

## VIRTUÁLNÍ SCREENING

DANIEL SVOZIL<sup>a,b</sup>

<sup>a</sup> CZ-OPENSREEN: Národní infrastruktura pro chemickou biologii, Ústav molekulární genetiky AV ČR v.v.i., Vítězská 1083, 142 20 Praha 4, <sup>b</sup> CZ-OPENSREEN: Národní infrastruktura pro chemickou biologii, Laboratoř informatiky a chemie, Fakulta chemické technologie, Vysoká škola chemicko-technologická v Praze, Technická 5, 166 28 Praha 6  
Daniel.Svozil@vscht.cz

Došlo 4.8.17, přijato 10.10.17.

Klíčová slova: virtuální screening, podobnostní vyhledávání, fúze dat, farmakoforové modelování

### Obsah

1. Úvod
2. Virtuální screening
  - 2.1. Filtry ve virtuálním screeningu
  - 2.2. Podobnostní metody
  - 2.3. Farmakoforové modelování
3. Závěr

### 1. Úvod

5. října 1981 zveřejnil časopis Fortune titulní článek s názvem „The Next Industrial Revolution: designing drugs by computer at Merck“<sup>1</sup>, který je mnohými považován za počátek intenzivního zájmu o počítačový návrh léčiv (angl. Computer Aided Drug Design, CADD). Po téměř 40 letech vývoje metod CADD je však více než jasné, že CADD nemůže v procesu návrhu nového léčiva nikdy nahradit experimentální přístupy a slouží jako sice užitečný, leč stále pouze doplňkový nástroj. Hlavní experimentální technologií při hledání biologicky aktivních látek je testování látek s vysokou propustností (HTS, angl. High Throughput Screening)<sup>2</sup>, které je obšírněji diskutováno na jiném místě tohoto čísla Chemických listů<sup>3,4</sup>. Tento přístup má výhodu v tom, že nevyžaduje apriorní znalost mechanismu účinku a umožňuje testovat chemicky rozmanité sloučeniny. Nicméně tzv. „hit rate“, tj. počet nalezených biologicky aktivních látek vztahený na velikost celé testované chemické knihovny, je v případě HTS velmi nízký, což značným způsobem limituje jeho využití. A zde právě spočívá největší síla CADD, který se dnes používá především jako metoda umožňující významně snížit počet látek, jež je nutno experimentálně testovat (tzv. virtuální scree-

ning). Mnoho sloučenin, které jsou v CADD predikovány jako neaktivní, lze přeskočit, a ty, které jsou předpovězeny jako aktivní, je pak možno při experimentálním testování prioritizovat. Tím dojde nejen ke snížení nákladů na samotný HTS, ale je možno také paralelně s CADD vyvíjet a validovat konkrétní uspořádání HTS experimentu (angl. HTS assay), čímž dojde k významné časové úspoře. Sílu CADD je možno demonstrovat na příkladu hledání inhibitorů tyrosin fosfatasy-1B, enzymu hrajícího roli při onemocnění cukrovkou. Výzkumníci z firmy Pharmacia navrhli pomocí CADD knihovnu 365 potenciálních inhibitorů tyrosin fosfatasy-1B (cit.<sup>5</sup>). Při následném experimentálním ověření 127 těchto látek vykázalo skutečnou inhibiční aktivitu, což představuje hit rate téměř 35 %. Souběžně s tím byl proveden tradiční HTS na stejném molekulovém cíli. Ze 400 000 testovaných látek jich 81 vykázalo inhibiční účinek, hit rate byla tedy pouhých 0,021 %.

CADD poskytuje, ve srovnání s HTS a kombinatoriální organickou syntézou, výrazně vyšší hit rate, neboť vyhledává potenciálně aktivní látky cílenějším způsobem. Metody CADD však neslouží pouze k prioritizaci látek pro experimentální testování, ale jsou typicky používány v následujících třech scénářích:

1. **Virtuální screening.** Cílem je z velké chemické knihovny získat menší množinu pravděpodobně aktivních látek, které je možno následně testovat experimentálně.
2. **Optimalizace leadu.** Lead (angl. též lead compound) je jakýmsi předchůdcem léčiva, vykazuje biologickou aktivitu, ale nesplňuje všechny vlastnosti kladené na finální léčivo. Optimalizace leadu je proces, při kterém se upravuje struktura leadu za účelem zvýšení jeho biologické aktivity nebo vylepšení farmakokinetických a ADME/Tox (absorpce léčiva organismem, distribuce léčiva v organismu, metabolismus léčiva, vylučování léčiva a toxicita léčiva) vlastností. Výsledkem optimalizace leadu je kandidátní terapeutická látka.
3. **De novo design.** Návrh zcela nových struktur, tzv. „de novo design“, buď „pěstováním“ sloučenin v aktivním místě molekulového cíle, nebo spojováním větších fragmentů za účelem nalezení zcela nových chemotypů.

CADD metody lze klasifikovat do dvou hlavních kategorií. Zatímco ve strukturním (SBCADD, angl. structure-based CADD) přístupu máme k dispozici 3D strukturu molekulového cíle, ligandový CADD (LBCADD, angl. ligand-based CADD) využívá toliko informací o známých látkách aktivních či neaktivních na zkoumaném molekulovém cíli. SBCADD je upřednostňován v případě dostupnosti vysoce rozlišených struktur snadno krystalizovatelných či rozpustných proteinů. Hlav-

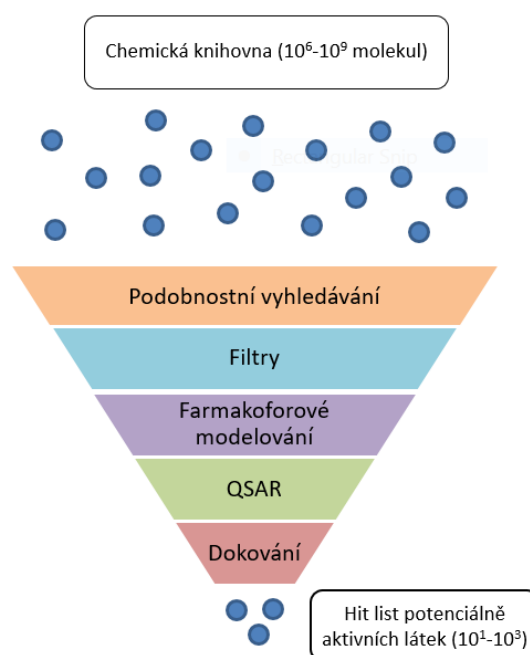
ním úkolem SBCADD, jehož typickým představitelem je molekulové dokování (angl. molecular docking), je navrhnout látky, které se váží silně na molekulární cíl, mají vylepšené ADME/Tox vlastnosti a jsou pro daný cíl specifické (tj. ovlivňují minimálně další cíle a nevykazují nežádoucí vedlejší účinky)<sup>6</sup>. V případě nedostupnosti strukturální informace o molekulovém cíli pak sáhneme po některé z metod LBCADD. Sem patří např. podobnostní vyhledávání či modelování vztahu mezi strukturou a aktivitou látky (QSAR, angl. Quantitative Structure-Activity Relationship).

## 2. Virtuální screening

Nejběžnějším použitím CADD je virtuální screening (VS), tedy screening (testování) virtuálních (počítačových) knihoven organických sloučenin. Jedna z možností, jak klasifikovat metody virtuálního screeningu, je založena na typu otázky, kterou se každá z metod snaží odpovědět, a na předpokladech, které jejich uživatel musí učinit o aktivitě sloučenin<sup>7</sup>:

- 1. Podobnostní metody.** Otázka: Které molekuly v databázi jsou podobné dotazované sloučenině/sloučeninám? Racionále: Sloučeniny podobné dotazované aktivní molekule mají pravděpodobně stejnou aktivitu. Vyžaduje: Jednu nebo více známých aktivních molekul.
- 2. Farmakoforové modelování.** Otázka: Které molekuly v databázi obsahují specifickou podstrukturu, tzv. farmakofor? Racionále: Látky, jež obsahují daný farmakofor, budou pravděpodobně aktivní. Vyžaduje: Farmakofor společný látkám aktivním na zkoumaném cíli.
- 3. Kvantitativní vztah mezi strukturou a aktivitou látky (QSAR).** Otázka: Které molekuly mají nejvyšší predikovanou aktivitu? Racionále: Sloučeniny s vysokou predikovanou aktivitou budou nejspíše aktivní. Vyžaduje: Aktivitní data o dostatečném množství sloučenin za účelem vybudování QSAR modelu.
- 4. Dokování.** Otázka: Které molekuly v databázi nejlépe padnou do aktivního místa zkoumaného molekulového cíle? Racionále: Látky, které nejlépe padnou do aktivního místa cíle, se na něj budou pravděpodobně vázat nejsilněji. Vyžaduje: Strukturu molekulového cíle v atomovém rozlišení nebo homologní model.

Jak již bylo řečeno výše, hlavním cílem virtuálního screeningu je vybrat z velké chemické knihovny menší množinu pravděpodobně aktivních látek, které je možno následně testovat experimentálně. Všechny výše uvedené metody slouží k identifikaci aktivních molekul a liší se nejen svojí výpočetní náročností, ale též přesností. Proto se typicky aplikují sekvenčně, kdy se rychlejší, leč méně přesné metody (např. podobnostní vyhledávání), používají na začátku procesu virtuálního screeningu a výpočetně dražší (např. dokování) na jeho konci (obr. 1).



Obr. 1. Sekvenční uspořádání VS protokolu. Při sekvenčním uspořádání jsou aplikovány výpočetně náročnější metody na postupně se zmenšující množinu látek. Kombinují se metody jak ligandového, tak strukturálního VS

Podobnostní vyhledávání a QSAR představují ligandové přístupy, zatímco dokování je typickým představitelem strukturálního přístupu. Farmakoforové modelování může být aplikováno buď jako ligandová, nebo jako strukturální metoda. Všechny výše uvedené metody, s výjimkou dokování, mohou používat 2D či 3D reprezentaci molekul. 2D reprezentace se označuje také jako topologická, neboť je odvozena z konektivní tabulky odrážející topologii molekul. S pojmem topologická reprezentace se také pojí pojem topologická vzdálenost mezi dvěma atomy, která je dána jako počet vazeb mezi nimi. 3D reprezentace pak vyžaduje znalost pozic atomů v prostoru (tedy jejich xyz souřadnice) a vzdálenost mezi dvěma atomy je dána jako jejich Euklidovská vzdálenost. Z výše uvedených metod je v tomto čísle Chemických listů věnována separátní kapitola dokování<sup>8</sup> a QSAR metodám<sup>9</sup>, tyto tedy nebudou v následujícím textu diskutovány.

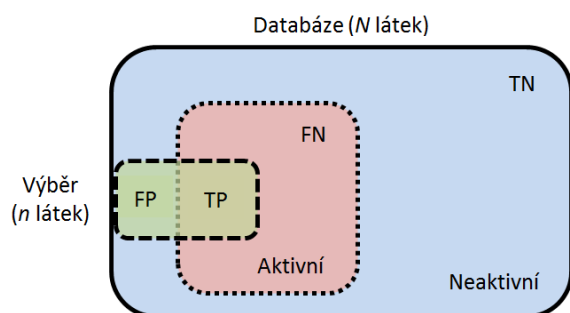
Obecným cílem VS metod je vybrat z molekulární databáze co nejvíce látek (ve srovnání s náhodným výběrem), jež se při následujícím experimentálním testování ukáží být opravdu aktivní. Takové látky se označují jako skutečně pozitivní (TP, angl. true positives). Neboť se však pracuje počítačovými predikcemi, jež mají vždy omezenou přesnost, běžně se do výsledného seznamu (angl. hit list) dostanou i látky experimentálně prokázané jako neaktivní (FP, angl. false positives). Zatímco výskyt FP pouze zbytečně navyšuje časovou a finanční náročnost násled-

ných experimentů, je větším problémem výskyt látek falešně negativních (FN, angl. false negatives). To jsou látky, jež jsou sice aktivní, ale počítačová metoda je jako takové neodhalí. Tyto látky tedy nejsou následně experimentálně testovány a jsou tak navždy ztraceny. Za účelem snížení počtu FN se proto kombinují různé VS metody s různými vnitřními nastaveními, nicméně ve výsledku si nikdy nemůžeme být jisti, kolik aktivních látek jsme vyloučili. Poslední skupinou sloučenin jsou skutečně negativní látky (TN, angl. true negatives), které jsou neaktivní a VS metoda je za neaktivní opravdu považuje. Celá situace je charakterizována na obr. 2.

Výsledkem VS bývá množina látek seřazených podle predikované vlastnosti (seřazená množina, angl. ranked list), jako je např. pravděpodobnost, že daná látka bude aktivní. VS protokol sestává z několika parametrů, které mohou být měněny a jejichž hodnoty mají vliv na celkový výsledek. Mezi tyto parametry patří např. různé způsoby reprezentace molekul (různé sady deskriptorů či různé fingerprinty, viz jiná kapitola v tomto čísle Chemických listů<sup>10</sup>) či použitá metoda VS a její vnitřní nastavení. Kvalita VS protokolu se vyhodnocuje pomocí různých metrik zkonstruovaných za použití počtu FP, FN, TP a TN<sup>11</sup>. Jedna z nejpoužívanějších metrik je faktor nabohacení (*EF*, angl. enrichment factor), který udává, kolikrát více aktivních látek bylo nalezeno v porovnání s náhodným výběrem. *EF* se vypočítá jako:

$$EF = \frac{TP/n}{A/N}$$

kde *A* je celkový počet aktivních látek v databázi a je roven *TP* + *FN*. *n* se určuje jako procento ze seřazené množiny; např. u *EF*<sub>5%</sub> se *TP* počítá z prvních 5 % seřazené množiny. Jednou z nevýhod *EF* je jeho závislost na poměru aktivních látek v testované databázi<sup>12</sup> a je možno ho použít pro porovnání různých VS protokolů pouze tehdy,



Obr. 2. Výběr *n* molekul (obdélník ohraničený čárkovanou čarou) z databáze obsahující *N* struktur pomocí VS protokolu. Databáze *N* struktur sestává z aktivních (obdélník ohraničený tečkovanou čarou) a neaktivních (obdélník ohraničený plnou čarou) látek. Při výběru pomocí VS protokolu mohou nastat následující situace: jako aktivní jsou vybrány látky, které jsou skutečně aktivní (TP), jako aktivní jsou vybrány látky, které jsou ve skutečnosti neaktivní (FP), VS protokol nevybral látky, které opravdu nejsou aktivní (TN) a VS protokol minul látky, které jsou ve skutečnosti aktivní (FN)

byla-li pomocí nich testována stejná databáze. Dalším problémem *EF* je skutečnost, že nebere do úvahy tzv. včasné rozpoznání (angl. early recognition), kdy by látky vyskytující se na vyšších pozicích seřazené množiny měly přispívat k metrice větší měrou. Pro zahrnutí včasného rozpoznání byly proto vyvinuty sofistikovanější metriky, jako např. RIE (angl. Robust Initial Enhancement)<sup>13</sup> či BEDROC (angl. Boltzmann-Enhanced Discrimination of the Receiver Operating Characteristics Curve (ROC))<sup>14</sup>.

## 2.1. Filtry ve virtuálním screeningu

Důležitou součástí VS jsou filtry, které umožňují rychle vyloučit sloučeniny nemající požadované fyzikální, chemické, biologické či farmakologické vlastnosti. Jedním z nejpoužívanějších filtrů je Lipinského pravidlo pěti<sup>15</sup> (Ro5, angl. rule of five), pomocí něhož se odhaduje šance látky stát se léčivem (angl. drug-likeness). Lipinského pravidlo pěti bylo odvozeno analýzou distribuce molekulové hmotnosti *M<sub>w</sub>*, logaritmu rozdělovacího koeficientu mezi oktanolem a vodou *logP* a počtu donorů *HBD* a akceptorů *HBA* vodíkových vazeb u 2245 sloučenin z databáze World Drug Index, které prošly minimálně do II. fáze klinického testování. Podle tohoto pravidla má většina orálně dostupných léčiv *M<sub>w</sub>* menší než 500 Da, *logP* menší než 5, *HBD* menší nebo rovno 5 a *HBA* menší nebo rovno 10. Ačkoliv jsou výše uvedená Lipinského kritéria dobře zavedená a nabízí relativně rychlý způsob posouzení farmakokinetických vlastností, nejsou schopna predikovat s jistotou, zda daná látka bude vhodná jako léčivo. Ostatně Ro5 bylo vždy chápáno spíše jako užitečná nápověda než jako pevně nastavitelný filtr. Zhruba 16 % orálně dostupných léčiv porušuje jedno z Lipinského pravidel a 6 % porušuje dvě a více pravidel<sup>16</sup>. Proto se látky, které porušují toliko jedno (libovolné) pravidlo, považují stále za drug-like, a až látky porušující dvě a více pravidel jsou odfiltrovány jako nevhodné. Lipinského pravidlo 5 bylo následně vylepšeno dodatečnými kritérii, kdy Veberova pravidla<sup>17</sup> dodávají k Ro5 počet rotovatelných vazeb (*nrot*) menší nebo rovný 10 a velikost polárního povrchu (*PSA*, angl. Polar Surface Area) menší než 140 Å<sup>2</sup>.

Ro5 kritéria byla odvozena z komerčně dostupných farmak a jsou tedy vhodná pouze pro výběr kandidátních léčiv určených k orálnímu podání a nikoliv pro výběr lead-like molekul<sup>18</sup>. Nicméně primárním cílem VS je identifikace lead-like molekul, proto má Ro5 jako filtr před vlastním VS omezenou použitelnost. Drug-like molekuly jsou navrhovány z lead-like molekul jejich optimalizací, v jejímž průběhu jsou typicky dodávány ke sloučenině velké hydrofobní skupiny vedoucí k nárůstu nejen biologické aktivity, ale také molekulové hmotnosti (o cca 200 kDa) a *logP* (o cca 0,5–4 jednotek)<sup>19</sup> a k poklesu rozpustnosti ve vodě (*logS<sub>w</sub>*)<sup>20</sup>. Pro filtrování lead-like molekul je možno použít např. Opreova pravidla (*M<sub>w</sub>* ≤ 460, −4 ≤ *logP* ≤ 4,5, *logS<sub>w</sub>* ≥ −5, *HBD* ≤ 5, *HBA* ≤ 9)<sup>21</sup> nebo přísnější pravidlo 3 (*M<sub>w</sub>* ≤ 300 Da, *logP* ≤ 3, *HBA/HBD* ≤ 3, *nrot* < 3)<sup>22</sup>.

Kromě výše uvedených fyzikálně-chemických filtrů máme ovšem k dispozici celou řadu dalších metod, jak

vyložit látky na základě jejich komplexnějších vlastností. Látky mohou být např. považovány za méně atraktivní, obsahují-li tzv. toxikofory, funkční skupiny spojené s metabolickou toxicitou<sup>23</sup>. Dále je vhodné odstranit molekuly obsahující reaktivní skupiny či tzv. PAINS (angl. Pan Assay Interference Compounds)<sup>24</sup>. PAINS jsou látky, které nesespecificky ovlivňují výsledek HTS experimentu např. tak, že vytváří agregáty, čímž zkrasují detegovaný signál a mohou tak být chybně považovány za biologicky aktivní. Pro filtrování lze s výhodou využít veřejně dostupných serverů, jako je např. FAF-Drugs4 server<sup>25</sup>. Ten umožňuje filtrovat sloučeniny na základě různých fyzikálně-chemických vlastností, řady kritérií na lead- a drug-likeness či vyloučit PAINS látky nebo látky obsahující toxikofory.

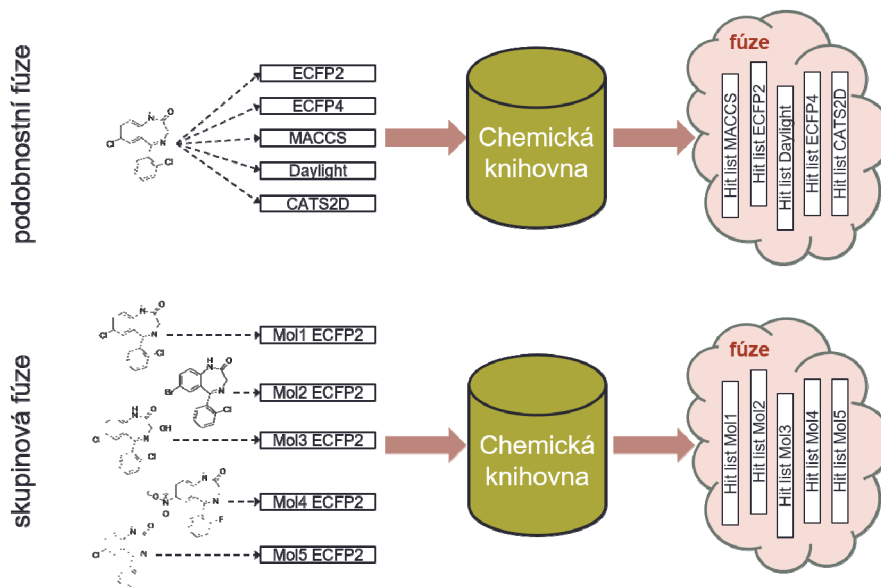
## 2.2. Podobnostní metody

Podobnostní metody<sup>26</sup>, které se vyvíjejí od 80. let 20. století, prokázaly již mnohokrát svoji užitečnost z následujících důvodů:

1. Pro formulování dotazu je potřeba relativně málo informace. Konkrétně není třeba definovat, která část molekuly je odpovědná za její biologickou aktivitu. Ve 2D verzi podobnostního vyhledávání není nutná znalost aktivní konformace molekuly. Z tohoto důvodu se podobnostní metody používají na začátku procesu hledání nového léčiva, kdy je k dispozici pouze pár molekul aktivních na tomto cíli.

2. Různé implementace podobnostního vyhledávání jsou výpočetně velmi efektivní a prohledávání velikých chemických databází je možno provádět zcela rutinně.
3. Předpoklad, že by strukturně podobné molekuly měly vykazovat podobnou aktivitu, tzv. „similar-property principle“<sup>27</sup>, se ukázal být obecně platný<sup>28</sup>.

Konceptuálně každá z podobnostních metod sestává ze dvou aspektů: jak reprezentovat molekuly a jak vyčíslit podobnost mezi nimi. Nejpoužívanějším způsobem reprezentace molekul je převést je na vektorovou reprezentaci a podobnost mezi nimi počítat pomocí některé z podobnostních měr, jak je popsáno na jiném místě tohoto čísla Chemických listů<sup>10</sup>. Velké množství vektorových reprezentací struktur (různé fingerprinty, rozličné sady deskriptorů apod.) a podobnostních měr vedlo ke snaze nalézt optimální nastavení podobnostního vyhledávání. Ačkoliv je o jistých kombinacích známo, že poskytují rozumnou míru efektivity, jako je např. použití ECFP4 fingerprintů<sup>29</sup> s Tanimotovým podobnostním koeficientem, je obecně akceptováno, že neexistuje jeden jediný způsob, který by poskytoval optimální výsledky za všech okolností<sup>30</sup>. A právě proto byly navrženy techniky tzv. fúze dat (angl. data fusion)<sup>31</sup> kombinující výsledky z dotazů provedených za použití různých reprezentací molekul a různých podobnostních měr. Metody fúze dat využívají tzv. „ranking“ (pořadí v seřazeném seznamu) a jsou aplikovatelné nejen na podobnostní vyhledávání, ale též např. na dokování, v kterémžto případě se označují jako konsensuální skórování (angl. consensus scoring)<sup>32</sup>.



Obr. 3. **Podobnostní a skupinová fúze.** Při podobnostní fúzi je dotazem do databáze jedna aktivní molekula (referenční molekula), která je kódována pomocí různých vektorových reprezentací (v daném příkladu byly použity fingerprinty ECFP2, ECFP4, MACCS, Daylight a CATS2D). Při skupinové fúzi je dotazem do databáze více různých struktur (označeny jako Mol1–Mol5), které jsou kódovány pomocí stejné vektorové reprezentace (zde ECFP2). V obou případech se pak různé výsledné množiny kombinují dle některého z fúzovacích pravidel

V podobnostním VS existují k fúzi dva přístupy<sup>33</sup>: podobnostní fúze (angl. similarity fusion) a skupinová fúze (angl. group fusion) (obr. 3).

Při datové fúzi se obecně postupuje tak, že je provedeno  $n$  dotazů do databáze, které poskytnou  $n$  seřazených seznamů molekul, jež jsou podobné dotazu. Poté je  $n$  seznamů sloučeno v jeden pomocí tzv. fúzovacího pravidla. Fúzovací pravidlo se aplikuje buď přímo na numerické vyjádření podobnosti látek, nebo na jejich pořadí v seřazeném seznamu. Většina existujících fúzovacích pravidel je aplikovatelná jak na podobnost, tak na pořadí. V praktických aplikacích se dává přednost fúzi podle pořadí, neboť fúze podle podobnosti může být, z důvodu odlišných rozdělení různých podobnostních měř, předpojatá (angl. biased). Mezi základní fúzovací pravidla patří MAX, MIN a SUM pravidlo. V MAX (MIN) pravidle je struktuře přiřazeno maximální (minimální) skóre (tedy buď číselné vyjádření podobnosti, nebo pořadí látky v seřazeném seznamu), které tato struktura získala v  $n$  vygenerovaných seznamech. V SUM pravidle se pak skóre ze všech  $n$  seznamů posčítají za vzniku nového skóre.

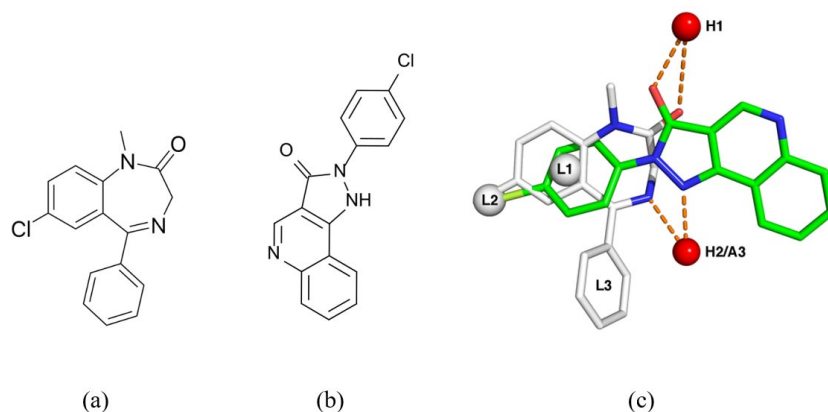
### 2.3. Farmakoforové modelování

Máme-li k dispozici jednu či více aktivních molekul, umožňuje podobnostní vyhledávání najít látky těmto molekulám strukturně podobné. Nicméně podobnostní vyhledávání již není schopno identifikovat látky, které jsou aktivní na stejném cíli, leč nejsou si strukturně podobné. Příkladem takové dvojice je diazepam a látka CGS-9896, které se, ač mají velmi rozdílné skelety (obr. 4a, b), oba váží na benzodiazepinový receptor  $GABA_A$  (cit.<sup>34</sup>). Pod pojmem skelet (angl. scaffold) se rozumí základní kostra molekuly

sestavající z kruhů pospojovaných tzv. „linkery“<sup>35</sup>. Nalezení nové aktivní látky s jiným skeletem než má např. přirozený ligand se označuje jako změna skeletu (angl. scaffold hopping)<sup>36</sup> a je žádoucí z několika důvodů. Např. umožňuje v chemické sérii nahradit skelet problematický z důvodu obtížnosti syntézy. Nejdůležitější motivací pro změnu skeletu je však duševní vlastnictví (IP, angl. intellectual property), v jehož kontextu se jedná o nalezení aktivní struktury, která není chráněna původním patentem a která není výsledkem přímočaré strukturní modifikace patentovaných struktur.

V příkladu diazepam/CGS-9896 je důvodem jejich aktivity nikoliv strukturní podobnost, ale skutečnost, že obě molekuly sdílí stejný tzv. farmakofor. Farmakofor je sada strukturních a fyzikálně chemických vlastností, jež jsou nutné pro vazbu látky do aktivního místa<sup>37</sup>. Farmakofor v rámci jedné molekuly popisuje třírozměrné uspořádání chemických vlastností zvaných farmakoforové rysy (angl. pharmacophore features), mezi něž patří např. pozice donorů a akceptorů vodíkových vazeb, aromatických kruhů, hydrofobních center a pozitivně/negativně nabitých či kyselých/bazických částí molekuly<sup>38</sup>. Farmakoforový model (angl. pharmacophore model) je pak konsensuální sada farmakoforových rysů ve 3D prostoru (obr. 4c) a může být zkonstruován buď překrytím farmakoforů aktivních látek (tedy ligandový přístup), nebo detegováním možných interakčních míst ve struktuře molekulového cíle (tedy strukturní přístup)<sup>39</sup>.

Nespornou výhodou ligandového farmakoforového modelu je skutečnost, že může být zkonstruován bez znalosti geometrie aktivního místa. Praxe je však komplikovanější, neboť neexistuje spolehlivý způsob, jak zjistit, která z možných konformací ligandu odpovídá té aktivní. Ještě



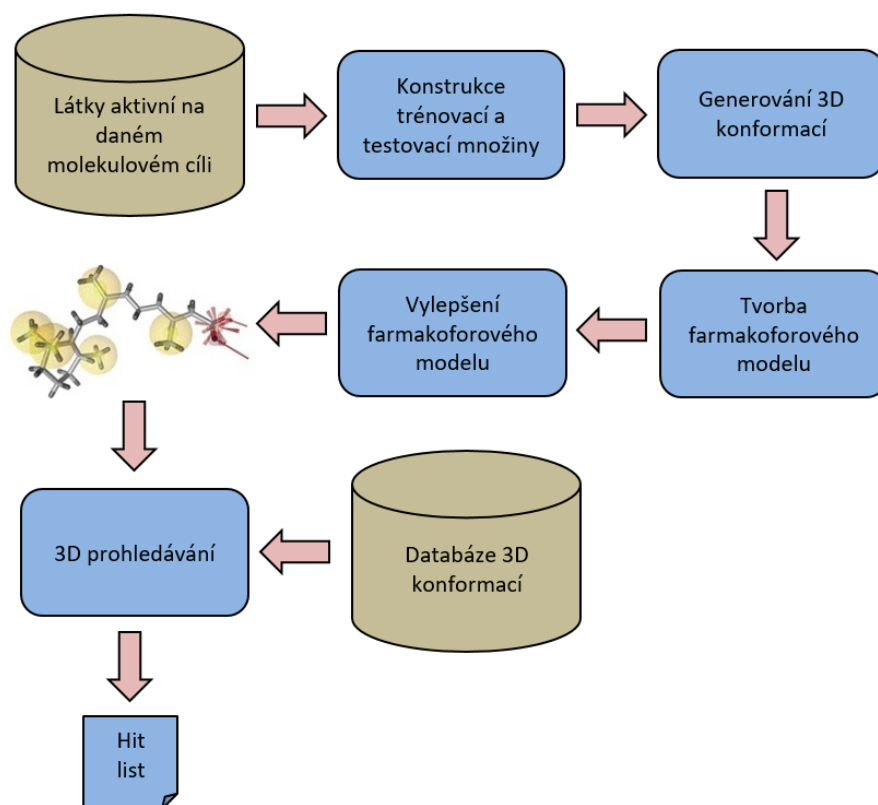
Obr. 4. **Ligandy benzodiazepinového receptoru.** Diazepam (a) patří do skupiny benzodiazepinů, jež sestávají z benzenového a diazepinového jádra. Jedná se o skupinu psychoaktivních léčiv, která interagují s benzodiazepinovým receptorem  $GABA_A$ <sup>34</sup> a vykazují anxyolitické (potlačují úzkost) účinky. Podobný biologický efekt má i léčivo CGS-9896 (b), které je ale strukturně odlišné a řadí se mezi nebenzodiazepinová anxyolitika. Farmakoforový model pro vazebné místo na  $GABA_A$  receptoru (c) vznikl překrytím 3D struktur diazepam a CGS-9896. H1 a H2/A3 jsou místa na receptoru fungující jako donory a akceptory vodíkových vazeb, zatímco L1, L2 a L3 označují lipofilní vazebná místa. Překrytí struktur (a) a (b) ve třírozměrném prostoru tak, jak je znázorněno na obrázku, povede k farmakoforovému modelu sestávajícímu z H1, H2/A3, L1 a L2 farmakoforových rysů<sup>51</sup>



větší překážku však představuje situace, kdy se v aktivním místě nachází důležitá funkční skupina, která není ve farmakoforovém modelu podchycena, neboť žádný z aktivních ligandů daný farmakoforový rys neobsahuje. Z tohoto důvodu je dobré konstruovat farmakoforový model z chemicky co nejvíce diverzní množiny struktur. Dále je třeba upozornit, že farmakoforový model obsahující pouze rysy přítomné v aktivních ligandech vrací při databázovém vyhledávání příliš mnoho falešně pozitivních výsledků. Důvodem je skutečnost, že nalezené ligandy sice budou mít žádané rysy na správných místech, ale kromě toho budou také obsahovat velké skupiny, jež zabrání interakci ligandu ve vazebném místě. Proto se do farmakoforových modelů ještě přidává jako další rys tzv. vyloučený objem (angl. excluded volume), který penalizuje atomy na pozicích, jež by kolidovaly s vazebným místem.

Tvorba ligandového farmakoforového modelu sestává ze tří hlavních kroků: 1) návrh množiny ligandů, které budou použity pro konstrukci farmakoforového modelu (tzv. trénovací množina), 2) generování energeticky výhodných 3D konformací pro každou látku z trénovací

množiny a 3) překrytí struktur trénovací množiny ve 3D prostoru (angl. 3D alignment) následované detekcí společných farmakoforových rysů. Ačkoliv existuje řada algoritmů pro automatickou tvorbu farmakoforového modelu<sup>40</sup>, nejsou tyto v současnosti dostatečně robustní a výsledný model je třeba vždy doladit ručně. Prvním problémem při automatické tvorbě farmakoforového modelu je modelování flexibility ligandu, tedy generování jeho 3D konformací. To by mělo být co nejrychlejší, vygenerované konformace by měly mít co nejnižší energii a měly by odpovídat těm, které může malá molekula nabýt při interakci s cílem. Pro generování 3D konformací malých molekul bylo publikováno mnoho různých algoritmů<sup>41</sup>, které jsou vhodné v různých situacích a z nichž žádný neposkytuje konzistentně nejlepší výsledky. Druhou výzvu při farmakoforovém modelování představuje zarovnání (angl. alignment) konformací v 3D prostoru<sup>42</sup>. Při zarovnání se nejčastěji postupuje tak, že se minimalizují vzdálenosti mezi jednotlivými atomy, molekulovými fragmenty nebo farmakoforovými rysy. Dalším aspektem tvorby farmakoforového modelu, kterému je třeba věnovat zvýše-



Obr. 5. **Schéma farmakoforového virtuálního screeningu.** Celý proces začíná nalezáním všech látek aktivních na daném molekulovém cíli. Ty jsou poté rozděleny na trénovací a testovací množiny a jsou vygenerovány jejich 3D konformace. Trénovací množina se používá pro tvorbu farmakoforového modelu, který je iterativně vylepšován na základě toho, jak moc je schopný rozlišovat mezi aktivními a neaktivními látkami v testovací množině. Výsledný farmakoforový model je pak použit jako dotaz do databáze obsahující 3D struktury širokého spektra molekul. Výsledkem je množina látek (hit list) splňujících daný farmakoforový model. Obrázek „Representation of the structure-based pharmacophores“ od autorů Legarde N. a spol.<sup>44</sup> je licencovaný pod CC Attribution 4.0 International License

nou pozornost, je výběr trénovací množiny. Ta musí být zkonstruována velmi pečlivě, neboť ligandy v množině přítomné, její velikost a chemická diverzita velmi významně ovlivňují kvalitu výsledného farmakoforového modelu.

Při strukturálním farmakoforovém modelování se analyzuje prostorové uspořádání farmakoforových rysů přímo z 3D struktury molekulového cíle. Metody strukturálního farmakoforového modelování je možno dále rozdělit do dvou kategorií podle toho, zda máme k dispozici pouze 3D strukturu molekulového cíle, nebo strukturu komplexu ligandu s molekulovým cílem. Strukturální farmakoforové metody založené na analýze komplexu jsou přirozeně přesnější, neboť umožňují spolehlivější detekci interakčních míst ve vazebném místě. Příkladem metody založené na analýze komplexu je aktuálně hojně používaný komerčně dostupný LigandScout<sup>43</sup>.

Získaný farmakoforový model je následně možné použít při virtuálním screeningu jako dotaz do databáze 3D struktur za účelem nalezení ligandů, které daný model splňují (obr. 5).

Některé z těchto ligandů mohou být podobné známým aktivním ligandům, u některých však mohlo dojít ke změně skeletu<sup>36</sup>. Při farmakoforovém screeningu je třeba mít k dispozici kromě kvalitního farmakoforového modelu také databázi s vygenerovanými 3D konformacemi ligandů a algoritmus identifikace struktury odpovídající danému farmakoforovému modelu. Volně přístupným zdrojem 3D konformací je databáze komerčně dostupných molekul ZINC<sup>45</sup>. Příkladem algoritmu pro nalezení struktury odpovídající danému farmakoforovému modelu jsou Ullmanův algoritmus<sup>46</sup> či backtracking<sup>47</sup>.

Navzdory všem výše jmenovaným obtížím představují farmakoforový VS, spolu s dokováním, hlavní proudy ve virtuálním screeningu. Ve srovnání s dokováním netrpí farmakofory problémy vzniklémi nedostatečným popisem flexibility molekulového cíle ani chybami plynoucími z nekvalitně navržené skórovací funkce. Farmakoforový VS je také rychlejší než dokování, a proto bývá aplikován v ranějších fázích projektu. Největším problémem farmakoforového screeningu je ovšem vysoké procento falešně pozitivních výsledků. K tomu přispívá mnoho různých faktorů, z nichž nejzřejmější jsou nedostatky ve farmakoforovém modelu. Dalším zdrojem chyb je způsob, jakým se pracuje s flexibilitou molekulového cíle. Ta je do farmakoforového screeningu zahrnuta pomocí zavedení poloměru tolerance každého farmakoforového rysu, kterýžto postup však v některých případech nemusí být dostatečný. Tomu napovídá i skutečnost, že farmakoforové modely generované z trajektorií molekulové dynamiky vykazují významně lepší reprezentaci flexibility farmakoforu<sup>48</sup>. Dalším zdrojem falešně pozitivních výsledků je fakt, že většina interakcí mezi ligandem a proteinem je silně závislá na vzdálenosti, což platí obzvláště pro krátkodosahové elektrostatické interakce. Tento fakt není farmakoforový model schopen popsat, a proto bylo vyvinuto několik hybridních protokolů úspěšně kombinujících farmakoforové modelování a dokování<sup>49</sup>.

### 3. Závěr

V posledních letech jsme svědky velkého nárůstu počtu úspěšných aplikací virtuálního screeningu v procesu identifikace malých biologicky aktivních sloučenin. VS je aplikován především v medicíně, kde je používán pro vytipování potenciálních lead molekul, nebo v chemické biologii, kde se hledají malé ligandy (tzv. próby), jež je možno použít pro zkoumání biologických jevů na molekulární úrovni. VS používá celou řadu metod založených buď na znalosti 2D struktur aktivních ligandů, nebo na znalosti 3D struktury molekulového cíle. Mezi tyto metody patří např. podobnostní vyhledávání, farmakoforové vyhledávání, modelování kvantitativního vztahu mezi strukturou a biologickou aktivitou látky (QSAR) či dokování. V procesu VS se také používají metody pro zjištění drug- či lead-likeness látek, jejich toxických účinků či snadnosti jejich syntézy, kteréžto problematice je věnován jiný článek tohoto čísla Chemických listů<sup>50</sup>. Za účelem vylepšení efektivity VS se tyto metody typicky kombinují hierarchickým způsobem, kdy výpočetně méně náročné přístupy (např. podobnostní či farmakoforové vyhledávání) se používají na začátku, zatímco ty náročnější (např. QSAR a především dokování) se aplikují až na konci celého procesu. Dnes je víc než zřejmé, že výpočetní metody se pevně etablovaly v procesu návrhu nového léčiva či próby a provázání mezi nimi a experimentálním výzkumem dosáhlo takové hloubky, že výstupy z jedné oblasti jsou nezbytné pro pokrok v oblasti druhé.

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

#### Seznam symbolů

ADME	absorpce, distribuce, metabolismus, exkrece, Tox – toxicita
BEDROC	Boltzmann-Enhanced Discrimination of ROC
CADD	počítačový návrh léčiv
EF	faktor nabohacení
FN	falešně negativní látka
FP	falešně pozitivní látka
HBA	počet akceptorů vodíkové vazby
HBD	počet donorů vodíkové vazby
HTS	testování s vysokou propustností
IP	duševní práva
LBCADD	ligandový CADD
logP	logaritmus rozdělovacího koeficientu mezi oktánolem a vodou
PSA	velikost polárního povrchu
PAINS	Pan Assay Interference Compounds
QSAR	kvantitativní vztah mezi strukturou a aktivitou látky
RIE	Robust Initial Enhancement
ROC	operační charakteristika zesilovače
Ro5	pravidlo pěti

SBCADD	strukturní CADD
TN	skutečně negativní látky
TP	skutečně pozitivní látky
VS	virtuální screening

## LITERATURA

1. Van Drie J. H.: *J. Comput. Aided. Mol. Des.* 21, 591 (2007).
2. Broach J. R., Thorner J.: *Nature* 384 (1996).
3. Popr M., Sedlák D., Bartůněk P.: *Chem. Listy* 111, 772 (2017).
4. Muller T., Sedlák D., Bartůněk P.: *Chem. Listy* 111, 766 (2017).
5. Doman T. N., McGovern S. L., Witherbee B. J., Kasten T. P., Kurumbail R., Stallings W. C., Connolly D. T., Shoichet B. K.: *J. Med. Chem.* 45, 2213 (2002).
6. Jorgensen W. L.: *Nature* 466, 42 (2010).
7. Sheridan R. P., Kearsley S. K.: *Drug Discovery Today* 7, 903 (2002).
8. Šicho M., Svozil D.: *Chem. Listy* 111, 754 (2017).
9. Škuta C., Svozil D.: *Chem. Listy* 111, 747 (2017).
10. Novotný J., Svozil D.: *Chem. Listy* 111, 716 (2017).
11. Kirchmair J., Markt P., Distinto S., Wolber G., Langer T.: *J. Comput.-Aided Mol. Des.* 22, 213 (2008).
12. Triballeau N., Acher F., Brabet I., Pin J. P., Bertrand H. O.: *J. Med. Chem.* 48, 2534 (2005).
13. Sheridan R. P., Singh S. B., Fluder E. M., Kearsley S. K.: *J. Chem. Inf. Comput. Sci.* 41, 1395 (2001).
14. Truchon J. F., Bayly C. I.: *J. Chem. Inf. Model.* 47, 488 (2007).
15. Lipinski C. A.: *Drug Discovery Today: Technol.* 1, 337 (2004).
16. Bickerton G. R., Paolini G. V., Besnard J., Muresan S., Hopkins A. L.: *Nat. Chem.* 4, 90 (2012).
17. Veber D. F., Johnson S. R., Cheng H. Y., Smith B. R., Ward K. W., Kopple K. D.: *J. Med. Chem.* 45, 2615 (2002).
18. Oprea T. I., Davis A. M., Teague S. J., Leeson P. D.: *J. Chem. Inf. Comput. Sci.* 41, 1308 (2001).
19. Teague S. J., Davis A. M., Leeson P. D., Oprea T.: *Angew. Chem., Int. Ed. Engl.* 38, 3743 (1999).
20. Oprea T. I.: *J. Comput.-Aided Mol. Des.* 16, 325 (2002).
21. Hann M. M., Oprea T. I.: *Curr. Opin. Chem. Biol.* 8, 255 (2004).
22. Congreve M., Carr R., Murray C., Jhoti H.: *Drug Discovery Today* 8, 876 (2003).
23. Williams D. P.: *Toxicology* 226, 1 (2006).
24. Baell J. B., Holloway G. A.: *J. Med. Chem.* 53, 2719 (2010).
25. Lagorce D., Sperandio O., Baell J. B., Miteva M. A., Villoutreix B. O.: *Nucleic Acids Res.* 43, W200 (2015).
26. Bajorath J.: *J. Chem. Inf. Comput. Sci.* 41, 233 (2001).
27. Johnson M. A., Maggiora G.: *Concepts and Applications of Molecular Similarity*. John Wiley & Sons Inc., New York 1990.
28. Patterson D. E., Cramer R. D., Ferguson A. M., Clark R. D., Weinberger L. E.: *J. Med. Chem.* 39, 3049 (1996).
29. Rogers D., Hahn M.: *J. Chem. Inf. Model.* 50, 742 (2010).
30. Bender A., Jenkins J. L., Scheiber J., Sukuru S. C., Glick M., Davies J. W.: *J. Chem. Inf. Model.* 49, 108 (2009).
31. Willett P.: *Comput. Struct. Biotechnol. J.* 5, e201302002 (2013).
32. Feher M.: *Drug Discovery Today* 11, 421 (2006).
33. Whittle M., Gillet V. J., Willett P., Loesel J.: *J. Chem. Inf. Model.* 46, 2206 (2006).
34. Kumar K., Sharma S., Kumar P., Deshmukh R.: *Pharmacol. Biochem. Behav.* 110, 174 (2013).
35. Bemis G. W., Murcko M. A.: *J. Med. Chem.* 39, 2887 (1996).
36. Zhao H.: *Drug Discovery Today* 12, 149 (2007).
37. Wermuth G., Ganellin C. R., Lindberg P., Mitscher L. A.: *Pure Appl. Chem.* 70, 1129 (1998).
38. Gillet V. J., v knize: *Physico-Chemical and Computational Approaches to Drug Discovery* (Luque J., Barril X., eds.), 151-170, RSC, Cambridge 2012.
39. Pirhadi S., Shiri F., Ghasemi J. B.: *Curr. Top. Med. Chem.* 13, 1036 (2013).
40. Yang S. Y.: *Drug Discovery Today* 15, 444 (2010).
41. Hawkins P. C. D.: *J. Chem. Inf. Model.* 57, 1747 (2017).
42. Lemmen C., Lengauer T.: *J. Comput.-Aided Mol. Des.* 14, 215 (2000).
43. Wolber G., Langer T.: *J. Chem. Inf. Model.* 45, 160 (2005).
44. Lagarde N., Delahaye S., Zagury J. F., Montes M.: *J. Cheminform.* 8, 43 (2016).
45. Sterling T., Irwin J. J.: *J. Chem. Inf. Model.* 55, 2324 (2015).
46. Ullmann J. R.: *J. ACM* 31 (1976).
47. Barnard J. M.: *J. Chem. Inf. Comp. Sci.* 33, 532 (1993).
48. Gimenez-Oya V., Villacanas O., Fernandez-Busquets X., Rubio-Martinez J., Imperial S.: *J. Mol. Model.* 15, 997 (2009).
49. Kirchmair J., Distinto S., Schuster D., Spitzer G., Langer T., Wolber G.: *Curr. Med. Chem.* 15, 2040 (2008).
50. Voršilák M., Svozil D.: *Chem. Listy* 111, 760 (2017).
51. <https://en.wikipedia.org/wiki/Pharmacophore>, staženo 5.8.2017.



**D. Svozil<sup>a,b</sup>** (<sup>a</sup> *CZ-OPENSREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Academy of Sciences of the Czech Republic, Prague,* <sup>b</sup> *CZ-OPENSREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague*): **Virtual Screening**

The dominant technique for the identification of potential drugs is the physical screening of large libraries of chemicals against a biological target (high-throughput screening, HTS). An alternative approach, known as virtual screening, is designed to computationally screen large libraries of chemicals for compounds that complement targets of known structure and experimentally test those

that are predicted to bind well. There are two broad categories of virtual screening techniques: the ligand-based and the structure-based approaches. Ligand-based methods use only the information about those ligands that are known to be active against a given target. Examples of ligand-based approaches are similarity search or pharmacophore modelling. Structure-based virtual screening requires the knowledge of 3D structure of target protein. A typical structure-based method is the so called docking that tries to predict the binding affinity between a ligand and its target. In this contribution, we provide an overview of both ligand- and structure-based virtual screening approaches and discuss their limitations in computer-aided molecular design. Selected methods are described in more detail in some other papers of this issue of *Chemické listy*.