus 2. lépésében a kimeneti-bemeneti adatok rögzítettek és a súlyok értékei változnak a hibák visszaterjesztése révén.

Természetesen az ANN alapú modellek nem tekinthetők univerzálisan alkalmazható megoldási módszernek. Felügyelt tanítási feladat esetén jellemző kihívás a megfelelő minőségű tanításhoz szükséges adatmennyiség előállítása. Továbbá fontos kiemelni, hogy a módszer, jellegéből adódóan, alapvetően interpoláció jellegű feladatok végrehajtására képes. Ezekből következik, hogy komoly kihívás a betanított neurális hálót szemléletes matematikai modellként interpretálni, továbbá kevésbé megbízható eredményt produkál extrapoláció jellegű, általánosítást igénylő kérdésekben.

4. További felügyelt tanítási módszerek

A felügyelt gépi tanítási modelleknél két alapvető típus (osztályozás, illetve regresszió) különböztethető meg, aszerint, hogy a kimeneti paraméterek diszkrét vagy folytonosak (pl: kovalens kötéstípusok szerinti klasszifikáció, illetve potenciál felület vizsgálata).

Az úgynevezett k -legközelebbi szomszéd módszer modellalkotás nélkül ad jóslatot, feltételezve, hogy a hasonló bemeneti minták azonos osztályba tartoznak. Az objektumok hasonlósága megfelelő metrikák segítségével számszerűsíthető és a kérdéses elemre a k szomszédos tréning adat alapján kaphatunk megoldást. A módszer klasszifikációra és regresszióra is alkalmazható: előbbi esetben a k szomszédos objektum tipikus osztálya, míg regresszió esetén a kimeneti értékük átlaga alapján kapunk becslést.

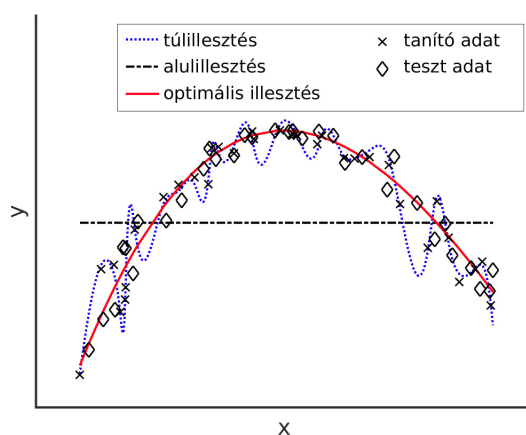
A Bayes-tétel alapján számos klasszifikációs módszert fejlesztettek, melyek segítségével a tanítóhalmazból meghatározható a legvalószínűbb modell. Ezen Bayes-hálózatok a legnagyobb valószínűség módszerét alkalmazzák a modell paraméterek becslésére a feltételes valószínűségekre vonatkozó Bayes-tétel kihasználásával. A tulajdonságok feltételes függetlensége esetén az úgynevezett naiv Bayes-hálózatok révén nagyon hatékonyan kezelhető a probléma numerikusan: az illesztendő paraméterek száma a jellemző tulajdonságok számával egyenes arányban nő.

A döntési fa az összetett osztályozási problémákat egyszerű eldöntendő kérdések sorozatára vezeti vissza. Egy vizsgálandó elem klasszifikálása során a döntési fa gyökeréből kiindulva a csomópontokban feltett kérdések válaszai alapján szállunk le a levelekig a fában, mely meghatározza az objektum címkéjét. A tanító adathalmazból rekurzívan állítható elő a döntési fa. Minden lépésben egy új osztályozást vezetünk be, mellyel az adatok több diszjunkt csoportra bonthatók. Az önkényesen bevezetett osztályozás akkor tekinthető hasznosnak, ha a generált csoportok hasonló méretűek és a vizsgált tulajdonság szórása az egyes csoportokban kisebb mint a teljes adatmintán. A döntési fákat gyakran használják együttesen (meta-algoritmus) a modellezés hatékonyságának növelésére. A döntési fa előnye, hogy a lényegtelen paramétereket automatikusan szűri. Ennek következménye, hogy kevésbé érzékenyek az adatok zajosságára, továbbá nagyméretű adathalmazokra is hatékonyan alkalmazhatók. A módszer további előnye a kapott modell interpretálhatósága, amit orvosdiagnosztikai és gazdasági feladatokban használnak ki.

A különböző osztályozási problémák sokszor nem oldhatók meg könnyen az elemek terében hipersíkok segítségével. A kernel módszerek (mint például a szupport vektor gép és a ridge regresszió) segítségével az adatokat egy magasabb dimenziós, úgynevezett jellemzőtérbe vetítve jelentősen javítható a szeparálhatóság.

5. Gépi tanulás alapú modellezés a gyakorlatban

Egy-egy újonnan felvetett alkalmazási probléma ML alapú vizsgálatát jellemzően kereszt-validációs stratégiát alkalmazva kezdik, azaz különböző tanítási módszereket tesztelve a gyakorlati tapasztalatok alapján választják ki a legjobban teljesítő tanító eljárást. További lehetőség, hogy eltérő algoritmusok, illetve különböző paraméterezésű modellek összességéből alkotnak robusztus leírást.



2. ábra. Az elfogultság-variancia kompromisszum illusztrációja. Nagy elfogultság esetén a releváns kapcsolatok nem modellezhetők: ilyen alulillesztett modell mind a tanító és teszt adatokra nagy hibát ad. Nagy variancia esetén a tanító adatok fluktuációjára érzékeny modell a teszt adatokra nagy hibával ad jóslatot (míg a tanító adatokra minimális a hiba). Ezen esetekkel szemben optimális modell mind a tanító és teszt adatokra elfogadható hibát ad.

Mivel az algoritmusok teljesítményének finomhangolása nagyon időigényes folyamat, így sokszor elsősorban a tanító adatok mennyiségének és minőségének növelésével próbálják a tanítás hatékonyságát fokozni. A tanítási modell megválasztásánál fontos szempont, hogy megfelelő komplexitású legyen, amit elfogultság-variancia kompromisszumnak (bias-variance tradeoff) neveznek a szakirodalomban. A kiegyensúlyozottság kényes problémája jól illusztrálható a 2. ábrán látható adatok illesztésével: túl sok illesztendő paraméter esetén a nyers adatokból származó zajosság óhatatlanul megjelenik a modellben, míg túl kevés paraméter nem képes visszaadni a rendszer részleteit.

A modellezés hatékonyságát sokszor javíthatja különböző tanítási módszerek kombinált alkalmazása is, mely a nyers adatot a felügyelet nélküli és a felügyelt tanulás elméletét is felhasználva több lépésben dolgozza fel.

6. Kvantumkémiai alkalmazások

A számítástechnikai kvantumkémia atomi rendszerek fizikai illetve kémiai tulajdonságait vizsgálja a kvantummechanika elveire építve. Ezen töltésrendszerek kölcsönható Schrödinger-egyenletük megoldásaként adódó hullámfüggvény révén írhatók le. Az egzakt megoldás komplexitása exponenciálisan nő az elektronok számával, így a bevezetően említett drámai számítástechnikai fejlődés ellenére a jelenlegi digitális számítógépekkel legfeljebb tucatnyi elektron kezelhető egzaktul. Jelentősen nagyobb rendszerek jellemzésére az elmúlt évtizedekben számtalan közelítő módszert fejlesztettek kompromisszumot kötvé a numerikus komplexitás és a precizitás között.

A számítástechnikai kémiában a következő években a Hohenberg–Kohn-tételekhez fogható forradalmi változásokat a robbanásszerűen fejlődő kvantuminformatika [3] mellett az ML paradigmája [2] hozhat. A következőkben pár példával illusztráljuk a technológia tipikus kémiai alkalmazásait.

6.1. Hullámfüggvény meghatározás

A hagyományos korrelációs módszerek (pl.: perturbatív módszerek, konfigurációs kölcsönhatás kifejtés, csatolt klaszter eljárás) az elektronrendszer hullámfüggvényét expliciten kifejtik a módszerek által megengedett, trunkált állapotterén. Ezekkel szemben a statisztikus, úgynevezett Monte-Carlo módszerek a konfigurációk halmazát megfelelően mintavételezve a vizsgált kvantumrendszer hullámfüggvényének valószínűségi eloszlását adja. A teljes energia, illetve egyéb mennyiségek (pl.: betöltési szám) a módszerrel feltérképezett eloszlás várható értékeként kaphatók. A standard variációs Monte-Carlo módszer fizikailag motivált próbafüggvény (pl.: Jastrow-függvény) segítségével írja fel az eloszlást, a modell paramétereit variálva minimalizálja a teljes energiát. A próbafüggvény általánosításának tekinthető az úgynevezett Boltzmann-gép, mely ANN modellel (lásd 1. ábra) kifejtett non-lineáris próbafüggvényt feltételezve keresi a kvantumrendszerek

alapállapotú megoldását [4]. Az eredetileg spin rácsmodellekre bevezetett módszer kölcsönható elektronrendszerekre is általánosítható megfelelő leképezések révén [5]. ANN és kernel alapú modellekkel a gerjesztett energiák is hatékonyan megbecsülhetővé válnak, mely számítás a hagyományos kvantumkémiai módszerekkel jellemzően rendkívül költséges.

6.2. Sűrűség-funkcionál optimalizálás

A gépi tanulás a sűrűség-funkcionál elméleti (DFT) számítások hatékonyságát is javíthatja: egyrészt az irodalmi kicserélődési-korrelációs funkcionál alakok finomíthatók a neurális modellek tükrében [6]. Továbbá az ANN a teljes funkcionál alak feltérképezésére is használható mind kvantumkémiai rendszerekben [7,8,9], mind szilárdtestfizikai rácsmodellekben (pl.: Hubbard-gyűrű) [10], mely a Hohenberg–Kohn-tételek szerinti, pályamentes leírást teszi lehetővé.

6.3. Potenciálfelület interpolálás

A gépi tanulás egyik leggyorsabban fejlődő kémiai alkalmazási köre a potenciális energia felület (PES) [11], illetve a rezgési-forgási szinképek interpolált leírása. Az ANN alapú módszerekkel akár *ab initio* minőségű potenciál is generálható [12]. Példaként kiemelhető, hogy ANN potenciál felhasználásával sikerült a grafit-gyémánt átalakulás nukleációs jellegét tisztázni [13]. A gyorsan fejlődő terület módszerei gyakorlati problémákban is hatékonyan alkalmazhatók és nem pusztán a számításos kémikusok, hanem a vegyészek szélesebb körének figyelmére is számot tarthat.

A PES gépi tanulás alapú illesztésére gyakorlatban kétféle megközelítést alkalmaznak: molekula és atom alapú hálókat [14-17]. A molekuláris ANN nagy pontosságú eredményeket adhat az adott molekulára, de egyéb rendszerek leírásához a hálót újra kell tanítani. Az atom alapú leírás során a $\mathbf{R}=\{\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N\}$ magkoordinátákkal jellemzett molekuláris rendszer teljes potenciális energiáját a felépítő atomok energijárulékaival segítségével állítják elő,

$$E(\mathbf{R}) = \sum_i E_{atom}(|\mathbf{R}_i - \mathbf{R}_j| < r) + E_{hosszú\ táv}(\mathbf{R}), \quad (2)$$

ahol az r sugarú környezettől függ az \mathbf{R}_i koordinátájú atom energiája. A vizsgált molekula egyes atomjait kémiai elem-specifikus hálókkal jellemzik. Az ANN alapú erőterrel meghatározott atomi energiákat összegezve a molekula rövid távú energiakomponense adható meg. A hosszú távú Coulomb-kölcsönhatási effektusokat is gyakran ANN segítségével határozzák meg. A módszerrel akár több ezres atomból álló rendszer energiája is hatékonyan leírható [11].

A koordinátatér komplexitása miatt rendkívül fontos a megfelelő minőségű és mennyiségű tanítóminta előállítása. Megmutatták, hogy egyrészt a kritikus pontok figyelembevételével [18], másrészt a tudatosan tervezett (pl.: struktúra alapú) mintavételezéssel [19] jelentősen javítható a modell pontossága. Az úgynevezett aktív, illetve adaptív tanítási

protokollok szintén segíthetik a hatékony tanítást, ilyenkor a tanulás hatékonyságát monitorozva a megfelelő elemekkel iteratíván bővül a minták halmaza [20]. Az elmúlt évtizedekben neurális hálókkal és kernel ridge regresszió (KRR) alapú módszerekkel is sikeresen illesztettek potenciálokat. A legfrissebb tanulmányok szerint a KRR módszer pontosabb leírást adhat alacsonyabb dimenziós problémák esetén [21].

6.4. Pontosabb modell alacsony szintű információkból

Az ML módszerek sajátossága, hogy egy beparaméterezett modell kizárólag az aktuálisan betanított tulajdonságok vizsgálatára alkalmas. Ezzel szemben még az egyszerűbb, közelítő elméleti leírások is számos különböző tulajdonság becslésére alkalmasak. A Δ -ML a két módszer előnyeit próbálja egyesíteni oly módon, hogy magas és alacsony szintű eredmények különbségére tanított modell a kis számítási igényű elmélet további jósatait (pl.: entalpia, entrópia, korrelációs energia, szabadenergia) is javíthatja. E kernel ridge regresszió alapú korrekciós módszerrel például a $\text{C}_7\text{H}_{10}\text{O}_2$ izomerjeinek kísérleti pontosságú termokémiai leírását tudták adni. A Δ -ML eljárással kapható modell kiemelkedő általánosító képességét igazolja a zérus hőmérsékletű atomizációs energiákon végzett regresszió révén javított pontosságú becslés magas hőmérsékletű atomizációs entalpiákra. A módszer részletei és további érdekes alkalmazásai a [22]-es referenciában olvashatók.

Másik megközelítésben magát az alacsony szintű elméletet, azaz annak szemi-empirikus paramétereit hangolják adaptívan ML technikák segítségével a pontosabb modellezés érdekében [23].

Összefoglalás

Figyelembe véve a kvantumkémia komplexitását a gépi tanulás alapú mesterséges intelligencia semmi esetre sem fogja kiváltani a hagyományos kémiai numerikus módszereket a közeljövőben. Ugyanakkor a standard technikák és az ML paradigma ötvözése új modellezési lehetőségeket teremthet. Várható, hogy a módszerek fokozatos kiforrásával a gyakorlati jelentőségű alkalmazások köre és száma továbbra is dinamikus módon fog nőni bizonyítva az új tudományterület „életképességét”.

Az érdeklődők számtalan bevezető jellegű könyvből [24-27] informálódhatnak a terület elméleti háttérének részleteiről. Számos ML módszer hatékony implementációja szabadon elérhető Python programcsomagok (pl.: Scikit-Learn, Mlflow, PyTorch, Keras, TensorFlow) részeként [28], melyek segítségével akár összetettebb feladatok is pár sornyi programkóddal megoldhatók. Az általános felhasználási célú csomagokon kívül már kifejezetten PES illesztésekre optimalizált kódok is szabad hozzáférésűek (pl.: AMP [29], ANI [30], Schnet [31], TensorMol [32], QUIP [17]).

Köszönetnyilvánítás

Köszönet Nemes Csabának és a bírálóknak a hasznos észrevételekért. Köszönet az MTA Bolyai János Kutatási Ösztöndíj támogatásáért. A közlemény az Információs és Technológiai Minisztérium ÚNKP-20-5 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Hivatkozások

- Schleder, G. R.; Padilha, A. C. M.; Mera Acosta; C.; Costa, M.; Fazzio, A. *J. Phys. Mater.* **2019**, 2, 032001. <https://doi.org/10.1088/2515-7639/ab084b>
- Dral, P. O. *J. Phys. Chem. Lett.* **2020**, 11, 6, 2336–2347. <https://doi.org/10.1021/acs.jpclett.9b03664>
- Cao, Y.; Romero, J.; Olson, J.; Degroote, M.; Johnson, P. D.; Kieferová, M.; Kivlichan, I. D.; Menke, T.; Peropadre, B.; Sawaya, N.; Sim, S.; Veis, L.; Aspuru-Guzik, A. *Chemical Reviews* **2019**, 119, 10856–10915. <https://doi.org/10.1021/acs.chemrev.8b00803>
- Carleo, G.; Troyer, M. *Science* **2017**, 335, 6325. <https://doi.org/10.1126/science.aag2302>
- Choo, K.; Mezzacapo, A.; Carleo, G. *Nat Commun* **2020** 11, 2368. <https://doi.org/10.1038/s41467-020-15724-9>
- Schütt, K.T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. *Nat Commun* **2019**, 10, 5024. <https://doi.org/10.1038/s41467-019-12875-2>
- Zheng, X.; Hu, L.; Wang, X.; Chen, G. *Chem. Phys. Lett.* **2004**, 390, 186–192. <https://doi.org/10.1016/j.cplett.2004.04.020>
- Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M.E.; Burke, K.; Müller, K.-R. *Nat Commun* **2017**, 8, 872. <https://doi.org/10.1038/s41467-017-00839-3>
- Nagai, R.; Akashi, R.; Sugino, O. *npj Comp. Mater.* **2020**, 6, 43. <https://doi.org/10.1038/s41524-020-0310-0>
- Nelson, J.; Tiwari, R.; Sanvito S. *Phys. Rev. B* **2019**, 99, 075132. <https://doi.org/10.1103/PhysRevB.99.075132>
- Höltzl, T.; Veszprémi T. *Kémiai szimulációk az atomoktól a vegyipari reaktorokig*, Akadémiai Kiadó, **2020**, ISBN:978-9630599726
- Behler, J. *Phys. Chem. Chem. Phys.* **2011**, 13, 17930–17955. <https://doi.org/10.1039/clcp21668f>
- Khaliullin, R. Z.; Eshet, H.; Kuhne, T. D.; Behler, J.; Parrinello, M. *Nat. Mater.* **2011**, 10, 693–697. <https://doi.org/10.1038/nmat3078>
- Behler, J. *Int. J. Quant. Chem.* **2015**, 115, 1032–1050. <https://doi.org/10.1002/qua.24890>
- Behler, J.; Parrinello M. *Phys. Rev. Lett.* **2007**, 98, 146401 <https://doi.org/10.1103/PhysRevLett.98.146401>
- Bartók, A.P.; Payne, M.C.; Kondor, R.; Csányi G. *Phys. Rev. Lett.* **2009**, 104, 13, 136403 <https://doi.org/10.1103/PhysRevLett.104.136403>
- Bartók, A.P.; Csányi G. *Int. J. Quant. Chem.* **2015**, 115, 16, 1051–1057 <https://doi.org/10.1002/qua.24927>
- Gastegger, M.; Marquetand, P. *J. Chem. Theory Comput.* **2015**, 11, 2187–2198. <https://doi.org/10.1021/acs.jctc.5b00211>
- Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. *J. Chem. Phys.* **2017**, 146, 244108. <https://doi.org/10.1063/1.4989536>
- Li, Z.; Kermode, J. R.; De Vita, A. *Phys. Rev. Lett.* **2015**, 114, 096405. <https://doi.org/10.1103/PhysRevLett.114.096405>
- Kamath, A.; Vargas-Hernandez, R. A.; Krems, R. V.; Carrington, T., Jr.; Manzhos, S. *J. Chem. Phys.* **2018**, 148, 241702. <https://doi.org/10.1063/1.5003074>
- Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2015**, 11, 2087–2096. <https://doi.org/10.1021/acs.jctc.5b00099>
- Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. *J. Chem. Theory Comput.* **2015**, 11, 2120–2125. <https://doi.org/10.1021/acs.jctc.5b00141>
- Goodfellow, I.; Bengio, Y.; Courville A. *Deep Learning*, The MIT Press, **2016**, ISBN:978-0-262-03561-3
- Mostafa, Y. A.; Magdon-Ismail, M.; Lin, H.-T. *Learning From Data*, AMLBook, **2017**, ISBN:978-1-60049-006-4
- Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning*, Cambridge University Press, **2014**, ISBN:978-1107057135
- Harrington, P. *Machine Learning in Action*, Manning Publications, **2012**, ISBN:9781617290183
- Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, **2017**, ISBN:978-1491962299
- Khorshidi, A.; Peterson, A. A. *Comp. Phys. Com.*, **2016**, 207, 310–324. <https://doi.org/10.1016/j.cpc.2016.05.010>
- Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chem. Sci.*, **2017**, 8, 3192–3203. <https://doi.org/10.1039/C6SC05720A>
- Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. *J. Chem. Phys.* **2018**, 148, 241722. <https://doi.org/10.1063/1.5019779>
- Yao, Y.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J., *Chemical Science*, **2018**, 9, 2261–2269. <https://doi.org/10.1039/C7SC04934J>

Applying machine learning in computational quantum chemistry

Recently, the computational quantum chemistry community is also relying more and more on machine learning-based techniques, i.e., new methods and applications are emerging which synthesize computational chemistry and artificial intelligence [1, 2]. In this introductory paper, we outline the basic idea of the most important teaching methods and give a brief overview of the latest and most interesting chemically relevant research topics.

The branch of artificial intelligence in which the model is synthesized using sample data without explicit programming of rules is called machine learning (ML). The system tries to identify regularities based on the patterns in the training data. In case of successful teaching, the model does not learn by rote, but it is able to give a correct solution to the problem in general. The data used for training can be divided into two types according to its structure: 1) typical input sample data and their correct "solution" to be modeled (known as label) is also part of the data set 2) a set of unlabeled input parameters. It is important to note that the production of the label may require machine or human resources. ML methods can be divided into three main categories according to the type of teaching data and the nature of the teaching task. In the following, we briefly present the three paradigms, which are implemented by a number of different numerical methods, including the basics known from statistics and physics. Supervised learning is based on labeled data aiming to map the input object to the output information as accurately as possible. One of the simplest solutions to such problems is linear regression. In unsupervised learning, the teaching set contains no labels, the purpose of the method is to explore the internal structure of the system without any preconception. Typically used solution methods include principal component analysis and singular value decomposition. Reinforcement learning is aiming to maximize a cost function. Stochastic Monte-Carlo methods, among others, fall into this category.

The mentioned methods cannot be considered the full arsenal of ML. Numerous new techniques with different application potentials have been introduced in the recent decades, e.g. decision trees, Bayesian methods, kernel algorithms, k -nearest neighbor approach. Nowadays, the revolutionizing tool of machine learning is the artificial neural network (ANN) which tries to mimic the basic behavior of biological neurons with the simplest possible mathematical tools. In practice, a number of different paradigm-specific artificial neural models have been developed. The basic structure of the feedforward network containing single hidden layer of neurons is illustrated in Fig. 1.

A generalization of standard variational Monte Carlo methods is realized by the so-called Boltzmann machines, which search for a ground state solution of quantum systems assuming a non-linear

ansatz expressed by the ANN model. The method originally introduced for spin lattice models [4] can also be generalized to interacting electron systems through appropriate mappings [5]. ANN and kernel ridge regression based methods also make it possible to efficiently estimate excited-state energies, which is particularly useful considering that such calculation is typically expensive using traditional quantum chemical methods.

ML can also improve the efficiency of density-functional theory (DFT) calculations, i. e., literature exchange-correlation functional forms can be refined in light of neural models [6]. Furthermore, ANN can be used to map the full functional form in both quantum chemical systems [7,8,9] and solid-state physics lattice models (e.g., Hubbard ring) [10].

Typical application of machine learning is fitting the potential energy surface (PES) and the vibration-rotation spectra [11]. For many molecular systems, *ab initio* quality potential has been generated using the neural network [12]. In practice, two approaches are used for ANN-based fitting of PES: molecular and atom-based networks [14-17]. Molecular ANN is able to provide high-precision results for a given molecule, but the network needs to be re-trained to describe alternative systems. In atom-based description, the total potential energy of a molecular system is produced from the energy contributions of the building atoms, which are determined from the ANN for each atom and its chemical environment.

ANN methods parameterize model which is capable to describe the actually trained properties. In contrast, even simpler approximate theoretical descriptions are suitable for estimating a number of different properties. Δ -ML attempts to combine the advantages of the two approaches in such a way that further predictions of low-level theory can be refined using a model trained for the difference between solutions calculated by high-level and low-level methods. Among others, the Δ -ML method has been used to provide an improved estimate of high-temperature atomization enthalpies from teaching on atomization energies [22]. In another approach, the low-level theory itself, i.e., its semi-empirical parameters, is adaptively tuned using ML techniques to provide more accurate model [23].

Given the complexity of quantum chemistry, machine learning-based artificial intelligence will by no means replace traditional computational chemistry methods in the near future. However, the combination of standard techniques and the ML paradigm can create new modeling opportunities. It is expected that with the gradual maturation of the methods, the range and the number of applications of practical importance will increase dynamically. Interested readers find the details of the theoretical background of the field in numerous introductory books [24-27].