QSAR – MODELOVÁNÍ KVANTITATIVNÍCH VZTAHŮ MEZI STRUKTUROU A AKTIVITOU CHEMICKÝCH LÁTEK

CTIBOR ŠKUTA[a] a DANIEL SVOZIL[a,b]

[a] CZ-OPENSCREEN: Národní infrastruktura pro chemickou biologii, Ústav molekulární genetiky AV ČR v.v.i. Vídeňská 1083, 142 20 Praha 4, [b] CZ-OPENSCREEN: Národní infrastruktura pro chemickou biologii, Laboratoř informatiky a chemie, Fakulta chemické technologie, Vysoká škola chemicko-technologická v Praze, Technická 5, 166 28 Praha 6
ctibor.skuta@img.cas.cz, daniel.svozil@img.cas.cz

Došlo 24.7.17, přijato 15.10.17.

Klíčová slova: QSAR, modelování biologické aktivity, vytěžování znalostí z dat, virtuální screening, oblast použitelnosti, konformní predikce

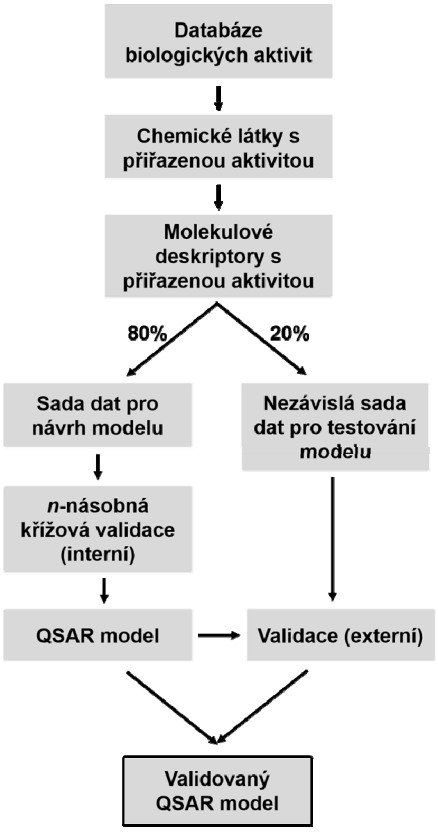
Obsah

1. Úvod
2. Data v QSAR modelování
   2.1. Zdroje dat
   2.2. Molekulové deskriptory
3. Metody strojového učení v QSAR
4. QSAR modelování
4.1 Oblast použitelnosti modelu
5. Použití QSAR modelu
6. Závěr

1. Úvod

V posledních letech dochází k velkému nárůstu množství volně dostupných chemicko-biologických dat. Tato data mají, ve spojení s řadou volně použitelných nástrojů pro jejich analýzu a obecně dostupnou výpočetní kapacitou, velký vliv na vývoj na poli modelování biologických vlastností chemických látek[1]. Jednou z nejvíce zkoumaných vlastností je biologická aktivita látky na konkrétním biologickém cíli (např. receptoru, enzymu aj.). Ta nám pomáhá nacházet potenciálně úspěšné kandidáty na nová léčiva, předpovídat vedlejší účinky látek nebo jejich potenciální toxicitu. Fyzikálně-chemické i biologické vlastnosti látek vychází z jejich struktury a na jejím základě mohou být tyto vlastnosti modelovány[2–9]. Modelování kvantitativních vztahů mezi strukturou a aktivitou (angl. Quantitative Structure-Activity Relationship, dále jen QSAR) může být popsáno jako využití statistických metod a metod pro analýzu dat za účelem výstavby predikčních modelů, které jsou schopny přesně předpovídat biologickou aktivitu resp. vlastnost látky (angl. Quantitative Structure-Property Relationship, QSPR) na základě její molekulární struktury. Každá QSAR metoda může být jednoduše zapsána ve formě $y = f(x)$, kde y je biologická aktivita (nebo vlastnost chemické látky), x sada vypočítaných nebo naměřených molekulárních deskriptorů (viz článek v tomto čísle Chemických listů[10]) a f empiricky stanovená matematická funkce. Cílem QSAR modelování je tedy vyhledání trendu/ů v sadě deskriptorů, které korelují s trendem v biologické aktivitě látky. Tento proces, stejně jako všechny ostatní metody virtuálního screeningu, vychází ze základního cheminformatického konceptu, že „podobné molekuly mívají podobné vlastnosti“ (angl. molecular similarity principle)[11], přeneseně v tomto případě, že „podobné molekuly jsou schopny působit na podobné biologické cíle“[12].

Postup pro vytváření QSAR modelů (obr. 1) ve své podstatě odpovídá postupům používaným při běžných



Obr. 1. **Základní schéma vytvoření a validace QSAR modelu**

modelovacích úlohách v rámci vytěžování znalostí z dat (angl. data mining) a skládá se z několika základních kroků: (1) Výběr a zpracování dat; (2) Výběr a výpočet molekulárních deskriptorů; (3) Trénování modelu; (4) Validace modelu; (5) Určení oblasti použitelnosti modelu.

Přestože se v tomto článku budeme zabývat primárně modelováním biologické aktivity (QSAR), je většina konceptů aplikovatelná také pro modelování fyzikálně-chemických vlastností (QSPR) látek[2–9, 13–18]. Modelování aktivity je ze své podstaty složitějším problémem, neboť závisí na dvou entitách, ligandu a jeho často velmi komplexním biologickém cíli, zatímco fyzikálně-chemické vlastnosti jsou závislé primárně na chemické struktuře látky[2,5].

2. Data v QSAR modelování

Data jsou obecně nejdůležitější součástí jakékoli modelovací úlohy a QSAR není výjimkou. Mnohé studie poukazují, že největší vliv na kvalitu QSAR modelu má právě množství a kvalita dostupných biologicko-chemických dat a použitých molekulárních deskriptorů, a že aplikovaná metoda strojového učení sice hraje svoji roli, ale v zásadě nerozhoduje o celkovém úspěchu či neúspěchu[19,20]. Jak ostatně říká známý princip „odpad na vstupu, odpad i na výstupu“ (angl. garbage in, garbage out), žádná sebelepší metoda nedokáže dosáhnout dobrých výsledků, když nemá k dispozici potřebná kvalitní data.

2.1. Zdroje dat

Pomineme-li vlastní (lokální) experimentální data a nespočet menších již předpřipravených množin dat, které se dají použít k samotnému modelování nebo jako benchmark pro již natrénované modely[21–23], tak za hlavní datové zdroje pro QSAR považujeme velké, volně přístupné databáze biologických aktivit jako jsou ChEMBL[24], PubChem BioAssay[25] nebo BindingDB[26]. Tyto databáze v současné době obsahují miliony chemických struktur a přiřazených, experimentálně naměřených biologických aktivit pro tisíce různorodých biologických cílů, od proteinů přes buněčné linie až po organismy.

Při práci s experimentálními daty z veřejných zdrojů jsme vždy odkázáni na jejich poskytovatele a musíme dávat pozor na chyby a variabilitu/nesrovnalosti, které se v nich mohou vyskytovat. Ty se mohou vyskytovat jak na straně chemické struktury, tak na straně biologické aktivity.

Studie z roku 2005 poukazuje na fakt, že v každém článku z oboru medicinální chemie jsou v průměru dvě chyby ve strukturách chemických látek, které se následně promítají i do chemických databází[27]. To potvrdila studie z roku 2008, která zjistila, že se chybovost ve struktuře v rámci těchto databází pohybuje mezi 0,1 až 3,4 % (cit.[28]). Uvážíme-li počet dostupných struktur, nejedná se o zanedbatelná čísla.

Především v případě, kdy strukturní data nepochází z jednoho zdroje, je nutné provádět jejich normalizaci, standardizaci a filtraci[29]. V průběhu zpracování by mělo docházet ke třem základním krokům, které by měly minimalizovat problémy ze strukturního hlediska:

1. Odstranění problematických typů látek (anorganické a organometalické látky, směsi) z důvodu omezených možností jejich zpracování pomocí cheminformatických nástrojů. Primárně pracujeme s organickými molekulami bez koordinačně-kovalentních vazeb, kde je jasně určena účinná látka.
2. Standardizace struktury (odstranění solných iontů a rozpouštědel, pokud možno neutralizace, odstranění stereochemie, sjednocení tautomerů). Různé druhy solí by neměly mít vliv na rozdílnou biologickou aktivitu látky. Naproti tomu mohou odlišné stereoisomery vykazovat odlišnou biologickou aktivitu, avšak při použití nejběžnějších 2D deskriptorů se tato informace o stereochemii do popisu molekuly nepromítá.
3. Odstranění/sloučení duplicit. Různé cheminformatické nástroje mohou identickou strukturu reprezentovat odlišnými zápisy (podle implementace konkrétního algoritmu). Tomuto jevu lze předcházet použitím strukturního formátu InChI (resp. InChIKey, jeho hašované podoby)[30], který by měl dávat identické výsledky bez ohledu na použitý software. U staršího, stále hojně používaného formátu SMILES[31,32] tato podmínka zajištěna není.

Vzhledem k množství dat není ve většině případů možné provádět kontrolu a zpracování dat manuálně, můžeme ale použít některé z dostupných cheminformatických nástrojů jako např. programovací knihovny RDkit[33] a OpenBabel[34], volně dostupnou analytickou platformu KNIME s rozšířením pro cheminformatiku[35] či komerční nástroje Standardizer (ChemAxon)[36] a Pipeline Pilot (Accelrys)[37].

Modelovaná biologická aktivita bývá nejčastěji vyjádřena ve formě tzv. potence vyjádřené jako koncentrace látky potřebné k dosažení určitého biologického efektu. Mezi nejběžnější modelované veličiny patří inhibiční, aktivační nebo obecně disociační konstanty (např. *IC50* – koncentrace potřebná k dosažení poloviny maximální inhibice cíle danou látkou, *EC50* – koncentrace potřebná k dosažení poloviny maximální aktivace cíle danou látkou, K_d (K_i) – disociační konstanta ligandu, cíle a jejich komplexu). Tyto koncentrace se obecně udávají ve formě záporného dekadického logaritmu molární koncentrace (hodnota = –log(molární koncentrace[mol dm^{-3}], kde jednotka mol dm^{-3} se též označuje jako M) z důvodu potřeby normální distribuce chyby v rámci modelování. Na této škále se používané veličiny zpravidla uvozují malým písmenem *p* před jejich názvem (např. *pIC50* = –log(*IC50*), tzn., že *IC50* hodnotě 0,000 001 mol dm^{-3} (tedy 1 μmol dm^{-3}) odpovídá bezrozměrná *pIC50* hodnota 6). V rámci klasifikace látek se nejčastěji používá binární logika (1/0), kde rozřazujeme látky např. na aktivní (1) a neaktivní (0). Za neaktivní v tomto případě považujeme látky, které na konkrétním biologickém cíli nevykazují žádný biologický účinek, nebo i látky, u kterých je k dosažení účinku potřeba použití vyšších koncentrací. Za

dostatečně potentní (aktivní) se často považují látky s dosaženým účinkem při použití 10 nM koncentrace nebo nižší (na záporné logaritmické škále tedy číslo 7 a vyšší).

Vzhledem k tomu, že experimentální měření aktivity látek je ovlivňováno velkou řadou faktorů (typ experimentu, použité látky a jejich čistota, laboratorní podmínky, lidský faktor aj.), hodnoty naměřené v různých experimentech/laboratořích se často mohou i řádově lišit[38]. Stává se tedy, že pro identickou kombinaci ligandu a biologického cíle máme k dispozici řadu hodnot s větším či menším rozptylem. V takovém případě je potřeba dávat pozor a v případě příliš velkého rozptylu data raději nepoužít, aby nesnižovala přesnost modelu.

2.2. Molekulové deskriptory

Pro popis látek v rámci QSAR modelování je možné použít libovolný druh molekulárních deskriptorů (datových vektorů) popisujících strukturní vlastnosti látek: od fyzikálně-chemických vlastností, přes popis 2D topologie, až po data popisující 3D strukturu včetně její dynamiky[3–5,39,40]. Pro modelování biologické aktivity jsou v současné době patrně nejpoužívanější binární 2D topologické deskriptory, tzv. otisky prstů (angl. fingerprints)[41]. Ty sice neposkytují takovou úroveň detailu jako popis struktury ve 3D, na druhou stranu jejich výpočet není zatížen složitými výpočty konformací molekuly v prostoru, které se mohou podle použitých nástrojů a jejich nastavení lišit. Ty navíc ve výsledku nemusí vystihovat správnou konformaci potřebnou pro dosažení biologického účinku. Naproti tomu je výpočet 2D deskriptorů zpravidla velmi rychlý a jeho výsledek vždy stejný. To může být v době, kdy máme k dispozici miliony dostupných struktur, jedním z rozhodujících kritérií. Obecně platí, že výběr molekulových deskriptorů by měl být závislý na modelované veličině a od ní odvozené složitosti problému, který řešíme. Problematice molekulových deskriptorů se podrobněji věnuje článek v tomto čísle Chemických listů[10].

3. Metody strojového učení v QSAR

QSAR modelování se liší od jiných typů modelování primárně použitými daty a je pro něj tedy možné použít prakticky libovolnou metodu strojového učení. Vzhledem k tomu, že se jedná o problém poměrně komplexní, jsou v dnešní době oblíbené hlavně sofisitkovanější metody jako neuronové sítě (angl. neural networks, NN)[42], metoda podpůrných vektorů (angl. support vector machines, SVM, společně se support vector regression, SVR)[43,44], náhodné lesy (angl. random forest, RF)[45], ale i mnohé další[46,47]. Všechny zmíněné metody jde použít jak pro klasifikační, tak pro regresní úlohy, bez toho, že by některá z nich výkonnostně vyčnívala nad ostatními. Zatímco NN mohou častěji trpět přeučením (angl. overfitting) pro konkrétní data, SVM a RF jsou v tomto ohledu považovány za poměrně robustní[48]. Především RF jsou v dnešní době oblíbeny pro svou výpočetní nenáročnost a z podstaty také pro jejich možnou interpretovatelnost (RF je množina rozhodovacích stromů)[49].

Vzhledem k tomu, že různé metody strojového učení v kombinaci s různými molekulovými deskriptory mohou zachycovat rozdílné aspekty modelované množiny dat, je dnes velmi populární tzv. konsenzuální modelování. V konsenzuálním přístupu můžeme vytvořit řadu odlišných modelů a na základě všech jejich výsledků spočítat výsledek finální. Konsenzuální přístup v QSAR modelování zpravidla vykazuje větší úspěšnost než samostatně použité modely[29,50].

Z hlediska softwarových nástrojů je pro modelování k dispozici řada knihoven pro většinu běžně používaných programovacích jazyků, např. scikit-learn pro programovací jazyk Python[51], Weka pro programovací jazyk Java[52] nebo nástroje od Bioconductor pro programovací prostředí R[53]. Uživatelsky přívětivá a v rámci modelování velmi dobře vybavená je také již zmíněná analytická platforma Knime[35].

4. QSAR modelování

V rámci QSAR modelování je vstupní datová množina (v našem případě množina chemických struktur s přiřazenou biologickou aktivitou) na počátku rozdělena na dvě nezávislé sady dat: základní sadu dat a nezávislou sadu dat. Základní sada dat se používá pro vnitřní nastavení parametrů modelu, tzv. trénování modelu, a proto se také označuje jako trénovací množina. Nezávislá sada dat pak slouží k testování kvality modelu a označuje se též jako testovací množina[3,54]. Nejčastěji se vstupní soubor dat rozděluje v poměru 70:30 nebo 80:20 ve prospěch trénovací množiny a toto rozdělení může být provedeno náhodným výběrem nebo tzv. stratifikovaně. Stratifikovaný výběr trénovací a testovací datové množiny by měl zaručovat rovnoměrné zastoupení (stejnou distribuci) závislé proměnné (aktivity) v obou množinách. V případě klasifikace stejné poměrné zastoupení kategorií aktivní/neaktivní, v případě regrese stejné rozložení a rozsah numerických hodnot.

V QSAR modelování se standardně používá dvojí typ validace modelu: interní a externí. Zatímco externí validace se provádí s již zmíněnou nezávislou testovací množinou pro finální validaci modelu, k interní validaci dochází v rámci samotného modelování s daty z trénovací množiny. Metoda nejčastěji využívaná pro interní validaci se nazývá *n*-násobná křížová validace. V rámci křížové validace se trénovací množina rozdělí na *n* stejně velkých podmnožin (např. $n = 10$) a postupně jednu množinu po druhé vynecháváme, model trénujeme na zbylých $n-1$ množinách a aktuálně vynechanou množinu používáme jako testovací množinu aktuálního modelu. Zároveň v každém cyklu zaznamenáváme vybrané charakteristiky modelu (např. chybu), které slouží jako část výsledku interní validace. U *n*-násobné křížové validace tak dochází k modelování *n* QSAR modelů. V limitním případě se *n*

může rovnat počtu prvků v množině (v takovém případě se jedná o tzv. „leave-one-out“ křížovou validaci). Finální QSAR model se nakonec trénuje na celé trénovací množině a je aplikován na počátku oddělenou nezávislou testovací množinu, čímž získáme výsledky externí validace.

Pro určení, zda model vyhovuje požadavkům pro aplikaci na nová data, můžeme použít řadu statistických ukazatelů, které zpravidla zachycují přesnost modelu a jeho úspěšnost v porovnání s náhodným scénářem.

Pro regresní modely je standardně používaným primárním kritériem kombinace křížově validovaného koeficientu determinace (q^2) (interní validace) a koeficientu determinace pro testovací set (R^2) (externí validace).

Koeficient determinace udává poměr čtvercové chyby modelu (součet čtverců odchylek) a čtvercové chyby při nahrazení modelu průměrnou hodnotou aktivity (obecně tedy míru přesnosti modelu oproti náhodnému scénáři). Můžeme ho vypočítat podle následující rovnice *(1)*:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \qquad (1)$$

kde N představuje velikost testovací množiny, y_i experimentální hodnotu, $\hat{y}_i$ předpovězenou hodnotu a $\bar{y}$ průměrnou hodnotu biologické aktivity pro všechny látky v testovací množině. Koeficient determinace nabývá hodnot mezi 0 a 1, kde 1 reprezentuje dokonale přesný model (nulovou chybu).

Nejběžnějšími kritérii používanými pro hodnocení klasifikačních modelů jsou tzv. plocha pod křivkou (angl. area under the curve, AUC) a faktor obohacení (angl. enrichment factor, EF). AUC vychází z tzv. ROC křivky (angl. receiver operating curve), která zachycuje počet skutečně pozitivních (angl. true positive rate, TPR) jako funkci falešně pozitivních (angl. false positive rate, TPR) klasifikovaných prvků (obr. 2). TPR na určité pozici vyjadřuje poměr určených aktivních látek a skutečně aktivních látek, FPR je poměr určených neaktivních látek a skutečných neaktivních látek. Z této křivky můžeme vypočítat AUC podle následující rovnice *(2)*:

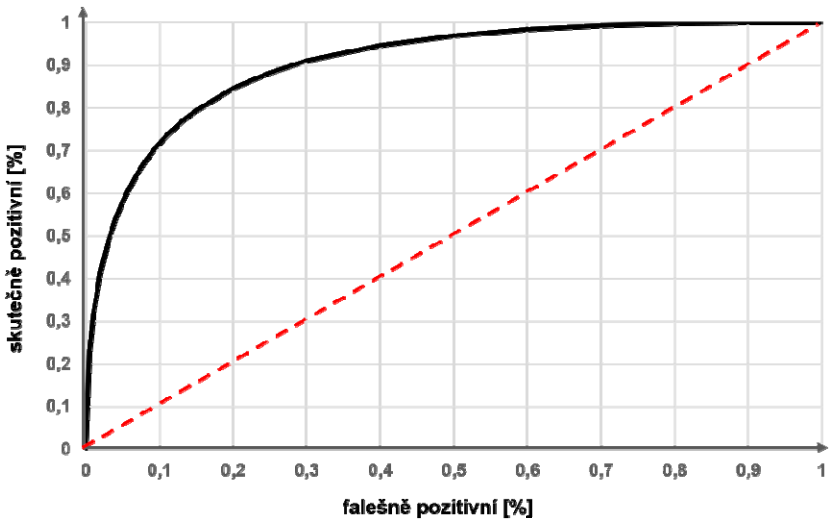
$$\mathrm{AUC} = \frac{1}{nN}\sum_{i=2}^{N} A_i\left(I_i - I_{i-1}\right) \qquad (2)$$

kde n je počet aktivních látek a N je celkový počet látek v testovacím setu, A je kumulovaný počet aktivních látek na pozici i a I kumulovaný počet neaktivních látek na pozici i. AUC nabývá hodnot mezi 0 a 1, kde hodnota 1 odpovídá nulovému FPR a hodnota 0,5 náhodnému scénáři.

EF se používá jako ukazatel schopnosti modelu časně rozpoznat (angl. early recognition) aktivní látky. Časné rozpoznání poukazuje na schopnost modelu nejen rozpoznat aktivní látky, ale udělat tak s vysokou jistotou a zařadit aktivní látku v pomyslném žebříčku hodnocených látek na co nejvyšší místo. EF pro určenou část klasifikované množiny (např. 5 % řazené od nejvyšší pravděpodobnosti, že látka je aktivní) udává poměr počtu nalezených aktivních látek oproti počtu aktivních látek, které by se v určené části množiny vyskytovaly při náhodném výběru. Výpočet EF probíhá podle následující rovnice *(3)*:

$$\mathrm{EF}(\chi) = \frac{\sum_{i=1}^{n}\delta(r_i)}{\chi^n}, s\ \delta(r_i) = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases} \qquad (3)$$

kde χ je velikost části testované množiny, n je počet aktivních látek, N je celkový počet látek v testovacím setu a r_i určuje pořadí i-té aktivní látky. Rozsah EF se pohybuje mezi 0 a hodnotou závislou na použitém χ a počtu aktivních látek v testované množině: $1/\chi$, pokud je $\chi \geq n/N$, v opačném případě N/n.



Obr. 2. **Ukázka ROC křivky zachycující úspěšnost klasifikačního modelu.** Výpočtem obsahu plochy pod touto křivkou získáme hodnotu AUC. Diagonála odpovídá náhodnému scénáři

Na závěr této části je potřeba si připomenout, že výborné charakteristiky modelu nikdy neznamenají, že je možné jej použít pro všechna možná vstupní data bez omezení, ale jen na ta, která je model schopen předpovědět s určitou přesností, tj. ta, která spadají do oblasti použitelnosti modelu.

4.1. Oblast použitelnosti modelu

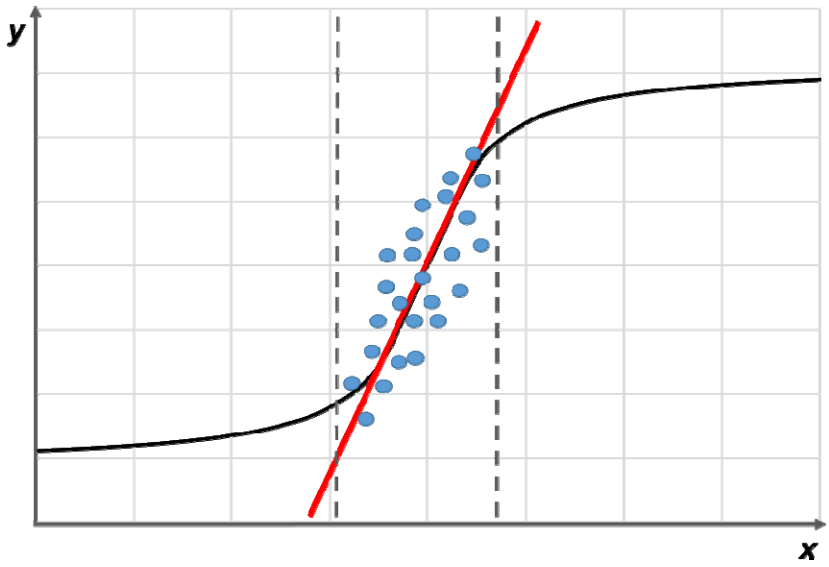
Prostor vstupních dat (vektorů), pro která je bezpečné model použít, se nazývá oblastí použitelnosti modelu (angl. applicability domain, AD)[13,55]. Stejně tak jako by např. model pro predikci druhů ovoce natrénovaný pouze na datech pro jablka nešel použít pro predikci banánů (resp. šel, ale výsledek by byl logicky špatný), tak ani QSAR modely by neměly být používány pro predikci vlastností látek, které se výrazně odlišují od látek v trénovací množině (obr. 3). Před použitím modelu je tedy nutné předem definovat možný prostor vstupních dat, na která je možné ho bezpečně aplikovat. Hledisek, která lze uplatňovat pro určení tohoto prostoru, je celá řada, primárně by se však mělo vycházet ze vstupních dat, resp. deskriptorů použitého QSAR modelu.

V případě použití fyzikálně-chemických vlastností jako vstupních dat by se tyto vlastnosti pro predikované látky neměly výrazně lišit, v ideálním případě by měly ležet v intervalu mezi maximální a minimální hodnotou vstupních dat. Při použití topologických deskriptorů by podobnost predikované látky k minimálně jedné látce z trénovací množiny měla dosáhnout určité stanovené prahové hodnoty (záleží na použitých deskriptorech a míře podobnosti). Často dochází k použití kombinace více kritérií.

Jednou z oblíbených technik postihujících oblast použitelnosti QSAR modelů je tzv. konformní predikce (angl. conformal prediction)[56]. Při použití této metody je na základě chyby naměřené při aplikaci QSAR modelu na trénovací množinu vypočítán přidružený chybový model. Na jeho základě je možné pro určenou míru pravděpodobnosti predikovat konfidenční interval udávající rozsah hodnoty predikované QSAR modelem. Modelovým výsledkem této metody tedy může být např. informace, že s 90% pravděpodobností je aktivita látky na definovaném biologickém cíli rovna hodnotě 7,0±1,3. Čím větší pravděpodobnost zvolíme, tím širší je obdržený konfidenční interval.

5. Použití QSAR modelu

Při používání QSAR modelů (resp. obecně jakéhokoli modelu) je třeba stále mít na paměti, že i přes potenciálně výborné statistiky a robustně definovanou oblast použitelnosti jsou získané výsledky vždy jen predikcí skutečných hodnot a dokud nejsou experimentálně potvrzené, nemůžeme je považovat za jisté. QSAR modelování vychází z podobnostního principu, často však mohou dvě velmi podobné látky vykazovat diametrálně odlišné vlastnosti**.** Důkazem mohou být rozdílné fyzikálně-chemické vlastnosti či biologické aktivity např. *cis*- a *trans*- isomerů nebo optických antipodů některých látek[12,57]. Z této skutečnosti vychází např. tzv. koncept aktivitních útesů (angl. activity cliffs) v aktivitní krajině (angl. activity landscape) (cit.[58,59]). Když si představíme množinu chemických látek a jejich aktivit jako krajinu v *n*-dimenzionálním prostoru (podle deskriptorů použitých k jejich popisu), nenajdeme v ní nejen roviny s mírnými kopci (pozvolně měnící se aktivitu společně se strukturou), ale také útesy (oblasti velkých změn), kde se malá změna ve struktuře látky signifikantně projeví na její biologické aktivitě.



Obr. 3. **Ukázka důležitosti definice oblasti použitelnosti QSAR modelu (oblast mezi přerušovanými čarami).** V tomto případě bude při jejím překročení chyba predikce narůstat se vzdáleností od trénovací množiny dat

QSAR modely se, stejně jako další metody virtuálního screeningu, primárně používají pro selekci vhodných kandidátů v počátečních fázích testování velkých knihoven chemických látek, kde snižují nejen potřebný čas, ale hlavně výslednou cenu experimentu. Obecnou výhodou modelů je fakt, že látky, na které model aplikujeme, nemusíme mít ve skutečnosti k dispozici nebo dokonce nemusí ani reálně existovat. QSAR model v tomto případě může poskytnout důležitou informaci při rozhodování, zda se látku máme vůbec pokusit nasyntetizovat.

Vzhledem k současnému množství nástrojů i dat pro QSAR modelování a silnému hlasu proti provádění testů na zvířatech jsou v dnešní době *in-silico* metody testování také součástí evropské legislativy pro regulaci chemických látek REACH (angl. Registartion, Evaluation, Authorisation and Restriction of Chemicals). Dedikovaná aplikace OECD QSAR toolbox je především nástrojem sdružujícím data a modely, které pomáhají predikovat ekotoxická rizika látek a jejich potenciální mechanismus účinku[60].

6. Závěr

QSAR modelování je jedním z hlavních zástupců metod virtuálního screeningu používaných pro predikci biologických či fyzikálně-chemických (QSPR) vlastností látek. S využitím libovolných molekulárních deskriptorů a metod strojového učení dokáže produkovat regresní a klasifikační modely, které jsou nejen velmi rychlé, ale v kombinaci s odpovědným přístupem k jejich validaci a dobře definovanou oblastí použitelnosti také velmi robustní a přesné. QSAR modely nacházejí pro tyto své vlastnosti uplatnění v rámci počítačového návrhu léčiv, objasňování principů funkce chemických látek v živých organismech a predikci jejich ekotoxické bezpečnosti. Roli v posuzování ekotoxických rizik hrají také v rámci evropské legislativy pro regulaci chemických látek (REACH), kde mohou být *in-silico* alternativou ke klasickým experimentům.

Tento článek vznikl za podpory MŠMT v rámci Národního programu udržitelnosti I projekt LO1220 (CZ-OPENSCREEN).

LITERATURA

1. Cherkasov A. + 19 spoluautorů: J. Med. Chem. *57*, 4977 (2014).
2. Kunal R., Supratik K., Das R. N.: *A primer on QSAR/ QSPR Modeling*. Springer, Cham 2015.
3. Poling B. E., Prausnitz J. M., O´Connell J. P.: *The Properties of Gases and Liquids, fifth edition*. McGraw-Hill, New York 2001.
4. Růžička V., Šobr J., Novák J., Bureš M., Cibulka I., Růžička K., Matouš J.: *Odhadové metody pro fyzikálně-chemické vlastnosti tekutin. Aplikace v technologii a chemii životního prostředí.* VŠCHT Praha, 1996.
5. Kolská Z., Zábranský M., Randová A., v knize: *Thermodynamics - Fundamentals and Its Application in Science* (Morales-Rodriguez R., ed.) kap. 6. Rijeka, Croatia 2012.
6. Kolská Z.: Chem. Listy *98*, 328 (2004).
7. Rucki M., Tichý M.: Chem. Listy *103*, 100 (2009).
8. Chuchvalec P., Novák J. P.: Chem. Listy *101*, 989 (2007).
9. Kolská Z., Růžička V., Gani R.: Ind. Eng. Chem. Res. *44*, 8436 (2005).
10. Novotný J., Svozil D.: Chem. Listy *111*, 716 (2017).
11. Hendrickson J. B.: Science *252*, 1189 (1991).
12. Martin Y. C., Kofron J. L., Traphagen L. M.: J. Med. Chem. *45*, 4350 (2002).
13. Kolská Z., Kukal J., Zábranský M., Růžička V.: Ind. Eng. Chem. Res. *47*, 2075 (2008).
14. Randová A., Bartovská L., Hovorka Š., Poloncarzová M., Kolská Z., Izák P.: J. Appl. Polym. Sci. *111*, 1745 (2009).
15. Chickos J. S.; Wilson J. A.: J. Chem. Eng. Data. *42*, 190 (1997).
16. Li P., Ma P., Yi S., Zhao Z., Cong L.: Fluid Phase Equilib. *101*, 1994 (1994).
17. Constantinou L., Prickett S. E., Mavrovouniotis M. L.: Ind. Eng. Chem. Res. *32*, 1734 (1993).
18. Marrero J., Gani R.: Fluid Phase Equilib. *183*, 183 (2001).
19. Tetko I. V., Sushko I., Pandey A. K., Zhu H., Tropsha A., Papa E., Oberg T., Todeschini R., Fourches D., Varnek A.: J. Chem. Inf. Model. *48*, 1733 (2008).
20. Zhu H., Tropsha A., Fourches D., Varnek A., Papa E., Gramatica P., Oberg T., Dao P., Cherkasov A., Tetko I. V.: J. Chem. Inf. Model. *48*, 766 (2008).
21. Huang N., Shoichet B. K., Irwin J. J.: J. Med. Chem. *49*, 6789 (2006).
22. Rohrer S. G., Baumann K.: J. Chem. Inf. Model. *49*, 169 (2009).
23. Heikamp K., Bajorath J.: J. Chem. Inf. Model. *51*, 1831 (2011).
24. Gaulton A. + 17 spoluautorů: Nucleic Acids Res. *45*, D945 (2017).
25. Wang Y., Bryant S. H., Cheng T., Wang J., Gindulyte A., Shoemaker B. A., Thiessen P. A., He S., Zhang J.: Nucleic Acids Res. *45*, D955 (2017).
26. Gilson M. K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J.: Nucleic Acids Res. *44*, D1045 (2016).
27. Olah M. + 10 spoluautorů: Methods Princ. Med. Chem. *22*, 223 (2005).
28. Young D., Martin T., Venkatapathy R., Harten P.: QSAR Comb. Sci. *27*, 1337 (2008).
29. Tropsha A.: Mol. Inf. *29*, 476 (2010).
30. Heller S. R., McNaught A., Pletnev I., Stein S., Tchekhovskoi D.: J. Cheminf. *2015*, 7.
31. Weininger D.: J. Chem. Inf. Comput. Sci. *28*, 31 (1988).
32. Jirát J., Svozil D.: Chem. Listy *111*, 710 (2017).
33. *RDKit: Open-source cheminformatics*. 2006.
34. O'Boyle N. M., Banck M., James C. A., Morley C.,

Vandermeersch T., Hutchison G. R.: J. Cheminf. *3*, 33 (2011).
35. Berthold M. R., Cebron N., Dill F., Gabriel T. R., Kotter T., Meinl T., Ohl P., Sieb C., Thiel K., Wiswedel B.: Stud. Class. Data Anal. *2008*, 319.
36. Weber L.: Chem. World-Uk. *5*, 65 (2008).
37. Stevenson J. M., Mulready P. D.: J. Am. Chem. Soc. *125*, 1437 (2003).
38. Kalliokoski T., Kramer C., Vulpetti A., Gedeck P.: Plos One *2013*, 8.
39. Todeschini R. C. V.: *Handbook of molecular descriptors*. Wiley, Freiburg 2008.
40. Kolská Z., Petrus P.: Collect. Czech. Chem. Commun. *75*, 393 (2010).
41. Willett P.: Drug Discov. Today *11*, 1046 (2006).
42. Baskin I. I., Winkler D., Tetko I. V.: Expert Opin. Drug Discovery *11*, 785 (2016).
43. Alvarsson J., Lampa S., Schaal W., Andersson C., Wikberg J. E., Spjuth O.: J. Cheminf. *8*, 39 (2016).
44. Czerminski R., Yasri A., Hartsough D.: Quant. Struct-Act. Rel. *20*, 227 (2001).
45. Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P., Feuston B. P.: J. Chem. Inf. Comput. Sci. *43*, 1947 (2003).
46. Varnek A., Baskin I.: J. Chem. Inf. Model. *52*, 1413 (2012).
47. Lavecchia A.: Drug Discov. Today *20*, 318 (2015).
48. Mitchell J. B. O.: Wires Comput. Mol. Sci. *4*, 468 (2014).
49. Kuz'min V. E., Polishchuk P. G., Artemenko A. G., Andronati S. A.: Mol. Inf. *30*, 593 (2011).
50. Riniker S., Fechner N., Landrum G. A.: J. Chem. Inf. Model. *53*, 2829 (2013).
51. Pedregosa F., Varoquaux G., Gramfort A., et al.: J. Mach. Learn. Res. *12*, 2825 (2011).
52. Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I. H., Trigg L.: *Data Mining and Knowledge Discovery Handbook, Second Edition.* Springer, New York 2010.
53. Gentleman R. C. + 24 spoluautorů: Genome. Biol. *2004*, 5.
54. Walker J. D., Carlsen L., Jaworska J.: QSAR Comb. Sci. *22*, 346 (2003).
55. Netzeva T. I., Gallegos Saliner A., Worth A. P.: Environ. Toxicol. Chem. *25*, 1223 (2006).
56. Norinder U., Carlsson L., Boyer S., Eklund M.: J. Chem. Inf. Model. *54*, 1596 (2014).
57. von Nussbaum F. + 21 spoluautorů: ChemMedChem *10*, 1163 (2015).
58. Maggiora G. M.: J. Chem. Inf. Model. *46*, 1535 (2006).
59. Wassermann A. M., Wawer M., Bajorath J.: J. Med. Chem. *53*, 8209 (2010).
60. Dimitrov S. D. + 14 spoluautorů: SAR QSAR Environ. Res. *2016*, 1.

C. Škuta[a] and D. Svozil[a,b] *([a] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Academy of Sciences of the Czech Republic, Prague, [b] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague):* **QSAR – Modelling of Quantitative Relations between Structure and Activity of Chemical Compounds**

Quantitative structure–activity relationship (QSAR) modelling is one of the most popular techniques of virtual screening used to predict the activity of a compound toward a biological target. While QSAR classification models are able to predict whether a compound is active or inactive (class) toward a target, regression models try to predict its exact activity value. To find the relationship between the structure and activity of a compound, common machine learning methods are employed (e.g., Support Vector Machines, Random Forest, Neural Networks etc.) together with diverse types of compound descriptors (e.g., physico-chemical properties, structural keys, binary fingerprints etc.). QSAR models are generally very fast and, when a correct approach to their validation and applicability domain setting is used, also reliable. They became a common part of computational drug design workflows employed to detect new drug candidates, elucidate their side/adverse effects or assess their potential toxicity risks.