JE STROJOVÉ UČENÍ BUDOUCNOSTÍ TEORETICKÉ CHEMIE?

KAREL BERKA[a], ŠTĚPÁN SRŠEŇ[b] a PETR SLAVÍČEK[b,c]

[a] Katedra fyzikální chemie, RCPTM, Přírodovědecká fakulta, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc, [b] Ústav fyzikální chemie, Vysoká škola chemicko-technologická v Praze, Technická 6, 166 28 Praha 6, [c] Ústav fyzikální chemie J. Heyrovského AV ČR, v.v.i., Dolejškova 2155, 182 23 Praha 8
petr.slavicek@vscht.cz

Došlo 6.8.18, přijato 21.8.2018.

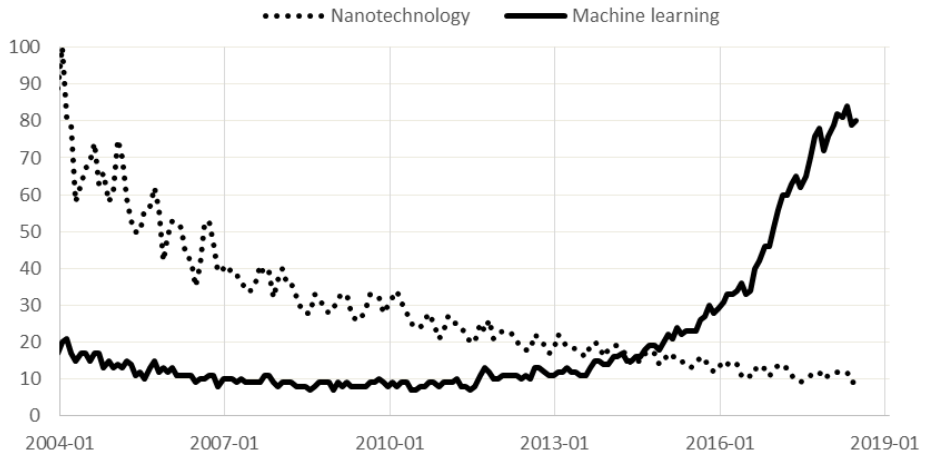
Klíčová slova: strojové učení, umělá inteligence, QSAR, kvantová chemie, teoretická chemie, neuronové sítě

Obsah

1. Úvod

Každá doba má svá klíčová, do úmoru opakovaná slovní spojení, u kterých navíc není přesně jasné, co vlastně znamenají. V oboru společného zájmu čtenářstva Chemických listů k takovýmto pojmům patřilo v posledních dekádách například slovo „nanotechnologie“. Oblíbenost určitých výrazů je možné sledovat nástrojem *Google Trends*[1] a na obr. 1 můžeme vidět, že zrovna pojem „nanotechnology“ je již za svým zenitem, minimálně z pohledu vyhledávání. Naproti tomu výraz „machine learning“ (tedy „strojové učení“, viz obr. 1) rychle nabývá na síle – podobnou závislost bychom nalezli také pro počet publikovaných vědeckých prací.

Frekvence užití tohoto výrazu pomalu v každé oblasti lidské činnosti již občas vyvolává alergickou reakci a podvědomou nechuť. Je k ní i určitý důvod. Umělá inteligence (AI, z angl. artificial intelligence) a strojové učení (ML, z angl. machine learning) představují jen vznosnější pojmenování statistických metod, používané algoritmy byly většinou publikovány již před delší dobou a zásadně nové myšlenky se v posledních letech vlastně neobjevily[2]. Nedá se ale zároveň přehlížet, že v oblasti AI dochází ke kvantitativním změnám, díky kterým jsou tyto techniky najednou použitelné. Je to dáno především neustále se zvětšujícím výpočetním výkonem a produkcí ohromného množství dat (dnes často označována jako „big data“), na kterých se dají jednotlivé algoritmy ML učit. Množství produkovaných dat roste exponenciálně a zároveň jsou tato data díky internetovým databázím dostupnější, přičemž veliké množství



Obr. 1. **Porovnání trendů vyhledávání slov „nanotechnology“ a „machine learning“ v *Google Trends* normované na počet vyhledávání slova „nanotechnology“ v roce 2004** (cit.[1])

učících dat je pro ML zcela zásadní. Důležitý je ovšem i moment psychologický: přístup umělé inteligence se postupně vylepšoval, v poslední době ale v mnoha oblastech dosahuje výsledků na úrovni člověka nebo jej překonává, ať už se to týká rozpoznávání obrazu[3], hry Go[4] nebo organické syntézy[5,6]. V tom je nejspíše současná situace revoluční, ve smyslu hegeliánského přechodu kvantity na kvalitu.

Naše krátká úvaha se zaměřuje na využití technik umělé inteligence v chemii, přičemž jako vodítko nám poslouží životní dráha jubilujícího profesora Rudolfa Zahradníka[7]. Rudolf Zahradník začínal svou kariéru v Ústavu hygieny práce a nemocí z povolání, kde se mimo jiné zabýval hledáním souvislostí mezi strukturou a biologickou aktivitou – v dnešní době typická oblast pro využití strojového učení. Hlavní renomé si ale vydobyl v oblasti kvantové chemie. I tato oblast možná dozná v brzké době výrazných změn díky technikám strojového učení.

2. Umělá inteligence, strojové učení a další pojmy

Člověk generace autorů tohoto textu si pod pojmem umělá inteligence v prvním okamžiku možná představí Arnolda Schwarzeneggera v roli terminátora, jehož využití v chemii je těžko myslitelné. Pojem zavedl John McCarthy, jeden ze zakladatelů oboru, v roce 1955 v žádosti o vědecký projekt – ve svébytném literárním žánru, ke kterému asi určitá nadsázka patří. V něm představil vizi, že každý aspekt lidského učení a inteligence může být v principu popsán tak podrobně, že mohou být vytvořeny stroje, které ho simulují[8]. V dnešní době existuje řada různých definic umělé inteligence, např. Oxford Dictionary[9] ji definuje jako teorii a vývoj počítačových systémů schopných vykonávat úkony, ke kterým je běžně potřeba lidská inteligence. Dle tohoto výkladového slovníku patří k těmto úkonům např. vizuální vjemy, rozpoznávání řeči, rozhodování či překlady. Pojem umělá inteligence ve smyslu, v jakém o ní píšeme v tomto textu, se týká zpracování dat*.

Definice umělé inteligence nespecifikuje, jak úkony vyžadující inteligenci strojově realizovat. V principu tak lze napsat dlouhý program plný podmínek, který bude říkat, jak se má systém chovat v každé konkrétní situaci. Praktičtější se však ukázalo nechat systém učit se na základě vlastních zkušeností, kdy není nutné v programu explicitně podchytit každou situaci, takovému přístupu se říká strojové učení. Algoritmy strojového učení dostanou učící data a na jejich základě jsou schopné pak činit závěry o datech nových. Ačkoli výše zmíněné pojmy jako umělá inteligence či strojové učení v nás evokují představy sofistikovaných technologií, v realitě se jedná pouze o aplikovanou statistiku. Jednoduchým příkladem strojového učení je regrese, kdy konečnou sadu bodů x, y prokládáme (známou) funkcí umožňující předpovědět hodnotu y pro dosud neměřenou veličinu x. V současnosti je strojové učení téměř výhradním přístupem k návrhu umělé inteligence a tyto pojmy tak často splývají.

Na základě toho, zda ke vstupním datům programu známe či neznáme výstup/odezvu, rozdělujeme strojové učení na učení s učitelem a učení bez učitele, existují však i jejich kombinace. Při učení s učitelem dostaneme např. hodnoty $\boldsymbol{x}$, y, kde $\boldsymbol{x}$ může představovat geometrii molekuly a y její excitační energii, a učíme se funkci f: $\boldsymbol{x} \rightarrow y$. Pro nové molekuly jsme pak schopni na základě jejich geometrie předpovědět jejich excitační energie. Učení s učitelem typicky zahrnuje klasifikační a již zmíněné regresní úlohy, těm se budeme věnovat nejvíce. Pokud by nás pouze zajímalo, zda molekula absorbuje v dané spektrální oblasti, tak by y nabývalo pouze hodnot ano/ne a jednalo by se o klasifikaci. Ve chvíli, kdy nás zajímají konkrétní hodnoty excitačních energií a y tedy již není diskrétní, tak se jedná o regresi.

V případě učení bez učitele dostaneme pouze vstupní hodnoty $\boldsymbol{x}$ a hledáme struktury v datech. Učení bez učitele zahrnuje nejčastěji úlohy shlukové analýzy (též klastrová analýza) či redukce dimenzionality. V případě shlukové analýzy třídíme jednotlivé vstupy do skupin tak, aby si vstupy v rámci každé skupiny byly co nejpodobnější, ale lišily se co nejvíce od vstupů z ostatních skupin. Jedná se o metodu příbuznou klasifikaci, ovšem zde neznáme, resp. nepoužíváme označení y pro učení a záleží tak pouze na podobnosti vstupů $\boldsymbol{x}$. Shluková analýza se používá např. ve velkých chemických databázích, kde je takto možno získat např. skupiny příbuzných molekul. Při redukci dimenzionality dat chceme z n-tice čísel vstupního vektoru $\boldsymbol{x}$ získat vybraný počet čísel menší než n, která jsou pro nás nějakým způsobem relevantní. Nejjednodušším případem je pouze výběr relevantních hodnot z vektoru $\boldsymbol{x}$. Pokud chceme např. simulovat systém metodou molekulové dynamiky, ale nemůžeme si dovolit modelovat všechny stupně volnosti, tak zafixujeme některé vazebné délky či úhly, čímž jsme vlastně provedli „ruční“ redukci dimenzionality dat. V pokročilejším případě nevybíráme pouze z vektoru $\boldsymbol{x}$, ale používáme nějakou transformaci vstupních dat. Příkladem je analýza hlavních komponent, kde jsou získány nové proměnné jako lineární kombinace hodnot vektoru $\boldsymbol{x}$

* Různé zdroje rozdělují umělou inteligenci do odlišných kategorií, obecněji lze však rozlišit dvě základní kategorie, obecnou a aplikovanou umělou inteligenci. Obecná umělá inteligence by měla v principu být schopná vykonávat jakékoliv úkony, aplikovaná umělá inteligence je pak navržena k jednomu či více konkrétním úkolům. Zatímco obecná umělá inteligence je pojem, který straší některé soudobé vědce, politiky a milovníky sci-fi žánru, aplikovaná umělá inteligence je běžně používaná, a setkáváme se s ní třeba v podobě rozpoznávání obličejů vaším fotoaparátem či v Google překladači.

tak, že dochází k co nejmenší ztrátě informace a jednotlivé proměnné jsou vzájemné dekorelované. Redukce dimenzionality lze např. použít před regresí či klasifikací, čímž se sníží počet odhadovaných parametrů.

Strojové učení můžeme realizovat velkou a stále rostoucí řadou konkrétních algoritmů, jejich kompletní výčet však přesahuje rozsah tohoto textu. Stojíme např. před výše zmíněným úkolem odhadnout ze struktury molekuly její excitační energii[10]. K učení máme k dispozici sadu molekul, jejichž excitační energie známe a k nimž jsme schopni přiřadit strukturní deskriptor $\boldsymbol{x}$ (což může být třeba soubor meziatomových vzdáleností dané molekuly). Kdybychom měli k dispozici úplně všechny molekuly, tak bychom pouze hledali v gigantické tabulce. S použitím konečného souboru učících dat můžeme provést regresi. Excitační energii E_{exc} můžeme pro nové geometrie s deskriptorem $\boldsymbol{x}$ hledat například ve formě:

$$E_{exc}(x) = \sum_{i=1}^{n} \alpha_i e^{-\frac{1}{\sigma}\|x - x_i\|} \qquad (1)$$

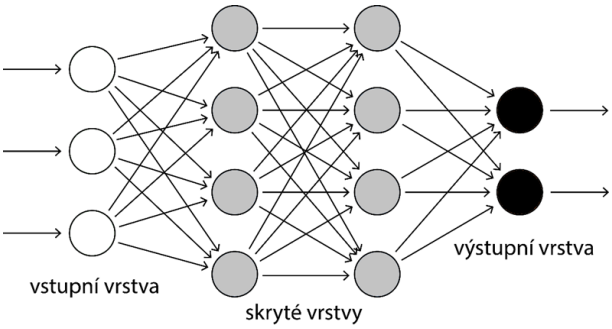
kde regresní koeficienty α_i získáme poměrně jednoduše minimalizací z učících dat a hyperparametr σ si zvolíme nebo získáme další optimalizací. Tato metoda se nazývá hřebenová regrese s jádrovou transformací (KRR, z angl. Kernel Ridge Regression) a je příkladem metod založených na tzv. jádrových transformacích. Exponenciální funkce s manhattanskou normou v rovnici (*1*) představuje ono jádro, které implicitně transformuje deskriptory do vyšší dimenze, díky čemuž lze odhadovat i nelineární funkce. Toto jádro se nazývá Laplaceovo, alternativně lze jako jádro použít i jiné funkce, např. polynomiální či Gaussovo jádro[11]. Do této skupiny metod dále patří např. metoda podpůrných vektorů (SVM, z angl. Support Vector Machines) či regrese gaussovskými procesy (GPR, z angl. Gaussian Process Regression)[12].

Aktuálně největšímu zájmu se těší algoritmy založené na neuronových sítích. Neuronová síť je biologicky inspirovaný výpočetní model, kde jsou umělé neurony navzájem pospojovány tak, že každý z nich má několik vstupů, ale pouze jeden výstup, který je posílán do neuronů v další vrstvě (viz obr. 2). Neuronové sítě lze použít jak pro řešení klasifikačních problémů, tak k regresi.

Každý neuron má následující jednoduchou strukturu:

$$y_j = f\left(b_j + \sum_i w_{ij} x_i\right) \qquad (2)$$

kde y_j představuje výstup j-tého neuronu, x_i jsou vstupy z předchozích vrstvy neuronů, w_{ij} jsou váhy odpovídající propojení příslušných dvou neuronů a b_j je tzv. prahová hodnota aktivace neuronu. Funkce f zde představuje aktivační nebo také přenosovou funkci neuronu, kterou může v nejjednodušších případech být například skoková či lineární funkce. V procesu učení jsou pak zpětnou propagací například pomocí gradientové metody upravovány parametry sítě, tedy váhy w_{ij} a prahové hodnoty b_j. Pokud vezmeme v úvahu jednoduchou neuronovou síť s jedním neuronem s jedním vstupem a identitou jako aktivační funkcí, tak pro výstupní hodnotu dostaneme následující tvar:



Obr. 2. **Znázornění struktury neuronové sítě**. Bílou barvou je znázorněna vstupní vrstva, poté následuje libovolný počet skrytých vrstev (šedá barva) a nakonec černě výstupní vrstva. Neurony ve vstupní vrstvě jsou typicky pasivní neboli pouze přeposílají vstupy do dalších neuronů. Každá vrstva může mít odlišný počet neuronů

$$y = b + wx \qquad (3)$$

což není nic jiného než předpis přímky a jedná se tak o lineární regresi.

S prvním modelem neuronu přišli již v roce 1943 Warren McCulloch a Walter Pitts[13], v roce 1957 pak Frank Rosenblatt vytvořil jednovrstvou neuronovou síť nazvanou perceptron, která sloužila k binární klasifikaci[14]. Po počátečním nadšení se však ukázalo, že perceptron je schopný klasifikovat pouze množinu dat, která jdou oddělit přímkou. Naneštěstí vydali Marvin Minsky a Seymour Papert knihu *Perceptrons*, kde jsou probírána mimo jiné úskalí perceptronu[15]. Důkazy v této knize byly často špatně interpretovány a citovány a větší část komunity uvěřila, že ani vícevrstvé neuronové sítě nedokáží implementovat některé logické funkce jako je nelineární XOR, kvůli čemuž na zhruba dvacet let zájem o neuronové sítě ochabl. Postupně se však začaly používat vícevrstvé neuronové sítě, spojité přenosové funkce a velký posun znamenalo zavedení již výše zmíněné metody zpětné propagace Paulem Werbosem[16].

Jako mělké jsou typicky označovány neuronové sítě s jednou skrytou vrstvou. Popisnost dat pomocí sítě je možné zvyšovat přidáváním neuronů do stávajících vrstev neboli rozšiřováním sítě do šířky. Čím více neuronů se pro síť použije, tím přesnější výsledky může poskytovat. S počtem neuronů však také roste výpočetní náročnost a potřebné množství učících dat. U širokých sítí dochází také často k přeučení (angl. overfitting), což je případ, kdy pro učící data obdržíme sice přesné výsledky, ale model neodpovídá realitě a pro ostatní vstupy obdržíme špatné výsledky. Jednoduchý příklad přeučení obdržíme, pokud bychom prokládali lehce zašuměná data pocházející z přímky pomocí polynomu vyššího řádu. Ukázalo se jako

žádoucí vstupní data pro mělké sítě nejprve připravit, což znamená extrahovat deskriptory (nebo též atributy či rysy, v angličtině se používá termín „features“) vhodné pro danou aplikaci. Přehled používaných chemických deskriptorů naleznete i ve zvláštním čísle Chemických listů z listopadu 2017 (cit.[17,18]).

V posledních letech se dočkal nevídaného rozvoje přístup tzv. hlubokých sítí neboli sítí s velkým množstvím vrstev (angl. označovány jako „deep learning“). Jedná se o alternativní přístup ke zlepšování popisu dat pomocí neuronové sítě. Oproti širokým mělkým sítím mají hluboké neuronové sítě několik výhod. Typicky mají díky moderním učícím algoritmům menší náchylnost k přeučení a vhodnou konstrukcí vrstev lze deskriptory automaticky extrahovat z původních dat. Při zachování popisnosti existují také pro některé problémy architektury hlubokých sítí, které obsahují celkově menší počet neuronů, resp. spojení než odpovídající široké mělké sítě, což typicky znamená kratší dobu učení a potřebu menšího množství dat[19]. Nicméně ne vždy takovou architekturu známe, resp. jsme schopni najít. V praxi ovšem mívají hluboké sítě většinou mnohem více parametrů a potřebují více učících dat než sítě mělké, ovšem mají díky tomu také větší popisnost. Pro hluboké sítě také existují speciální učící mechanismy, kdy lze jednotlivé vrstvy „předučit“, čímž se stává učení efektivnější. Jak jsme již naznačili, nevýhodou takových sítí je jejich složitější konstrukce, ale pro menší datové sady není často mnoho lepších možností[20].

Neuronové sítě představují zcela obecný rámec, který v principu umožňuje aproximovat jakoukoliv funkci, nicméně složitost takové sítě může vylučovat její praktické použití. Záleží tedy i na fyzikální podstatě problému. Pokud existují rysy, na jejichž základě je relativně jednoduché vyjádřit hledanou funkci/veličinu, tak nám mohou dobře posloužit i menší mělké sítě[19]. Tuto závislost samozřejmě typicky neznáme, jinak bychom nemuseli používat strojové učení, ale často víme, že existuje. S jednotlivými technikami strojového učení se v poslední době nadšeně experimentuje i v řadě oblastí chemie. Podívejme se teď na dvě z nich.

3. Struktura a biologická aktivita: Od QSAR ke strojovému učení

Vlastnosti látek jsou zakódovány ve struktuře jejich molekul, a proto se neustále navrhují, vyvíjejí a testují nové molekuly k nejrůznějším účelům. Tzv. chemický prostor, tj. množina molekul, které můžeme syntetizovat, je přitom prakticky nekonečný[21]. Například počet molekul s potenciálním farmakologickým účinkem je Lipinským[22] odhadován na 10^{60}, což je číslo mnohonásobně přesahující odhady počtu hvězd ve vesmíru (zhruba 10^{24}, cit.[23]) či spojení mezi neurony v našem mozku (zhruba 10^{15}, cit.[24]). Toto množství je tak naprosto zjevně daleko za hranicemi možností jakékoliv, byť vysoce propustné syntézy s následnou charakterizací. Proto je stále atraktivnější přístup založený na racionálním návrhu látek.

Předpověď vlastností molekul na základě jejich struktury má dlouhou historii, počínající na přelomu 19. a 20. století, kdy vznikly první studie o narkotických účincích látek od Overtona[25,26] a Meyera[27]. O strojovém učení v té době pochopitelně nikdo nevěděl. Tito autoři přišli se vztahem mezi hydrofobními vlastnostmi těkavých látek a jejich narkotickými účinky. V podstatě šlo o lineární regresi, jakožto nejjednodušší statistickou metodu vyhledávající vztah mezi proměnnými, a tedy i hledání kvantitativních vztahů mezi strukturou a funkcí (QSAR, z angl. quantitative structure-activity relationship).

Výzkumy tohoto typu stály ovšem po následující dekády spíše na okraji zájmu, neboť hledání jednoduchých souvislostí pro složité jevy bylo považováno za příliš troufalé. Zájem o tyto otázky se v plné síle probudil až v šedesátých letech. K pionýrům při studiu vlivu struktury na biologickou aktivitu z té doby patří i Rudolf Zahradník, který na přelomu padesátých a šedesátých let zkoumal se svým kolegou MUDr. Chvapilem vliv různě dlouhých alifatických řetězců na biologickou funkci v homologních řadách látek se shodnou funkční skupinou. Použil přitom rovnici převzatou z fyzikální organické chemie[28]:

$$\log\left(\frac{\tau_i}{\tau_{Ref}}\right) = \alpha \cdot \beta_i \qquad (4)$$

kde τ_i značí hodnotu biologické aktivity molekuly i v homologické řadě, τ_{Ref} je pak hodnota pro referenční molekulu, např. molekulu ethylového derivátu. Konstanta biologické aktivity substituentu β_i popisuje vlastnosti alkylového substituentu a její hodnota nezáleží na funkční skupině homologické řady ani na konkrétní biologické aktivitě. Konstanta α naopak charakterizuje biologický systém. Platnost této rovnice ověřili Chvapil se Zahradníkem na 5 biologických systémech, pro 3 různé funkční skupiny a 14 alifatických řetězců. V následující studii[29] pak již Zahradník uvažoval, že by biologické účinky mohly být v principu řízeny partičním koeficientem mezi polární a nepolární fází, ale nadále používal jen jednu konstantu α pro popis biologického systému pomocí vztahu $\alpha\beta$-typu a povšiml si, že jde vlastně o lineární vztah pro volnou energii, který se již dříve používal v teoretické organické chemii pro aromatické[30] či alifatické[31] systémy. Pokud se nějaká molekula pro určitou regresní řadu odlišovala, předpokládal, že tato látka působí jiným mechanismem. Následně Zahradník společně s Kouteckým začali uvažovat o využití kvantové chemie pro výpočty parametrů pro rovnice $\alpha\beta$-typu[32]. Profesor Zahradník se postupně po přechodu z Ústavu hygieny práce a chorob z povolání na Ústav fyzikální chemie ČSAV ponořil do vývoje kvantové chemie a spoluzaložil školu československé teoretické chemie, k níž ve třetí, resp. čtvrté generaci patří i autoři tohoto textu.

Jak se později ukázalo, Zahradníkovy konstanty β korelovaly s hodnotami hydrofobních konstant navržených Corwinem Hanschem, který v roce 1962 publikoval v časopise Nature studii o biologické aktivitě popsané Hammetovými substitučními konstantami a partičními

koeficienty[33]. Hansch posléze odstartoval rozvoj QSAR metod zobecněním svého přístupu do složitější rovnice[34]:

$$\log C = k_1(\log P_{ow})^2 + k_2 \log P_{ow} + k_3\sigma + k_4 E_S + k_5 \qquad (5)$$

kde C je hodnota (biologické) aktivity (např. LD_{50}, EC_{50}), P_{ow} je rozdělovací koeficient systému oktanol-voda, σ je Hammettova konstanta pro substituent, E_s je Taftova sterická konstanta pro substituent a k_1, k_2, k_3, k_4 *a* k_5 jsou regresní konstanty. Druhá mocnina simuluje nelineární průběh závislosti parabolickou funkcí, nicméně často se používá i aproximace přímkou, pokud není soubor látek dostatečně rozsáhlý, aby pokryl i nelineární části závislosti. Rudolf Zahradník tedy stál již blízko Hanschově rovnici, stačilo by pouze zavést další parametry.

Metody QSAR se následně rychle rozvíjely[35]. Kolem 90. let se staly standardem pro organickou či farmaceutickou chemii. Statistické techniky, které se zpočátku používaly, byly především klasická regresní analýza, analýza hlavních komponent a podobně. Tyto techniky jsou ale často příliš zjednodušující, a tak se postupně začaly nahrazovat pokročilejšími metodami, které používají místo jedné definované regresní rovnice spíše kombinaci pravděpodobnostních modelů využívajících neuronových sítí a dalších metod strojového učení. Všechny tyto novější modely nemají za výsledek jen jednu rovnici, ale tvoří je kombinace různě vybraných rovnic, které jsou automaticky poskládány s různými váhami a větším množstvím parametrů, jak jsme již ukázali v předchozím oddíle. Jinými slovy, nemusíme už předpokládat určitý model, metody strojového učení si jej najdou samy. Pokud mezi strukturou molekuly a biologickou aktivitou existuje nějaký vztah a správně zkonstruujeme neuronovou síť, tuto souvislost odhalíme. Platíme za to ovšem určitou ztrátou porozumění – korelační rovnice Hammetova typu vzbuzuje naši obraznost a naznačuje nám hlubší souvislosti, neuronové sítě nám místo toho dají (co možná nejpřesnější) čísla a je zde problém s jejich interpretací[36].

Jako příklad použití metod strojového učení v oblasti QSAR může sloužit predikce průchodnosti látek kůží, kdy v roce 1992 přišli Potts a Guy s úspěšnou rovnicí[37], které stačily jako deskriptory pouze molární hmotnost a rozdělovací koeficient oktanol-voda. Dokázali tak popsat 67 % variability v experimentálních datech pro průchodnost látek kůží. Tato data měla rozsah měření přes pět řádů a model Pottse a Guye je predikoval s průměrnou chybou 0,69 logaritmických jednotek, tedy v rozsahu přes 1 řád. Novější a přesnější model z roku 2015 už používá strojového učení s použitím 9 deskriptorů a vylepšil popis na 86 % variability na dostupných experimentálních datech. Průměrnou chybu se podařilo stlačit na 0,39 logaritmických jednotek, a tak jsou průchodnosti látek kůží určovány alespoň řádově správně[38].

Další výhodou pro nasazení strojového učení je také větší množství dostupných biologických i chemických dat, které dnes máme k dispozici díky použití kombinatoriální chemie a robotických metod i teoretických výpočtů s vysokou propustností. Již delší dobu se používají metody strojového učení na predikci fyzikálně-chemických vlastností látek, jako jsou odhady rozpustnosti ve vodě[39] či partičních koeficientů oktanol/voda[40], a jejich uplatnění v tomto odvětví chemie se neustále rozšiřuje[5,41]. Další příklady nasazení strojového učení pro predikci vlastností v biologických systémech a v kondenzované fázi lze nalézt např. v souhrnném článku[42].

Nasazování metod strojového učení pro predikci vlastností látek ale občas trpí několika nešvary – nedostatkem dat a jejich častou nekonzistencí. Kdo byl někdy v laboratoři, ví, že získávat konzistentně stejné výsledky téhož měření je záležitostí mnoha let učení a provádění správných laboratorních postupů. Nejsou neobvyklé situace, kdy měření aktivity stejné látky na stejném biologickém systému dvěma laboratořemi dává dva zcela neporovnatelné výsledky kvůli mnoha drobným nuancím používaných citlivých metod. Tím se bohužel (nebo pro chemiky bohudík) chemie stává pro metody strojového učení obtížnější, neboť datové soubory jsou většinou menší než např. u rozpoznávání obličeje, kde jsou data sbírána víceméně automaticky. Velmi těžko se pak porovnávají jednotlivé metody strojového učení vůči sobě, protože se jen málokdy tytéž metody porovnávají na stejných datech. Proto se začaly shromažďovat standardizované sady dat s metrikami přesnosti a jejich největší sbírkou je dnes nejspíše MoleculeNet[43], což je sada pro výpočty ať biologických, fyzikálně-chemických či i kvantových vlastností molekul, které podobné porovnání metod strojového učení umožňují.

4. Strojové učení v teoretické chemii

Odhadovat biologickou aktivitu ze struktury molekul je nejspíše příliš obtížný úkol, už jen proto, že část informace nemusí být ve struktuře molekuly vůbec zakódovaná. Například prostupnost molekul skrze kůži bude záviset nejspíše i na vlastnostech kůže. Můžeme být ale skromnější a ptát se po jednodušších vlastnostech molekul, jako je třeba slučovací entalpie molekuly, ionizační energie či třeba její elektronové spektrum. Tyto informace můžeme typicky získat vyřešením elektronové Schrödingerovy rovnice pro molekulu s geometrií danou souřadnicemi $\boldsymbol{R}$:

$$\hat{H}_{el}\,\Psi_i^{el} = E_i^{el}\Psi_i^{el} \qquad (6)$$

kde index i označuje příslušný elektronový stav. (Většinou přibližné) řešení této rovnice je doménou kvantové chemie. Veškerá informace o molekule je obsažena v elektronovém hamiltoniánu $\hat{H}_{el}$, zde se objevuje geometrie molekuly (tedy polohy jader $\boldsymbol{R}_I$) a nábojová čísla jader Z_I. Schrödingerova rovice představuje nástroj, pomocí kterého získáme z geometrií a nábojů jader energii elektronů E_i^{el}. Řešení této rovnice je známo svými obtížemi, které kvantoví chemikové s jistými úspěchy překonávají již po několik generací. Naskýtá se ale otázka, zda tuto komplikovanou diferenciální rovnici vůbec řešit potřebujeme.

Z matematického hlediska totiž stačí vědět, že elektronová energie je funkcí souboru poloh a nábojových čísel jader:

$$E_i^{el} = E_i^{el}\left(\{\boldsymbol{R}_I\},\{Z_I\}\right) \qquad (7)$$

Konkrétní tvar této funkce je pak možné získat metodami strojového učení, např. ve formě neuronových sítí – existuje určitá naděje, že strojové učení může být rychlejší než vývoj nových kvantově-chemických metod, resp. nárůst počítačového výkonu. Věc má samozřejmě háček. Na první pohled není zřejmé, zda je možné závislost (*7*) prakticky uchopit. Výsledek také silně závisí na reprezentaci molekulární struktury. Zdá se ale, že tento přímočarý způsob k získání chemických vlastností molekul, který je přitom ukotven v kvantové chemii, může být úspěšný[44].

Podobně jako v případě využití strojového učení pro odhad vztahů mezi strukturou a biologickou aktivitou nejde ale v zásadě o nic nového. Vlastnosti látek se z molekulární struktury v chemii odhadují již skoro dvě století. Tak třeba hustota kapalin se odhadovala z atomárních příspěvků již v polovině 19. století[45], kdy samotné existenci molekul řada lidí nevěřila. Jiným příkladem je odhad termochemických vlastností sloučenin pomocí disociačních entalpií vazeb[46]. Tato nesmírně jednoduchá technika často funguje, a pokud někdy selhala – jako v případě benzenu – směřovalo to k objevu nových jevů. Techniky strojového učení automatizují odhadové metody, znovu je ale nutné připomenout, že tak s přesností ztrácíme intuici.

Je možný také kompromis mezi kvantovou teorií molekul a statistickými přístupy, kdy výpočet pomocí „levné“ a rychlé metody následně doplníme také velmi rychlou korekcí natrénovanou strojovým učením[47]. Myšlenkou tohoto přístupu je fakt, že většina fyzikální podstaty je zahrnuta i ve vysoce aproximativních metodách a za posledních pár procent přesnosti získaných kvalitnější metodou zaplatíme neúměrně větší cenu, přičemž rozdíl mezi dvěma kvantovými metodami se lze naučit pomocí strojového učení jednodušeji. Ramakrishnan a spol. ukázali již v roce 2015, že jsou schopni tímto přístupem dosáhnout chemické přesnosti (≈1 kcal mol^{-1}). Pojem chemická přesnost zde ovšem neznamená přesnost vůči experimentu, ale vůči exaktnější metodě, která je použita k učení. Přesnost strojového učení nemůže (cíleně) převýšit přesnost metody, pomocí které se učí, neboli pokud použijeme k naučení DFT metodu, nemůžeme očekávat výsledky na úrovni například metody CCSD(T). V rámci výše uvedeného přístupu se možná dočkáme revitalizace starších levných metod, jako jsou např. semiempirické metody. Tyto metody vychází se struktury kvantové mechaniky, řada parametrů je zde ale určována z experimentálních dat, jinými slovy metodu musíme opět nejdříve naučit na určité skupině molekul obdobně jako v moderních přístupech strojového učení.

Numerické algoritmy strojového učení ale mohou změnit přímo i *ab initio* kvantovou chemii. Přesné výpočty jsou zde komplikovány korelacemi mezi pohyby jednotlivých částic, což vede k exponenciálně se zvyšujícím nárokům s rostoucí velikostí systému. Techniky strojového učení aplikované na samotnou vlnovou funkci mohou být v některých případech schopny výrazně výpočty urychlit[48].

V dnešní době se již techniky strojového učení používají pro predikci rozličných molekulárních vlastností jako atomizačních energií[49], slučovacích tepel[50], dipólových momentů, vibrační energie nulového bodu, tepelných kapacit[51], rozdílu mezi energiemi HOMO-LUMO či polarizovatenosti[52], a to s chemickou přesností. Další oblastí, kde začíná strojové učení nacházet uplatnění, je spektroskopie. Gastegger a spol. dokázali pomocí přístupu založeném na neuronových sítích velmi přesně předpovídat infračervená spektra pro alkany až o 200 atomech a protonovaný tripeptid alaninu[53]. Ramakrishnan a spol. dokázali zase úspěšně natrénovat rozdíl excitačních energií mezi metodou TD-PBE0 s malou bází a metodou CC2 s rozsáhlou bází, nicméně u oscilátorových sil již dosáhli pouze minimálního zlepšení oproti TD-PBE0, nejspíše kvůli častému rozdílnému pořadí stavů u zmíněných metod[10]. Z toho je zřejmé, že ani pomocí strojového učení nejsme zatím schopni získat libovolné vlastnosti molekul. Hledání korelací mezi metodami není koneckonců myšlenka příslušející pouze strojovému učení, obdobný přístup mimo rámec strojového učení byl použit např. k výpočtu elektronových spekter problematických Criegeeho intermediátů[54].

Největší význam mohou mít ale techniky umělé inteligence v oblastech, kde *ab initio* teorie zatím neposkytuje dostatečně přesné výsledky. Například odhad rozpustnosti pro *ab initio* metody patří k náročným úkolům, neboť v sobě zahrnuje nejen vlastnosti samotné molekuly, ale i vlastnosti rozpouštědla. Odhady rozpustnosti za pomoci ML jsou rozumně dobré, využití hlubokých sítí přitom umožňuje automatickou extrakci rysů, jejichž optimální množina není v tomto případě známá[55]. K podobným oříškům patří odhad krystalových struktur a jejich energetické pořadí pro složitější krystaly. Často je v této souvislosti citován výrok Johna Maddoxe, tehdejšího editora časopisu Nature: „*Jedním z trvalých skandálů ve fyzikálních vědách je pokračující neschopnost předpovědi krystalické struktury látky z jejího složení*.“[56] Techniky strojového učení začínají být v této souvislosti úspěšné jak pro molekulární[57], tak i anorganické krystaly[58].

Strojové učení bývá často nasazeno v oblasti materiálů, kde našlo uplatnění např. při hledání organických polymerů pro solární články, které mají řadu výhod oproti anorganickým materiálům, avšak jejich účinnost zatím nedosahuje takových hodnot[59]. Dalším příkladem je predikce indexu lomu organických polymerů, která najde uplatnění v oblasti optiky a optoelektroniky[60], nebo design anorganických scintilátorů[61]. Do některých odvětví se strojové učení teprve dostává kvůli malému počtu učících dat. K takovým oblastem patří např. predikce výbušných vlastností materiálů, jako jsou detonační tlak a rychlost, energie exploze, slučovací tepla či hustoty. Elton a spol. byli schopni předpovídat tyto veličiny s chybou v rozmezí 4 až 11 % při použití pouhých 87 molekul použitých pro natrénování hřebenové regrese[62].

Rychle se rozvíjející oblastí strojového učení jsou i molekulárně dynamické simulace. Síly působící na atomy během dynamických simulací mohou být počítány v každém časovém kroku na *ab initio* úrovni nebo mohou být získány z empiricky designované hyperplochy potenciální energie (silového pole). *Ab initio* molekulová dynamika je v posledních desetiletích na zjevném vzestupu, přesto naprostá většina simulací je stále prováděna pomocí klasických silových polí. Výsledky takovýchto simulací jsou však pouze tak dobré jako odpovídající potenciály. Silová pole jsou nastavována převážně pomocí výsledků náročnějších metod kvantové mechaniky či pomocí experimentálních dat, to se nemění ani s nástupem metod strojového učení, nicméně to umožňuje konstrukci mnohem flexibilnějších silových polí, kde nejsou zahrnuty žádné fyzikální aproximace (vyjma těch zahrnutých v referenční kvantově chemické metodě)[63]. V principu jsme tak schopni získat libovolně přesné (vzhledem k referenční kvantové metodě) silové pole ve formě modelu strojového učení (např. neuronové sítě). Chmiela a spol. byli takto schopní provádět dynamické simulace pro molekuly až o několika desítkách atomů s přesností na úrovni CCSD(T) metody[64]. I přes tyto úspěchy je však konstrukce přesných silových polí stále velmi náročná a potenciály musí být pečlivě testovány vzhledem k absenci jejich fyzikálního předpisu.

5. Závěr a výhled

Ačkoliv mnoho teoretických chemiků začínalo svou cestu k chemii pod vlivem pokusů se sodíkem házeným do vody či černým střelným prachem, většina z nich bude asi přesvědčena o transformativní roli výpočetní techniky v chemii. Věrohodná předpověď struktury a reaktivit molekul je dávným snem teoretické chemie a naděje zde byly spojovány zejména s kvantovou chemií. Přes dechberoucí pokrok je třeba přiznat, že metody teoretické chemie v tuto chvíli chemii nedominují, byť se postupně stávají téměř povinnou součástí chemického výzkumu. Začíná být ale představitelné, že dávný sen teoretické chemie bude naplněn statistickými metodami nebo alespoň s jejich pomocí namísto deterministického přístupu kvantové teorie molekul.

Strojové učení se dnes běžně používá v kvantové chemii, farmakologii, materiálovém inženýrství i dalších odvětvích. Umělá inteligence podstatně zasahuje i do oblastí, kde zatím počítače měly spíše pomocnou roli. Již od 60. let se intenzivně rozvíjí koncept počítačového návrhu syntéz. Výzkum v této oblasti vykazoval střídavé úspěchy[65,66], nyní jsou ale techniky umělé inteligence schopny dosáhnout v návrhu retrosyntézy výsledků srovnatelných s graduovanými chemiky[67,68]. Souběžně s tím se začíná prosazovat i robotizace chemických syntéz[69]. I syntetická chemie tak možná již zanedlouho zcela změní svou tvář[68,70]. Snad tedy není daleko doba, kdy si chemik zasedne k terminálu Syntetizátoru 4.0™ a zadá následující úkol: „Podle známých sloučenin vázajících se na melatoninový receptor navrhni syntetickou přípravu nových účinných látek zlepšujících spánek a prostupujících kůží tak, aby se daly nanést ve formě krému a měly minimum vedlejších účinků, a následně je nasyntetizuj z dostupných levných surovin.“ Je také možné, že zmíněná syntéza bude fungovat jako cloudová služba, kterou si bude moci kdokoliv pronajmout, stejně jako si již dnes můžeme pronajmout např. diskový prostor či výpočetní výkon.

Štěpán Sršeň je podporován z účelové podpory na specifický vysokoškolský výzkum (MŠMT č.21-SVV/2018) a je součástí International Max Planck Research School for „Many Particle Systems in Structured Environments.“ (IMPRS-MPSSE). Karel Berka je podporován GAČR 17-21122S.

LITERATURA

1. https://trends.google.com/trends/explore?date=all&q=nanotechnology,machine%20learning, staženo 16.7.2018.
2. Barták R.: *Co je nového v umělé inteligenci?* Nová beseda, Praha 2017.
3. Russakovsky O. a 11 spoluautorů: Int. J. Comput. Vis. *115*, 211 (2015).
4. Silver D. a 19 spoluautorů: Nature *529*, 484 (2016).
5. Chen H., Engkvist O., Wang Y., Olivecrona M., Blaschke T.: Drug Discovery Today *23*, 1241 (2018).
6. Arús-Pous J., Probst D., Reymond J. L.: Chimia (Aarau) *72*, 70 (2018).
7. Zahradník R.: *Laboratorní deník*. Academia, Praha 2008.
8. Rajaraman V.: Resonance *19*, 198 (2014).
9. https://en.oxforddictionaries.com/definition/artificial_intelligence , staženo 18.7.2018.
10. Ramakrishnan R., Hartmann M., Tapavicza E., von Lilienfeld O. A.: J. Chem. Phys. *143*, 084111 (2015).
11. Hansen K., Montavon G., Biegler F., Fazli S., Rupp M., Scheffler M., von Lilienfeld O. A., Tkatchenko A., Müller K.-R.: J. Chem. Theory Comput. *9*, 3404 (2013).
12. Schölkopf B. a Smola A. J.: *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press, 2002.
13. McCulloch W. S., Pitts W.: Bull. Math. Biophys. *5*, 115 (1943).
14. Rosenblatt M: Ann. Math. Stat. *27*, 832 (1956).
15. Minsky M., Papert S.: *Perceptrons: an introduction to computational geometry*. Mit Press, 1972.
16. Werbos P. J.: Proc. IEEE *78*, 1550 (1990).
17. Svozil D., Bartonek P.: Chem. Listy *111*, 709 (2017).
18. Novotný J., Svozil D.: Chem. Listy *111*, 716 (2017).
19. Lin H. W., Tegmark M., Rolnick D.: J. Stat. Phys. *168*, 1223 (2017).
20. Goh G. B., Hodas N. O., Vishnu A.: J. Comput. Chem. *38*, 1291 (2017).
21. Čmelo I., Svozil D.: Chem. Listy *111*, 724 (2017).
22. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J.: Adv. Drug Deliv. Rev. *46*, 3 (2001).

23. Conselice C. J., Wilkinson A., Duncan K., Mortlock A.: Astrophys. J. *830*, 83 (2016).
24. Drachman D. A.: Neurology *64*, 2004 (2005).
25. Overton E.: Ges. Zuerich *44*, 88 (1899).
26. Overton E.: *Studium uber die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie*. G. Fischer, Jena 1901.
27. Meyer H.: Arch. Experim. Pathol. Pharmacol. *42*,109 (1899).
28. Zahradník R., Chvapil M.: Experientia *16*, 511 (1960).
29. Zahradník R.: Experientia *18*, 534 (1962).
30. Hammett L. P.: J. Am. Chem. Soc. *59*, 96 (1937).
31. Taft R. W.: J. Am. Chem. Soc., *75*, 4231 (1953).
32. Koutecky J., Zahradnik R.: Cancer Res. *21*, 457 (1961).
33. Hansch C., Maloney P. P., Fujita T., Muir R. M.: Nature *194*, 178 (1962).
34. Hansch C., Fujita T.: J. Am. Chem. Soc. *86*, 1616 (1964).
35. Škuta C., Svozil D.: Chem. Listy *111*, 747 (2017).
36. Polishchuk P.: J. Chem. Inf. Model *57*, 2618 (2017).
37. Potts R. O, Guy R. H.: Pharm. Res. *9*, 663 (1992).
38. Baba H., Takahara J., Yamashita F., Hashida M.: Pharm. Res. *32*, 3604 (2015).
39. Hewitt M., Cronin M. T. D, Enoch S. J., Madden J. C., Roberts D. W., Dearden J. C.: J. Chem. Inf. Model *49*, 2572 (2009).
40. Hughes L. D., Palmer D. S., Nigsch F., Mitchell J. B. O.: J. Chem. Inf. Model *48*, 220 (2008).
41. Zhang L., Tan J., Han D., Zhu H.: Drug Discovery Today *22*, 1680 (2017).
42. Mitchell J. B. O.: WIREs Comput. Mol. Sci. *4*, 468 (2014).
43. Wu Z., Ramsundar B., Feinberg E. N., Gomes J., Geniesse C., Pappu A. S., Leswing K., Pande V.: Chem. Sci. *9*, 513 (2018).
44. Huang B., von Lilienfeld O. A.: J. Chem. Phys. *145*, 161102 (2016).
45. Kopp H.: Dublin Philos. Mag. J. Sci. *20*, 187 (1842).
46. Benson S. W.: *Thermochemical kinetics: Methods for the estimation of thermochemical data and rate parameters*. Wiley, New York 1976.
47. Ramakrishnan R., Dral P. O., Rupp M., von Lilienfeld O. A.: *J. Chem. Theory Comput. 11*, 2087 (2015)
48. Carleo G., Troyer M.: Science *355*, 602 (2017).
49. Eickenberg M., Exarchakis G., Hirn M., Mallat S., Thiry L.: J. Chem. Phys. *148*, 241732 (2018).
50. Yang G., Wu J., Chen S., Zhou W., Sun J., Chen G.: J. Chem. Phys. *148*, 241738 (2018).
51. Faber F. A., Christensen A. S., Huang B., von Lilienfeld O. A.: J. Chem. Phys. *148*, 241717 (2018).
52. Gubaev K., Podryabinkin E. V., Shapeev A. V.: J. Chem. Phys. *148*, 241727 (2018).
53. Gastegger M., Behler J., Marquetand P.: Chem. Sci. *8*, 6924 (2017).
54. Sršeň Š., Hollas D., Slavíček P.: Phys. Chem. Chem. Phys. *20*, 6421 (2018).
55. Lusci A., Pollastri G., Baldi P.: J. Chem. Inf. Model. *53*, 1563 (2013).
56. Maddox J.: Nature *335*, 201 (1988).
57. Li X., Curtis F. S., Rose T., Schober C., Vazquez-Mayagoitia A., Reuter K., Oberhofer H., Marom N.: J. Chem. Phys. *148*, 241701 (2018).
58. Graser J., Kauwe S. K., Sparks T. D.: Chem. Mater. *30*, 3601 (2018).
59. Jørgensen P. B., Mesta M., Shil S., García Lastra J. M., Jacobsen K. W., Thygesen K. S., Schmidt M. N.: J. Chem. Phys. *148*, 241735 (2018).
60. Afzal M. A. F., Cheng C., Hachmann J.: J. Chem. Phys. *148*, 241712 (2018).
61. Pilania G., McClellan K. J., Stanek C. R., Uberuaga B. P. J. Chem. Phys. *148*, 241729 (2018).
62. Elton D. C., Boukouvalas Z., Butrico M. S., Fuge M. D., Chung P. W.: Sci. Rep. *8*, 9059 (2018).
63. Behler J.: J. Chem. Phys. *145*, 170901 (2016).
64. Chmiela S., Sauceda H. E., Müller K.-R., Tkatchenko A.: arXiv:1802.09238 (2018).
65. Corey E. J., Wipke W. T.: Science *166*, 178 (1969).
66. Corey E. J., Long A. K., Rubenstein S. D.: Science *228*, 408 (1985).
67. Klucznik T. a 19 spoluautorů: Chem *4*, 522 (2018).
68. Segler M. H. S., Preuss M., Waller M. P.: Nature *555*, 604 (2018).
69. Granda J. M., Donina L., Dragone V., Long D.-L., Cronin L.: Nature *559*, 377 (2018).
70. Aspuru-Guzik A., Lindh R., Reiher M.: ACS Cent. Sci. *4*, 144 (2018).

K. Berka[a], Š. Sršeň[b], and P. Slavíček[b,c]
([a] Department of Physical Chemistry, RCPTM, Faculty of Science, Palacký University Olomouc, Olomouc, [b] Department of Physical Chemistry, University of Chemistry and Technology, Prague, [c] J. Heyrovský Institute of Physical Chemistry of the CAS, Prague): **Is Machine Learning the Future of Theoretical Chemistry?**

The application of the methods of machine learning in chemistry is briefly summarized in the present work. We first explain the basic concepts of artificial intelligence and machine learning. Next, the applications in two particular areas are discussed: searching relations between the structure and biological activity of molecules and using the techniques of machine learning in quantum chemistry as well as in other fields of theoretical chemistry. The evolutionary character of the machine learning approaches is emphasized. A fast development is witnessed in the field which, however, gradually follows the previous development in using statistical techniques in chemistry.

Keywords: machine learning, artificial intelligence, QSAR, quantum chemistry, theoretical chemistry, neural networks