Chem. Listy *111*, 724–730 (2017) Referát

CHEMICKÝ PROSTOR

IVAN ČMELO[a,b] 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*
cmeloi@vscht.cz

Došlo 24.7.17, přijato 5.10.2017.

Klíčová slova: chemický prostor; chemické knihovny; počítačový návrh molekul

Obsah

1. Rozsah chemického prostoru

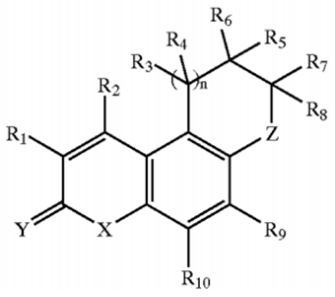
Chemický prostor je, ve svém nejširším slova smyslu, množinou všech molekul dostatečně stálých na to, aby mohly za reálných podmínek existovat[1]. Množství takovýchto stálých molekul je ze všech praktických hledisek nezměrné. I odhady týkající se toliko množství struktur potenciálních léčiv se řádově liší v závislosti na zdroji [2–4], bez výjimky však jde o astronomicky velká čísla. Jen počet možných struktur s molekulovou hmotností pod 500 Da je odhadován na 10^{60} molekul[1]. Ani takto omezenou množinu struktur nelze reálně připravit celou. Pokud by se celý proces návrhu, syntézy a následné izolace nové molekuly zrychlil v průměru na jednu vteřinu, příprava takové „kompletní“ chemické knihovny malých molekul by trvala přes 10^{52} let. Pokud bychom nevyžadovali syntézu sloučenin v chemickém prostoru a spokojili bychom se pouze s vyčíslením všech struktur, které ho tvoří, tak ani s dnešními výpočetními prostředky bychom nedosáhli kýženého výsledku v reálně dosažitelném čase. Dosud největším projektem v tomto směru je databáze GDB-17 (cit.[5]), kompletní vyčíslení všech chemických struktur s velikostí 17 a méně atomů C, N, O, S, F, Cl, Br a I obsahující $166 \cdot 10^9$ struktur.

Pro značnou část prováděného výzkumu jsou navíc z časových a ekonomických důvodů relevantní pouze molekuly, které je možno zakoupit či jiným způsobem získat z katalogu výrobce či distributora. Tuto množinu látek nazýváme komerčně dostupným chemickým prostorem. Představu o jeho velikosti nám může poskytnout databáze ZINC, která agreguje látky z katalogů většiny výrobců a dodavatelů[6]. ZINC v současné době (červen 2017) obsahuje 387 milionů komerčně dostupných látek, což velikostí odpovídá 2,3 ‰ databáze GDB-17, a prakticky nekonečně malému zlomku odhadované velikosti celého chemického prostoru.

Chemický prostor je tedy ze všech praktických hledisek nespočetně velký a je tudíž použitelný toliko jako abstrakce a kontext sjednocující v praxi využívaná zobrazení, postupy a metody. V tomto článku stručně popíšeme ústřední paradigmata založená na chemickém prostoru spolu s metodami, které je využívají.

2. Strukturní reprezentace chemického prostoru

Chemiky běžně používanou reprezentací částí chemického prostoru jsou generické (Markushovy) vzorce, jejichž použití v patentech úspěšně obhájil Eugene A. Markush již v roce 1924 (cit.[7]). Generické vzorce sestávají z hlavní (rodičovské) struktury, na kterou se ve vyznačených oblastech váží substituenty z předdefinovaných množin (obr. 1). Generické vzorce tak definují místo jedné konkrétní struktury celou množinu chemických struktur. Z hlediska chemického prostoru tedy Markushovy vzorce



Obr. 1. **Markushův vzorec převzatý z patentové přihlášky na tricyklické modulátory androgenního receptoru[8].** Množiny možných substituentů byly pro názornost zjednodušeny: R_1 = {F, Cl, Br, I, NO_2, OR_2, $SC_{1\text{-}8}$, $SOC_{1\text{-}8}$, $SO_2C_{1\text{-}8}$, $N(C_{1\text{-}8})C_{1\text{-}8}$}, $R_{2\text{-}8}$ = {F, Cl, Br, I, CH_3, CF_3, CHF_2, CH_2F, CF_2Cl, CN, $CF_2OC_{1\text{-}8}$, $CH_2OC_{1\text{-}8}$, $OC_{1\text{-}8}$, $SC_{1\text{-}8}$, $SOC_{1\text{-}8}$, $SO_2C_{1\text{-}8}$, $C_{1\text{-}8}$, $N(C_{1\text{-}8})C_{1\text{-}8}$}, $R_{9\text{-}10}$ = {{F, Cl, Br, I, CN, $OC_{1\text{-}8}$, $N(C_{1\text{-}8})C_{1\text{-}8}$, $C_m(C_{1\text{-}8})_{2m}$ m = (0;1;2)}}, $OC_{1\text{-}8}$, $SC_{1\text{-}8}$, $SOC_{1\text{-}8}$, $SO_2C_{1\text{-}8}$, $N(C_{1\text{-}8})C(O)C_{1\text{-}8}$, $C_{1\text{-}8}$}, X, Z = {O, S, $NC_{1\text{-}8}$}, Y = {O, S, $NC_{1\text{-}8}$, $C(C_{1\text{-}8})C_{1\text{-}8}$}, n = {1, 2}

umisťují pod patentovou ochranu místo jednotlivých bodů (struktur) celé oblasti chemického prostoru. Patentové krytí Markushovými vzorci definovaných oblastí chemického prostoru umožňuje najednou patentovat celou připravenou chemickou sérii, přímým důsledkem jejich používání je však také možnost patentovat navrhovatelem dosud nesyntetizované, tj. zcela virtuální struktury[7]. Například Markushův vzorec na obr. 1 pochází z patentové přihlášky na tricyklické modulátory androgenního receptoru[8] a, ačkoliv má zde pro názornost zjednodušenou množinu možných substituentů, může i v této formě kombinatoricky vzato pokrývat více než 10^{20} jednotlivých struktur. Nelze očekávat, že by držitel tohoto patentu skutečně syntetizoval a izoloval každou z tímto patentem chráněných struktur.

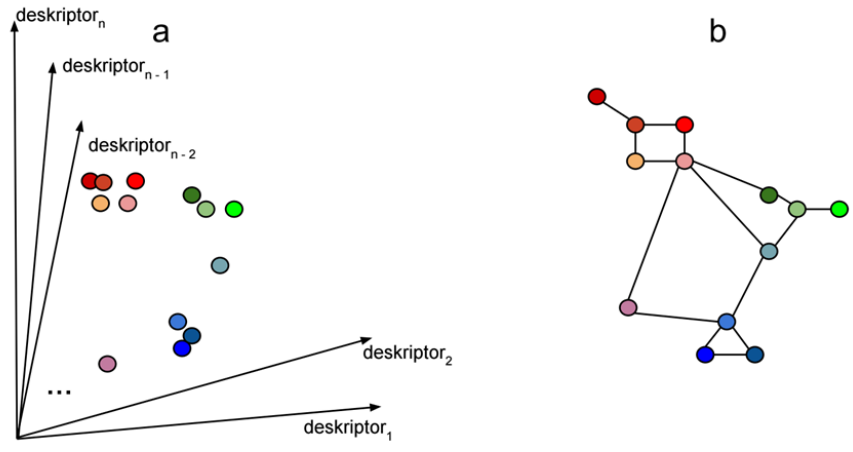
Užívání Markushových vzorců v patentech má také dopad na prohledávání chemického prostoru v patentových databázích. Úplná kombinatorická enumerace každého registrovaného Markushova vzorce by byla z hlediska úložných kapacit velmi náročná, do databází se tedy většinou ukládá samotný Markushův vzorec a jen několik z mnoha možných explicitních struktur[7]. Patentem chráněná oblast chemického prostoru se tedy v databázi projeví jen jako několik bodů (explicitních struktur), což může při vyhledávání patentů podle uživatelem zadané struktury vést k falešně negativním výsledkům. I navzdory stále se vyvíjející metodologii zůstává vyhledávání struktur v patentových databázích, a tudíž i strukturně orientované patentové rešerše, netriviálním problémem[7].

Reprezentace chemického prostoru založená na samotných chemických strukturách je historicky nejstarší, pracuje s nejmenší mírou abstrakce a nejlépe tak reflektuje realitu. Na ostatní reprezentace chemického prostoru, zejména na souřadnicové a síťové, lze pohlížet jako na vyšší abstrakce nad strukturní reprezentací chemického prostoru. Strukturní reprezentace má přímý chemický význam, zatímco ostatní reprezentace bývají zaváděny a používány primárně pro zpřístupnění chemického prostoru výpočetním a statistickým metodám.

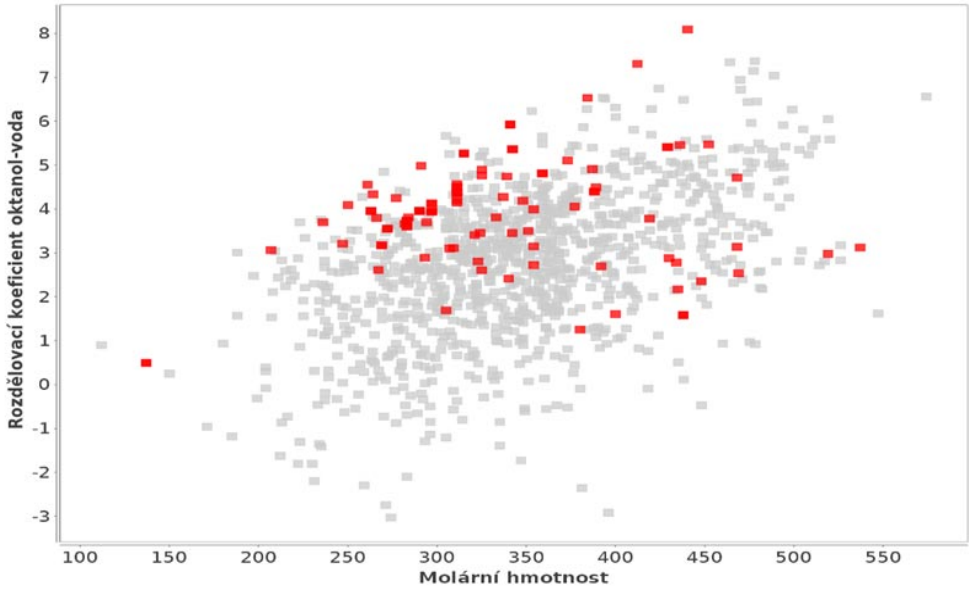
3. Souřadnicová reprezentace chemického prostoru

Chemický prostor je jakožto množina všech stálých molekul sám o sobě bezrozměrný. Při vizualizaci částí chemického prostoru a tvorbě predikčních modelů je však praktické tuto množinu nebo její části nějakým způsobem dimenzionalizovat. Intuitivním přístupem k dimenzionalizaci chemického prostoru je rozmístění molekul jako bodů do prostoru podle jejich deskriptorů[9] (obr. 2a). Známých deskriptorů je však příliš mnoho a pokud by každý z nich tvořil jednu osu v naší vizualizaci chemického prostoru, vzniklý prostor by obsahoval velké množství dimenzí. Velké množství dimenzí prostoru pak způsobuje řídké rozmístění bodů (v našem případě molekul), což velmi ztěžuje aplikaci výpočetních metod opírajících se o statistickou významnost. Tento jev je znám napříč mnoha se statistikou pracujícími obory jako tzv. „prokletí dimenzionality“ (curse of dimensionality). Přirozeným způsobem redukce počtu dimenzí je cílený výběr rysů, které poskytují pro daný model nebo vizualizaci nejvíce relevantní informace (feature selection). O výběru rysů lze uvažovat jako o výběru pro daný účel nejvíce relevantních dimenzí (tedy deskriptorů) z nespočetného množství potenciálních dimenzí chemického prostoru, čímž vzniká jeho méněrozměrná reprezentace, ve které se bude daná vizualizace nebo model pohybovat (obr. 3).

Výběr rysů z mnohadimenzionálních dat se v oblasti strojového učení běžně provádí například obalovacími (wrapper) a filtračními (filter) metodami[10]. Při výběru



Obr. 2. **Schematické znázornění souřadnicové (a) a síťové (b) reprezentace chemického prostoru.** Body zde zastupují chemické struktury. Osy v souřadnicové reprezentaci chemického prostoru znázorňují arbitrární počet kvantitativních vlastností (deskriptorů) molekul, podle kterých je možno chemické struktury promítat do prostoru. Spojnice mezi molekulami v síťové (b) reprezentaci chemického prostoru (hrany grafu) značí jejich významný vztah



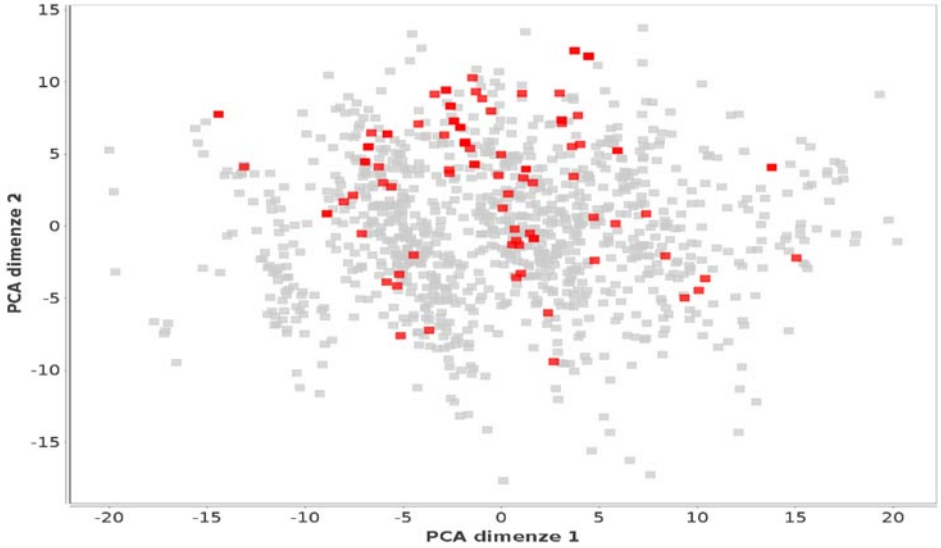
Obr. 3. **Jednoduchá dvourozměrná reprezentace části chemického prostoru.** Vyobrazeno je 100 inhibitorů androgenního receptoru (zvýrazněné čtverce) mezi 1000 náhodně vybranými komerčně dostupnými látkami. Vynesenými veličinami jsou molární hmotnost a rozdělovací koeficient oktanol-voda, obě jsou asociované s Lipinského pravidlem pěti pro biologickou dostupnost látek při perorálním podání[12]

vhodných deskriptorů se však přihlíží i k jejich chemickému významu. Jedním z takto významných deskriptorů je polární povrch molekuly (polar surface area; PSA), který silně souvisí se schopností molekuly překonávat buněčné membrány, a navíc je poměrně dobře odhadnutelný přímo ze strukturního vzorce molekuly jako tzv. „topological polar surface area“ (TPSA)[9]. PSA je tedy z farmakokinetického hlediska zvláště zajímavý deskriptor, a je proto přednostně zahrnován do modelů a vizualizací spojených s návrhem léčiv či jiných biologicky aktivních látek. Další podobně farmakokineticky významné deskriptory jsou např. molekulová hmotnost, rozdělovací koeficient oktanol-voda (obr. 3), molární refraktivita, počet donorů a akceptorů vodíkové vazby či počet rotovatelných vazeb. Existují také celé předdefinované množiny vzájemně se doplňujících deskriptorů, sestavené za účelem zjednodušení interpretace na nich založených vizualizacích a modelech. Mezi ně patří např. tzv. „molecular quantum numbers“ (MQN), což je 42 celočíselných deskriptorů reflektujících počty různých typů atomů, vazeb, kruhů a nábojů v charakterizované molekule[11].

Reprezentace chemického prostoru o řádově desítkách dimenzí, tedy i zmíněný 42-dimenzionální prostor MQN, je již možno výpočetně zpracovávat a vytvářet v nich predikční modely. Problémem však zůstává vizualizace chemického prostoru, která by kvůli uživatelské čitelnosti neměla mít více než tři rozměry. Přímý výběr maximálně tří vlastností molekul zřídka poskytuje odpovídající reprezentaci chemického prostoru. Přistupuje se tedy k tvorbě nových dimenzí tvořených vhodnými kombinacemi reprezentovaných dimenzí. Zavedenou statistickou metodou pro tvorbu takovýchto nových dimenzí je analýza hlavních komponent (principal component analysis, PCA). V PCA jsou původní dimenze lineární kombinací převedeny na nové dimenze (tzv. hlavní komponenty) s cílem postihnout co největší množství informace obsažené v původním n-dimenzionálním prostoru pomocí co nejmenšího množství hlavních komponent. Prostřednictvím PCA je tedy možno pro danou množinu chemických struktur popsaných větším množstvím kvantitativních deskriptorů (např. 42 MQN) sestavit jejich nejvěrnější možnou dvou- či třídimenzionální vizualizaci (obr. 4).

V souřadnicové reprezentaci chemického prostoru se kromě vizualizací pohybuje i široká škála cheminformatických metod. Jednoduché filtry založené na mezních hodnotách kvantitativních rysů molekul, např. obecně známé Lipinského pravidlo pěti[12], je možno v souřadnicové reprezentaci chemického prostoru vnímat jako plochy rozdělující chemický prostor na filtru odpovídající a neodpovídající (z hlediska filtru vhodné a nevhodné) části.

Na mnoho cheminformatických metod lze v obecném kontextu chemického prostoru pohlížet jako na modely, které na základě definované výchozí množiny kvantitativních rysů molekul (známých souřadnic) v chemickém prostoru odhadují jejich další rysy (neznámé souřadnice). Odhadované rysy molekul mohou být téměř jakékoliv, od ryze fyzikálně-chemických metrik jako bod tání až po emergentní rysy jako prostupnost biologickými membránami či dokonce afinitu k danému biologickému receptoru (často farmakologickému cíli)[13].



Obr. 4. **Chemický prostor definovaný 42 celočíselnými deskriptory MQN zredukovaný do dvoudimenzionálního zobrazení prostřednictvím PCA.** Vyobrazena je stejná množina jako u obr. 3, tedy 100 inhibitorů androgenního receptoru (zvýrazněné čtverce) mezi 1000 náhodně vybranými komerčně dostupnými látkami. Z farmakologického hlediska jsou zajímavé oblasti chemického prostoru s největším hustotou aktivních látek relativně k neaktivním, v této zjednodušené projekci by takovou oblast mohl vymezovat shluk aktivních látek v okolí [2;0]

4. Síťová reprezentace chemického prostoru

Alternativou k souřadnicové reprezentaci chemického prostoru je reprezentace chemického prostoru jako sítě (chemical space network, CSN), ve které jsou molekuly uspořádány nikoliv podle svých kvantitativních rysů, ale podle vzájemných relací[14]. Síťová reprezentace chemického prostoru umožňuje nasadit na chemickou problematiku aktuální metody a modely vycházející z teorie grafů, do jejichž vývoje a popisu bylo v posledních letech investováno mnoho prostředků vlivem rozvoje webové analytiky a sociálních sítí[14].

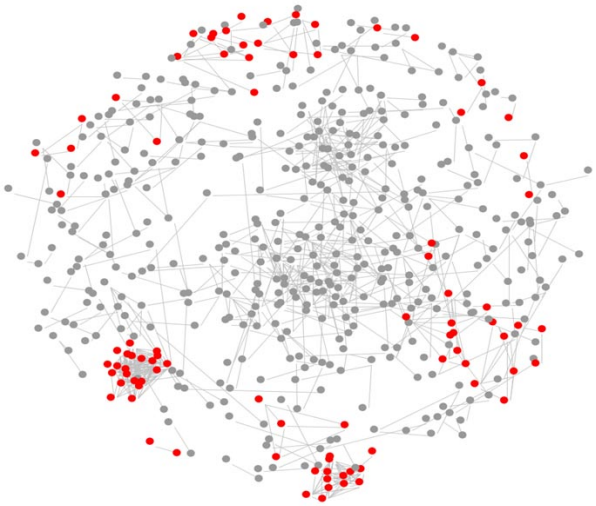
Vzdálenost dvojice molekul v síťové reprezentaci není určena rozdíly jejich kvantitativních rysů, jak tomu je u souřadnicové reprezentace, nýbrž charakterem a/nebo mírou jejich vztahu (obr. 2b). Velmi často používaným vztahem je strukturní podobnost molekul, kterou lze popsat mnoha různými způsoby[9]. I široce využívanou funkci podobnostního vyhledávání struktur implementovanou v řadě chemických databází je možno interpretovat jako procházení síťové reprezentace prohledávané databáze, kde se v síti vzájemných podobností vybírají nejbližší sousedé zadané struktury.

Strukturní podobnost molekul samozřejmě není jediným používaným základem síťové reprezentace chemického prostoru. Molekuly je možno v síťové reprezentaci asociovat na základě libovolných kvantitativních i kvalitativních metrik. Příkladem v praxi používané metriky přímo nesouvisející s chemickou strukturou je asociace molekul založená na jejich sdílených emergentních vlastnostech, zvláště pak jejich biologické aktivitě. Takto definovaná síťová reprezentace chemického prostoru tedy spolu asociuje molekuly, které mají podobné interakce se specifickými farmakologickými cíli a/nebo biologickými systémy obecně. Veřejně dostupnou databází agregující takovéto pozorované i predikované interakce mezi chemickými látkami a bílkovinami je STITCH (cit.[15]). STITCH tedy popisuje agregované interakce prostřednictvím heterogenní síťové reprezentace chemického prostoru.

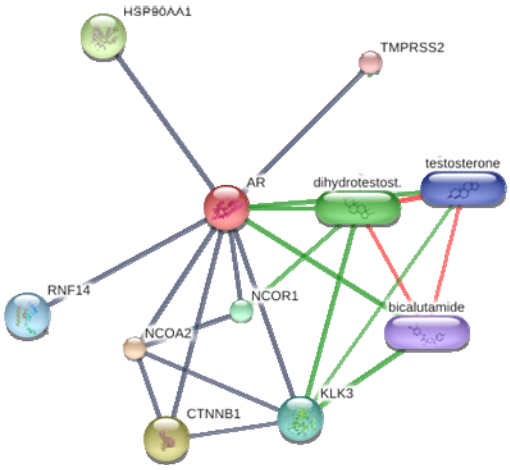
Vzájemné vztahy molekul v síťové reprezentaci chemického prostoru je možno přímo vizualizovat jako graf, ve kterém jsou jednotlivé molekuly znázorněny jako vrcholy a jejich vztahy jako hrany grafu (obr. 5). Hrany grafu často reprezentují pouze vztahy, jejichž velikost přesahuje zvolenou mezní hodnotu[14]. Vrcholy i hrany takto vizualizovaného grafu mohou svou zobrazovanou tloušťkou nebo kolorací nést dodatečnou informaci o samotných látkách a o síle či charakteru vztahů mezi nimi (obr. 6).

5. Významné oblasti chemického prostoru

Objev farmakologicky či jinak zajímavých oblastí, jejich omezení prostřednictvím strukturní, souřadnicové nebo síťové reprezentace a následné vzorkování takto vymezených oblastí je konceptuálním základem mnoha cheminformatických postupů, zejména pak virtuálního screeningu pro sestavování chemických knihoven zaměřených (focused) na konkrétní biologickou aktivitu[16]. Navrhované chemické knihovny musí často setrvávat v komerčně do-



Obr. 5. **Vizualizace vzájemných strukturních podobností ve stejné množině jako na obr. 3 a obr. 4, tedy inhibitorů androgenního receptoru (zvýrazněné body) mezi náhodně vybranými komerčně dostupnými látkami.** Jsou zde patrné ostře ohraničené oblasti strukturně podobných aktivních látek. Strukturní podobnost je zde definována jako Tanimotova podobnost strukturních fingerprintů s mezní hodnotou 0,75. V podobné reprezentaci chemického prostoru se pohybuje i algoritmus Molpher, který by zde bylo možno využít k bližšímu zmapování shluků aktivních látek. Podobnostní síť je dvourozměrně vizualizována metodou Kamada-Kawai



Obr. 6. **Vizualizace heterogenní síťové reprezentace chemického prostoru založené na vzájemných interakcích chemických látek a proteinů.** Tento graf byl generován databází STITCH[15] pro nejsilnější interakce lidského androgenního receptoru (AR). Síla hran grafu koresponduje se silou asociace jí spojených vrcholů, odstín odráží charakter interakce

stupné oblasti chemického prostoru, tedy v množině všech molekul dostupných v katalozích výrobců nebo dodavatelů. Setrvání v této části chemického prostoru je nutné, pokud do pracovního postupu nemůže být začleněna syntéza zcela nových látek z finančních, časových nebo organizačních důvodů. Jako množina komerčně dostupných látek je v mnoha pracovních postupech používána výše zmíněná databáze ZINC (cit.[6]).

Nadmnožinou komerčně dostupné oblasti chemického prostoru je tzv. synteticky dostupná oblast chemického prostoru. Jde o oblast chemického prostoru adekvátně dosažitelnou stávajícími syntetickými metodami. Tu je třeba zohlednit při počítačovém návrhu nových chemických struktur, které, ač stabilní, mohou být velmi náročné na syntézu. Nejhodnotnějším odhadem syntetické dostupnosti daných struktur je přirozeně odborné stanovisko samotných chemiků. Některé současné výpočetní metody však mohou navrhovat řádově desetitisíce struktur denně[17], jejichž ruční zpracování by představovalo neúnosnou časovou a finanční zátěž. Pro adresování tohoto problému jsou vyvíjeny výpočetní metody odhadu chemické dostupnosti struktur, které na základě v nich obsažených strukturních fragmentů odhadují jejich syntetickou dostupnost jako numerické skóre[18].

Z hlediska vývoje léčiv je velmi zajímavá oblast chemického prostoru obsahující látky biologicky dostupné při ústním podání, jejíž nejznámější jednoduchou aproximací je již výše popsané Lipinského pravidlo pěti[12]. Tato definice oblasti molekul biologicky dostupných při perorálním podání vyniká jednoduchostí aplikace a interpretace, svým velmi zobecňujícím charakterem však nevyhnutelně opomíjí mnoho dalších faktorů souvisejících s biologickou dostupností. Jeho příliš striktní aplikace může vést k promarněným farmakologickým příležitostem, což dokazuje další oblast chemického prostoru hodná zvláštní pozornosti – oblast přírodních látek. Přírodní látky jsou často biologicky aktivní i dostupné, ač nesplňují Lipinského pravidla[19].

Objev a vymezení oblastí chemického prostoru se silným zastoupením molekul zcela nevhodných pro dané použití má také hodnotu v podobě prostředků, které je možno ušetřit časným vyřazením těchto oblastí z pracovního postupu. Pro účely farmakologického výzkumu jsou takovéto nevhodné oblasti typicky asociované s vysokou toxicitou. Důvody vyřazování oblastí chemického prostoru však nemusí spočívat jen ve fyzikálně-chemických vlastnostech nebo nežádoucí biologické aktivitě v nich obsažených molekul. Zejména vyhýbání se patenty chráněným oblastem chemického prostoru je často nevyhnutelnou součástí farmakochemické praxe.

Oblast chemického prostoru definovanou ve strukturní, souřadnicové nebo síťové reprezentaci bývá v některých případech vhodné převést na množinu konkrétních struktur. Pokud je např. při vývoji nových léčiv identifiko-

vána oblast se zvýšenou hustotou molekul s požadovanou biologickou aktivitou, je pro ni třeba sestavit reprezentativní množinu struktur dostatečně malou na to, aby bylo možno ji posunout do náročnějších kroků pracovního postupu. Tato množina musí identifikovanou oblast chemického prostoru reprezentovat co možná nejvěrněji omezeným počtem struktur. Reprezentativní množinu je možno snadno sestavit náhodným vzorkováním, lepších výsledků lze však dosáhnout použitím vzorkování maximalizujícího rozmanitost vybrané reprezentativní množiny (diversity sampling).

Metoda tvorby výchozí množiny ke vzorkování se odvíjí od používané reprezentace chemického prostoru. Explicitní výčty struktur jako chemické databáze či jejich části lze vzorkovat přímo. Ze strukturních reprezentací oblastí chemického prostoru využívajících Markushovy vzorce nebo strukturní fragmenty je možno výchozí množinu struktur vytvářet kombinatoricky. Ze souřadnicově a síťově reprezentovaných oblastí chemického prostoru je výchozí množinu struktur nutno generovat algoritmicky. Níže popsané metody algoritmického procházení chemického prostoru jsou jedním z možných přístupů k tvorbě výchozí množiny struktur ze specifické oblasti chemického prostoru.

6. Pohyb v chemickém prostoru

Chemický prostor je ze své podstaty nespojitý a pohyb v něm lze provést pouze přechodem z jedné stálé chemické struktury na jinou. Jediným skutečným způsobem pohybu v chemickém prostoru jsou tak chemické reakce. Přechod reaktantů na produkty znamená strukturní změnu, která se v chemickém prostoru projeví jako změna pozice, tedy pohyb. Možnosti reálného pohybu v chemickém prostoru jsou tudíž limitovány přímo uskutečnitelností odpovídajících chemických reakcí. Náročnost jednotlivých reakcí v praxi velmi kolísá a návrh vhodného sledu posunů chemickým prostorem do zamýšleného místa (konečného produktu) je pro chemiky do značné míry intuitivním procesem, který je obtížné algoritmicky uchopit. Přesto však již existují automatizované služby, které na základě dat o komerčně dostupných látkách a zdokumentovaných reakcích asistují chemikům především při tzv. retrosyntetické analýze, tedy hledání vhodného postupu syntézy zamýšlené molekuly z komerčně dostupných látek[20]. V perspektivě chemického prostoru je retrosyntetická analýza snahou najít co nejjednodušší cestu odkudkoliv z komerčně dostupného chemického prostoru do jednoho cílového bodu v synteticky dostupné oblasti chemického prostoru.

Kontrastem k retrosyntetické analýze je syntéza nových látek cílená na maximalizaci jejich chemické rozmanitosti (diversity-oriented synthesis, DOS)[21]. Úkolem DOS je co nejlépe využít limitovaných výchozích látek (bodů) v komerčně dostupném chemickém prostoru tak, aby produkty jejich vzájemných reakcí rovnoměrně pokryly předdefinovanou zájmovou oblast chemického prostoru. Pokrytí zájmové oblasti je možno vizuálně reprezentovat a navádět prostřednictvím výpočetních metod (např. PCA) zvolených podle používané definice chemické rozmanitosti. Tento postup se uplatňuje především při tvorbě nových chemických knihoven[21].

V chemickém prostoru se lze pohybovat i zcela virtuálně, tedy provádět strukturní změny, které neodpovídají žádné uskutečnitelné chemické reakci. Tento přístup obětuje syntetickou interpretovatelnost reálného pohybu v chemickém prostoru za umožnění jednoduššího a svobodnějšího procházení zájmových oblastí chemického prostoru. Procházení chemickým prostorem zde bývá realizováno algoritmicky řízenými strukturními změnami jako přidání či odebrání atomu nebo tvorba, zánik a přesun chemických vazeb.

Jedním z novějších algoritmů pro takovýto *de novo* průzkum chemického prostoru je DAECS (Design Algorithm for Exploring Chemical Space)[22]. DAECS se pohybuje v mnohorozměrné souřadnicové reprezentaci chemického prostoru, jehož iterativní procházení je řízeno přímo uživatelem, který za pomoci průběžně aktualizované dvourozměrné projekce zpracovávané části chemického prostoru v každém kroku vybírá struktury (body v chemickém prostoru), jejichž okolí chce prozkoumat. DAECS potom z vybraných struktur prostřednictvím malých strukturních modifikací vytváří nové struktury, které v dalším kroku začleňuje do uživateli zobrazované projekce chemického prostoru.

Odlišně zaměřenou metodou algoritmického průzkumu chemického prostoru je evoluční algoritmus Molpher, který iterativně prochází chemický prostor mezi zadanými dvojicemi chemických struktur[17]. Pohybuje se převážně v síťové reprezentaci chemického prostoru (obr. 2b, obr. 5) a pokouší se inkrementálními strukturními modifikacemi první z dvojice vstupních struktur dosáhnout chemické struktury druhé, čímž je spojí „cestou“ chemickým prostorem tvořenou sledem nových chemických mezistruktur (tzv. „morfů“). Morfy nesou v různé míře strukturní rysy počáteční i cílové struktury.

Hlavní motivací algoritmického procházení chemického prostoru je návrh zcela nových chemických struktur ležících v již identifikované zájmové oblasti chemického prostoru[23]. Automaticky generované struktury mohou být prohibitivně náročné na syntézu[18], mnoho algoritmů na procházení chemického prostoru má proto kontrolu syntetické dostupnosti generovaných struktur integrovanou přímo ve svém pracovním procesu[17]. Takto validované chemické struktury je možno dále zpracovat virtuálním screeningem[16] a vytvořit z nich chemickou knihovnu zaměřenou (focused) na zájmovou oblast chemického prostoru.

7. Závěr

Koncept chemického prostoru je možno považovat za právně kodifikovaný již od raných let minulého století, kdy byl uznán původní Markushův patent. Kontakt chemiků s abstrakcí zde nazývanou jako chemický prostor však

nesetrvává jen v oblasti ochrany duševního vlastnictví: vyhledávání v chemické databázi, výpočetní modely, chemickou rozmanitost i vlastní chemickou syntézu je také možno vnímat v širším kontextu chemického prostoru.

Odhady velikosti chemického prostoru naznačují, že byl dosud připraven a charakterizován jen velmi malý zlomek všech možných chemických struktur. Chemický prostor tedy zůstává převážně neprozkoumán a lze očekávat, že stále obsahuje nezměrné množství molekul s dosud neobjevenými užitečnými vlastnostmi. Z hlediska chemického prostoru tak chemie neztratila za dobu své existence nic ze svého převratného potenciálu.

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

Seznam symbolů

PSA	polární povrch molekuly (polar surface area)
TPSA	topologický PSA
PCA	analýza hlavních komponent (principal component analysis)
MQN	„molecular quantum numbers“
CSN	síťová reprezentace chemického prostoru (chemical space network)
DOS	syntéza zaměřená na chemickou rozmanitost (diversity-oriented synthesis)
DAECS	algoritmus pro průzkum chemického prostoru (design algorithm for exploring chemical space)

LITERATURA

1. Dobson C. M.: Nature *432*, 824 (2004).
2. Polishchuk P. G., Madzhidov T. I., Varnek A.: J Comput Aided Mol Des *27*, 675 (2013).
3. Ertl, P.: J Chem Inf Comput Sci *43*, 374 (2003).
4. Walters W. P., Stahl M. T., Murcko M. A.: Drug Discov Today *3*, 374 (1998).
5. Ruddigkeit L., van Deursen R., Blum L. C., Reymond J. L.: J Chem Inf Model *52*, 2864 (2012).
6. Irwin J. J., Sterling T., Mysinger M. M., Bolstad E. S., Coleman R. G.: J Chem Inf Model *52*, 1757 (2012). http://zinc15.docking.org, staženo 26.9.2017.
7. Šilhánek J., Benešová E., Kačer P.: Chem. Listy *110*, 885 (2016).
8. Zhi L., van Oeveren C. A., Chen J., Higuchi R. I.: US 12/661,610 (2010).
9. Novotný J., Svozil D.: Chem. Listy. *111*, 716 (2017).
10. Guyon I., Elisseeff A.: J. Mach. Learn. Res. 3, 1157 (2003).
11. Nguyen K. T., Blum L. C., van Deursen R., Reymond J. L.: ChemMedChem *4*, 1803 (2009).
12. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J.: Adv Drug Deliv Rev *46*, 3 (2001).
13. Škuta C., Svozil D.: Chem. Listy. *111*, 747 (2017).
14. Maggiora G. M., Bajorath J.: J Comput Aided Mol Des *28*, 795 (2014).
15. Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., von Mering, C., Jensen, L. J., Beyer, A., Bork, P.: Nucleic Acids Res *38*, D552 (2010). http://stitch.embl.de, staženo 26.9.2017.
16. Svozil D.: Chem. Listy. *111*, 738 (2017).
17. Hoksza D., Škoda P., Voršilák M., Svozil D.: J Cheminform *6*, 7 (2014). https://www.assembla.com/spaces/molpher, staženo 26.9.2017.
18. Voršilák M., Svozil D.: Chem. Listy. *111*, 760 (2017).
19. Doak B. C., Over B., Giordanetto F., Kihlberg J.: Chem Biol *21*, 1115 (2014).
20. Corey E. J.: Chem. Soc. Rev. *17*, 111 (1988).
21. Thomas G. L., Wyatt E. E., Spring D. R.: Curr Opin Drug Discov Devel *9*, 700 (2006).
22. Takeda S., Kaneko H., Funatsu K.: J. Chem. Inf. Model. *56*, 1885 (2016).
23. van Deursen R., Reymond J. L.: ChemMedChem *2*, 636 (2007).

I. Čmelo[a,b] and D. Svozil[a,b] *([a] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, Academy of the Sciences of the Czech Republic, [b] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, University of Chemistry and Technology, Prague):* **Chemical Space**

The chemical space is a concept used primarily in cheminformatics, but it indirectly relates to all fields that deal with chemical structures, from general chemistry up to patent law. It is a theoretical space consisting of all energetically stable (and thus possible) isomers of all chemical structures. This space can be arranged by any number of various criteria to create a projection, within which interesting areas can subsequently be searched, delimited and sampled. In practice, such interesting areas are usually those associated with desired biological activity, synthetic accessibility or patent coverage of contained structures. This article introduces the concept of chemical space in an interdisciplinary context and describes the commonly used forms of its representation.