THÈSE présentée par :

Yuliana ZABOLOTNA

Soutenue le : **17 septembre 2021**

Pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : **Chimie informatique et théorique**

Exploration par chémographie des espaces chimiques ultra-larges pour la chimie médicinale

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. HORVATH Dragos	Directeur de recherche CNRS, Université de Strasbourg

RAPPORTEURS :

Mme. DOUGUET Dominique	Chargée de recherche Inserm (HDR), Université Côte d'Azur
M. TABOUREAU Olivier	Professeur, Université de Paris

AUTRES MEMBRES DU JURY :

M. ERTL Peter	Docteur, Instituts Novartis pour la recherche biomédicale
M. VOLOCHNYUK Dmitriy	Professeur, Académie nationale des sciences d'Ukraine

Abstract

This thesis is dedicated to the detailed GTM-based analysis of the chemical space of ultra-large libraries and development of the online tool for navigation through up to billions of compounds, called ChemSpace Atlas. The efficiency and polyfunctionality of GTM allowed producing a detailed picture of the chemical space currently available to medicinal chemists. Fragment-, lead-, drug-, PPI- and NP-like compounds, genuine NPs, purchasable building blocks, and DNA-encoded libraries were systematically analyzed using hierarchical GTM. The resulting tens of thousands of maps were employed as the main basis of the ChemSpace Atlas. This tool enables efficient exploration of the ultra-large chemical space from different perspectives: chemotypes, various physicochemical properties, biological activities, etc. Moreover, the hierarchy of maps provides multiple levels of detalization: from a global bird's eye view of the whole dataset on the universal map to the structural pattern detection in separate areas of the region-dedicated zoomed maps.

Acknowledgements

I would like to express my sincere gratitude to all my colleagues from the Laboratory of Chemoinformatics at the University of Strasbourg - it has been a privilege to work with such talented and creative scientists. In particular, I am extremely grateful to my supervisors Prof. Alexandre Varnek and Dr. Dragos Horvath, for their patient guidance, thoughtful suggestions, and their invaluable experience that they have shared with me. Their continuous support and encouragement kept me motivated even during the most challenging times. My heartfelt thanks go to Prof. Dmitriy Volochnyuk and Dr. Peter Ertl for finding the time to guide me through our collaborative projects, providing me with helpful feedback, and their professional expertise in medicinal chemistry, chemistry of natural products and chemoinformatics. I would like to express my deep appreciation to Dr. Arkadii Lin for the programming training, insightful comments, friendly support and valuable advices that significantly improved this work. Special thanks also go to Dr. Fanny Bonachera for developing the web interface of ChemSpace Atlas that provides users with access to the results of this thesis. I am grateful to Dr. Gilles Marcou for mathematical and methodological help and productive discussions and Dr. Olga Klimchuk for the constant support in all areas during these three years. I would like to show my gratitude to Dr. Iuri Casciuc, Dr. Alexey Orlov, Dr. Timur Madzhidov, Regina Pikalyova, Karina Pikalyova, Shamkhal Baybekov, Maxim Shevelev, William Bort, Tagir Akhmetshin and all other (former and current) lab members for their friendship, support, and our scientific discussions.

I would also like to express my sincere appreciation to all my teachers from the University of Strasbourg and Taras Shevchenko National University of Kyiv, who provided me with a solid theoretical basis for the scientific research.

Direction de la **recherche**
et de la **valorisation**
Collège **doctoral**
Université de Strasbourg

Déclaration sur l'honneur
Declaration of Honour

J'affirme être informé que le plagiat est une faute grave susceptible de mener à des sanctions administratives et disciplinaires pouvant aller jusqu'au renvoi de l'Université de Strasbourg et passible de poursuites devant les tribunaux de la République Française.

Je suis conscient(e) que l'absence de citation claire et transparente d'une source empruntée à un tiers (texte, idée, raisonnement ou autre création) est constitutive de plagiat.

Au vu de ce qui précède, **j'atteste sur l'honneur que le travail décrit dans mon manuscrit de thèse est un travail original et que je n'ai pas eu recours au plagiat ou à toute autre forme de fraude.**

I affirm that I am aware that plagiarism is a serious misconduct that may lead to administrative and disciplinary sanctions up to dismissal from the University of Strasbourg and liable to prosecution in the courts of the French Republic.

I am aware that the absence of a clear and transparent citation of a source borrowed from a third party (text, idea, reasoning or other creation) is constitutive of plagiarism.

*In view of the foregoing, **I hereby certify that the work described in my thesis manuscript is original work and that I have not resorted to plagiarism or any other form of fraud.***

Nom :   ZABOLOTNA   **Prénom :**   Yuliana

Ecole doctorale :   ED222

Laboratoire :   Laboratoire de Chémoinformatique, UMR 7140 CNRS

Date :   22/09/2021

Signature :

Contents

1   Résumé en français

1.1   Introduction

L'ère des mégadonnées en chimie médicinale est marquée par une explosion des nouvelles informations chimiques et biologiques rapportées quotidiennement[1]. Les nouvelles informations sont désormais produites à une vitesse supérieure à celle à laquelle elles peuvent être analysées et interprétées par les acteurs humains sur le terrain. Par conséquent, il existe un besoin urgent d'outils de calcul efficaces et compatibles avec les mégadonnées pour l'exploration de l'espace chimique des très grandes chimiothèques. Cette exploration devrait inclure la visualisation interactive, la diversité, l'analyse des propriétés et des chémotypes, la comparaison des chimiothèques, la prédiction *in silico* de l'activité et ADMETox, etc.

La cartographie topographique générative, ou GTM, répond parfaitement à toutes ces exigences. La GTM est une méthode de réduction de dimensionnalité qui convertit les composés depuis l'espace initial des descripteurs multidimensionnels vers un espace latent 2D, appelé carte 2D[2]. Contrairement aux autres méthodes de chémographie, la GTM distribue les projections des molécules sur la carte avec des probabilités spécifiques aux nœuds (responsabilités) au lieu d'attribuer sans ambiguïté chaque composé à un seul point de la carte. Cette fluidité permet la création de paysages GTM - des cartes, colorées par des valeurs moyennes de différentes propriétés, e. g. densité, activité biologique, classe assignée, etc. Pourtant, les cartes 2D ne peuvent pas accueillir un grand nombre de composés tout en capturant de fines différences entre des voisins proches. La Hierarchical GTM (hGTM) [3, 4], alias «Zooming» est une technique qui entraîne une nouvelle carte sur un ensemble de composés extraits d'une zone donnée sur la carte mère, afin d'assurer une cartographie localement optimale. L'empilement hiérarchique des GTM, de la carte générale

«universelle» aux cartes détaillées des clusters locaux rend cette stratégie compatible «mégadonnées».

Ainsi, cette thèse est dédiée au développement de «ChemSpace Atlas» - un outil polyfonctionnel qui permet de naviguer et d'analyser l'espace chimique de très grandes chimiothèques pour la chimie médicinale. Il est basé sur des dizaines de milliers de GTM organisés hiérarchiquement qui permettent une visualisation significative et une navigation facile à travers les centaines de millions de composés, d'une vue globale à vol d'oiseau à la détection de motifs structurels.

1.2 Cartes universelles de l'espace biologiquement pertinent

Les GTM universelles peuvent être définies comme des cartes du «meilleur compromis», offrant des performances prédictives satisfaisantes par rapport à des propriétés biologiques très diverses[5]. Sept cartes universelles de l'espace chimique de ChEMBL, définies par des descripteurs de fragments ISIDA, ont été évoluées par un algorithme génétique (GA) dans l'espace des paramètres de la carte comme degrés de liberté clés (y compris le choix du descripteur, la taille de la grille, les contrôles de flexibilité multiples, etc.). Une performance prédictive moyenne sur des centaines d'activités biologiques a été utilisée comme fonction objectif dans la recherche des meilleurs paramètres GTM pour sept cartes universelles. Il a été prouvé que ces GTM servent avec succès d'hôtes pour 618 paysages d'activités, associés à des séries de composés ChEMBL ayant des structures-activités spécifiques aux cibles respectives (**Figure 1**). En considérant que ChEMBL couvre une majeure partie des données structure-activité disponibles publiquement, les cartes universelles distinctes construites à l'aide de cette chimiothèque représentent des vues complémentaires et fortement synergiques de l'espace chimique biologiquement pertinent. Elles peuvent être utilisées non seulement comme outil prédictif, mais aussi comme cadre d'analyse de grandes chimiothèques chimiques dans le contexte de la chimie médicinale et de la conception de médicaments.

Dans cette thèse, les sept cartes universelles ont été combinées dans un modèle prédictif consensus pour le profilage de la bioactivité. La première carte universelle a été utilisée pour l'analyse de l'espace chimique défini par des composés biologiquement testés de ChEMBL (1,4M), des molécules disponibles dans le commerce pour le criblage à haut débit provenant de la base de données ZINC (presque 1B) et des chimiothèques codées par ADN (2.5B) énumérées à l'aide de BB commercialement disponibles. Cependant, en raison

du fait qu'il existe un nombre limité de NP dans ChEMBL, une NP-umap spécifique a été construite en utilisant des composés de la collection de NP COCONUT. De même, une carte universelle dédiée des synthons a été créée (sans prendre en compte les groupes partants dans les réactifs réels). Cette carte a été formée sur des synthons générés à la fois à partir de réactifs disponibles dans le commerce et de composés ChEMBL (via leur fragmentation).
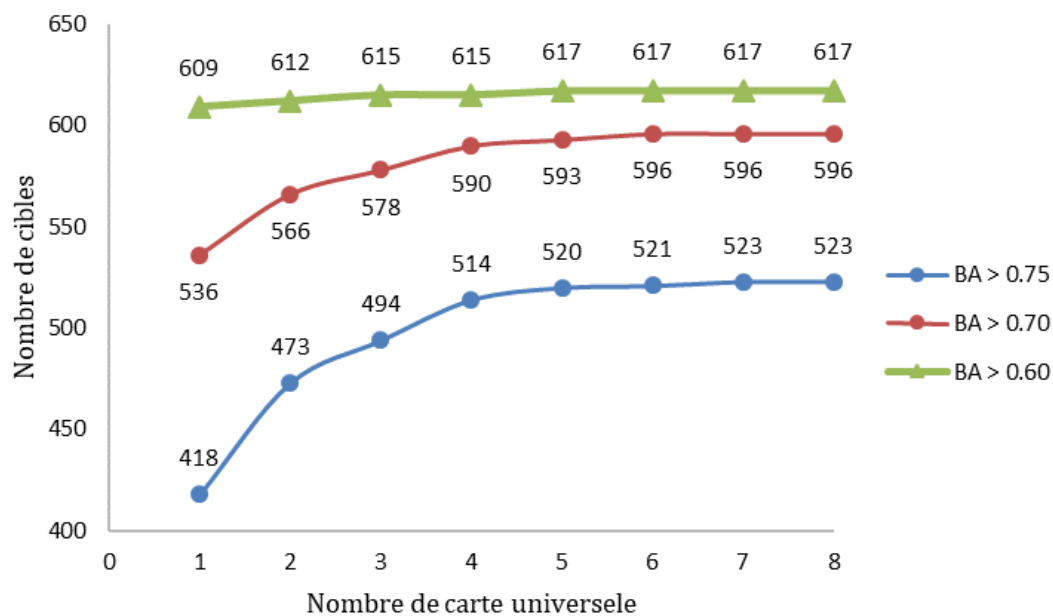

Figure 1. Performance cumulée des cartes universelles exprimée en nombre des activités prédites avec un BA supérieur au seuil établi vs nombre de cartes utilisées.

1.3   Exploration et analyse d'espaces chimiques ultra-larges

1.3.1   Chimiothèques de criblage

De nos jours, les composés disponibles dans le commerce sont l'une des principales sources de médicaments potentiels. Cependant, l'espace chimique actuellement connu est loin d'être entièrement étudié et appréhendé par les chimistes médicinaux. En essayant de combler cette lacune, nous avons comparé près d'un milliard de composés commercialement disponibles de la chimiothèque ZINC avec 1,4 million de molécules biologiquement testées de ChEMBL en utilisant la GTM hiérarchique (hGTM). En fonction de la stratégie d'identification de hits choisie, les composés ZINC et ChEMBL ont été divisés en quatre groupes ou sous-familles: *fragment-like*[6, 7]*, lead-like*[8, 9]*, drug-like*[10, 11]*, and PPI-like*[12]. La disponibilité des molécules de ZINC a également été évalué : les composés commerciaux ont ensuite été divisés en sous-ensembles «ZINC-Real» et «ZINC-Tangible». Ce dernier concerne des composés non encore synthétisés mais pouvant être préparés sur demande avec un taux d'achetabilité de 80%.

Les paysages comparatifs entre l'espace chimique disponible dans le commerce et la bibliothèque de référence contenant des composés testés biologiquement permettent d'évaluer l'étendue de la pertinence biologique des bibliothèques achetables. Afin d'améliorer la résolution et le niveau de détail de cette analyse, le GTM hiérarchique (hGTM) a été utilisé pour atteindre les plus petits clusters dans l'espace chimique. La comparaison structurelle des composés ChEMBL et ZINC au dernier niveau de cette hiérarchie permet de détecter des caractéristiques précédemment cachées de chaque bibliothèque, d'identifier ce qui a été manqué par les fournisseurs de produits chimiques dans la course à l'amélioration de leurs catalogues et par les chimistes médicinaux au cours de l'exploration biologique expérimentale de l'espace chimique disponible.

Environ 40 000 cartes hiérarchiques de l'espace chimique ont été construites. L'utilisation de hGTM a permis de multiplier par 40 la taille des bibliothèques analysées par rapport aux rapports publiés précédemment (800M contre 20M analysés dans les travaux de Lin et al..[13]). L'analyse détaillée de l'espace chimique à cette échelle a permis de mieux comprendre les caractéristiques structurelles de l'espace chimique achetable ainsi que sa pertinence biologique.

Les motifs structurels inhérents à une seule chimiothèque ont été identifiés (**Figure 2**). En conséquence, il a été découvert qu'il manque de nombreuses familles de composés connues pour inclure des membres biologiquement actifs - des inhibiteurs très puissants de cibles biologiques importantes - dans les chimiothèques disponibles dans le commerce. Ces ≈20 000 familles de composés ChEMBL hors marché sont une motivation pour enrichir les catalogues commerciaux. Par ailleurs, 100 000 familles de composés spécifiques au ZINC sont en attente d'évaluation dans le cadre de programmes de recherche de dépistage.

	ChEMBL	ZINC		ChEMBL	ZINC
Zone1	9 262	72 984	Zone3	35	0
Zone2	384	3 846	Zone4	0	70
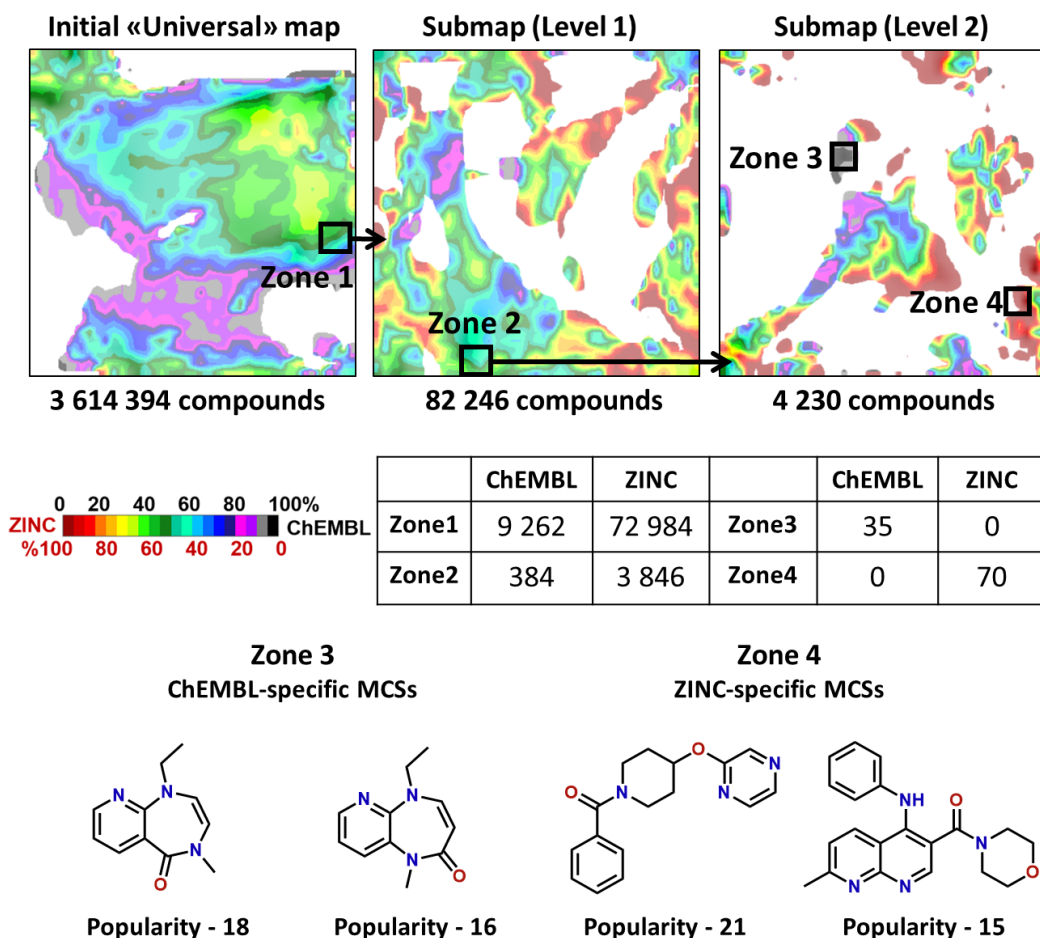
Figure 2. Navigation hGTM des zones fortement peuplées de l'espace chimique: exemple Lead-Like ChEMBL vs ZINC-Real. Le tableau fournit la composition de chaque zone en surbrillance. En partant de la zone mixte dense 1, en passant par les deux niveaux de zoom, de petites sous-zones purement ChEMBL (Zone3) et ZINC (Zone4) sont détectées. Les sous-structures communes maximales correspondantes (MCS) et leur popularité (nombre de composés contenant chaque fragment structurel) sont également signalées.

La faisabilité de la compilation d'un ensemble «idéal» de 1 million de composés diversifiés (50 000 échafaudages, minimum de 20 composés par échafaudage) a également été évaluée[14]. Une telle banque peut être très utile pour le criblage biologique primaire contre une nouvelle cible avec une structure inconnue, avec seulement quelques chémotypes actifs connus, ou sans modulateurs de petites molécules existants. Cependant, il est apparu qu'actuellement, il n'est pas possible de l'acheter même en combinant les catalogues de 33 vendeurs. En revanche, l'ensemble «idéal» de 500 000 peut être obtenu auprès de seulement six fournisseurs, avec un ensemble de 350 000 disponible auprès de seulement trois fournisseurs. Ces divers ensembles de données «idéales» ont été comparés à l'espace chimique biologiquement pertinent (chimiothèque ChEMBL) à l'aide de trois GTM universelles (**Figure 3**).
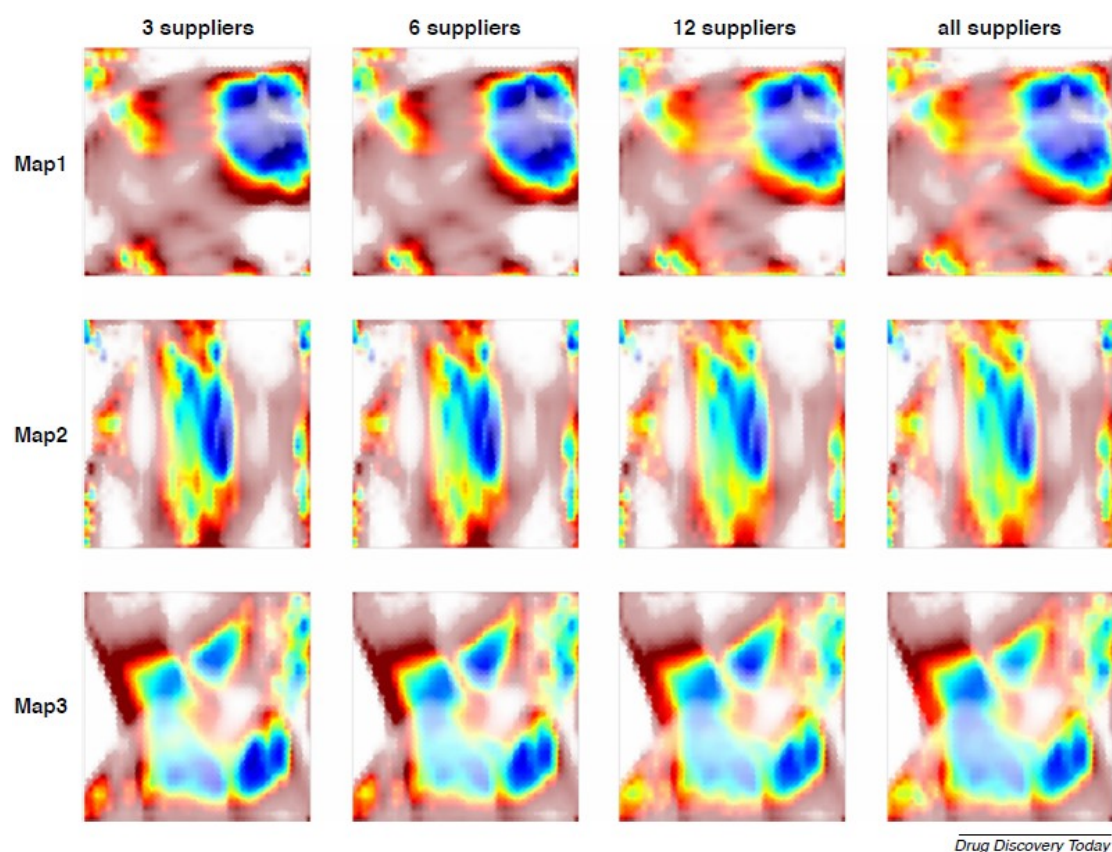
Figure 3. Cartes GTM de quatre ensembles "idéaux" de composés correspondant à trois, six, 12 et 33 fournisseurs (bleu) sur le fond de composés ChEMBL (rouge).

1.3.2 Chimiothèques codées par AND (DEL)

Outre les techniques classiques bien étudiées d'identification des hits, comme le HTS, plusieurs nouvelles méthodologies sont devenues disponibles récemment. L'une des plus prometteuses d'entre elles est la sélection par affinité avec les chimiothèques codées par ADN (DEL)[15]. Cette technologie est moins chère, plus rapide et parfois plus efficace - elle permet de cribler jusqu'à des milliards de composés à la fois. Cependant, il n'y a presque aucun rapport d'analyse chimio-informatique de l'espace chimique DEL.
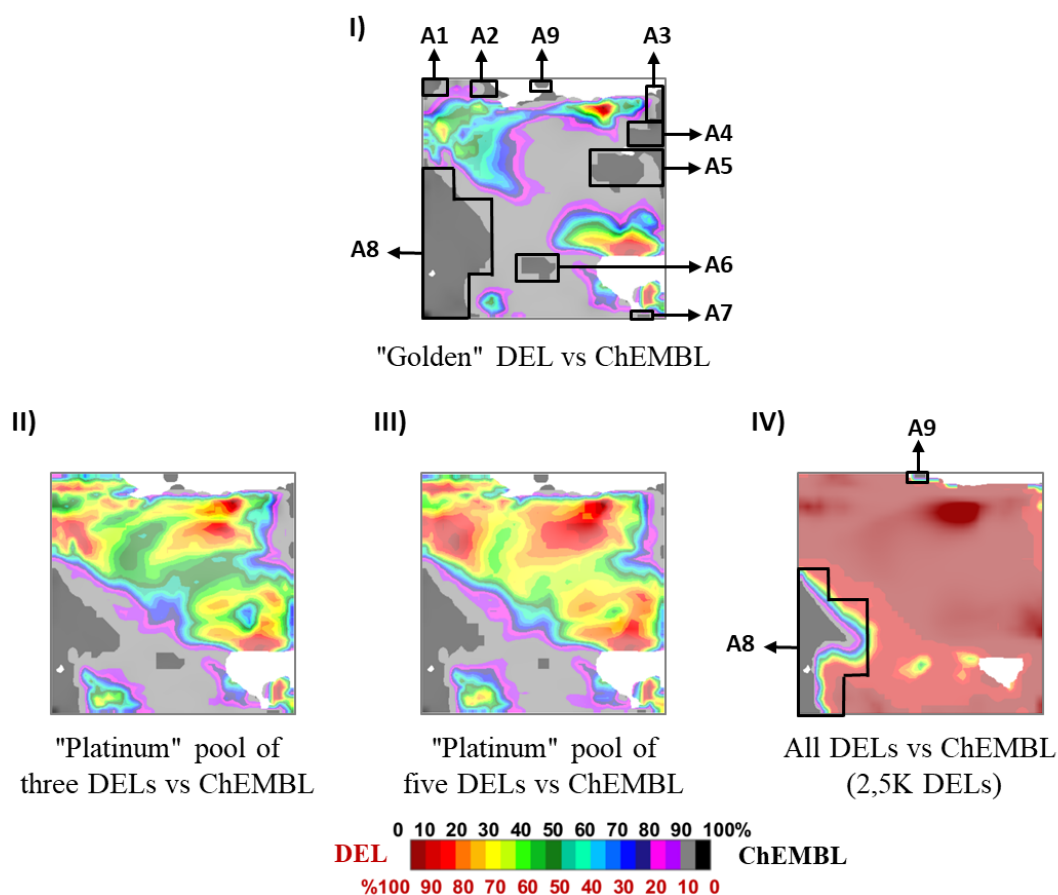
Figure 4. Comparaison de l'espace chimique des DEL (rouge) avec des composés biologiquement pertinents de ChEMBL (noir). La DEL «dorée» et les ensembles de 3 et 5 DEL ont été sélectionnés en maximisant la portion de ChEMBL couverte par ces chimiothèques.

A1: Thiophene-containing compounds
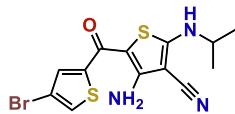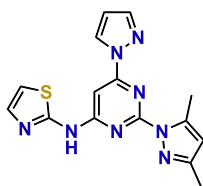
CHEMBL4454199

A2: Thiazoles and thiadiazoles

CHEMBL3100167

A3: Benzosulfonamides (with two or more PhSO2N groups)

CHEMBL1729230

A4: Sulfonamides

CHEMBL1346964

A5: Polyamides, ureas, and carbamates
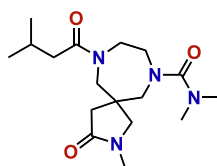
CHEMBL3444791

A6: Aromatic compounds with long conjugated systems
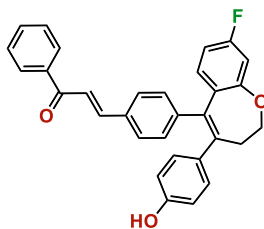
CHEMBL4225431

A7: Dihydropyrazoles and hydrazones with sulfonamide group

CHEMBL1950243

A8: Natural products and NP-like compounds
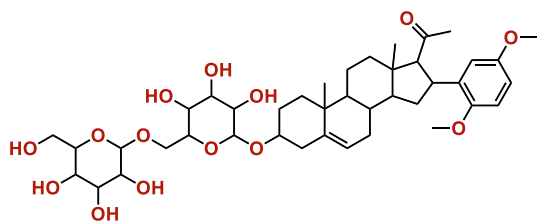
CHEMBL2096828

A9: Nucleotides
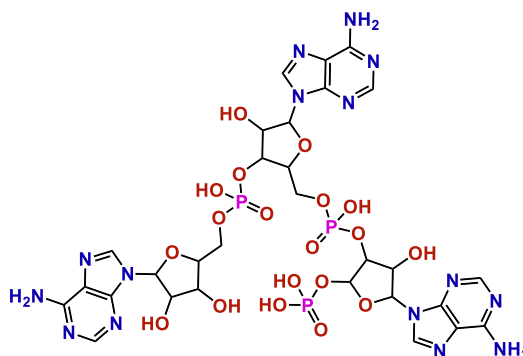
CHEMBL605454

Figure 5. Exemples de composés du CHEMBL peuplant les zones de A1 à A9 mis en évidence dans les paysages en **Figure 4**.

Dans ce projet, environ 2 500 DEL de différentes tailles (de 1M à 100M) ont été conçues à l'aide de BB disponibles dans le commerce. Un sous-ensemble représentatif de composés (1M pour chaque chimiothèque) a été généré, normalisé et projeté sur la première carte universelle. L'espace chimique de 2,5B DEL résultant a été comparé à des composés biologiquement pertinents de ChEMBL (**Figure 4**). Il semble qu'il existe plusieurs régions spécifiques à ChEMBL - des zones qui ne sont occupées par aucune DEL. Elles sont peuplées de NP complexes, comme des stéroïdes, des macrolides, des peptides, des

nucléotides, etc. (**Figure 5**). Ainsi, en général, la technologie DEL donne un accès à l'espace chimique biologiquement pertinent à l'exception tout à fait attendue des NP complexes.

Cependant, dans une campagne de dépistage, une seule DEL sera utilisée. Ainsi, une DEL «dorée» (ou un ensemble de quelques DEL complémentaires) qui fournit la couverture la plus élevée de l'espace chimique ChEMBL doit être trouvée. Avec l'aide de la GTM, il a été démontré qu'une seule chimiothèque peut couvrir environ 60% des composés ChEMBL. Dans le cas de 3 DEL complémentaires combinées, cette couverture augmente jusqu'à 72%, tandis que l'utilisation simultanée de 5 DEL fournit une couverture de 82%.

1.3.3 Building blocks

Comme la qualité et la diversité des composés de criblage dépendent inévitablement des BBs utilisés pour leur synthèse, leur sélection rationnelle peut considérablement améliorer le processus de conception des médicaments en se concentrant au préalable sur les sous-structures et les propriétés qui garantiront l'activité et le profil ADMETox souhaitables des candidats de médicaments potentiels.[16] Bien que ce fait soit largement reconnu par les chimistes médicinaux, le nombre de rapports scientifiques, ciblant l'analyse de la qualité des BBs existants achetables (PBB) et les stratégies potentielles pour l'amélioration des bibliothèques correspondantes, est significativement inférieur à celui pour les composés de criblage disponibles dans le commerce.

Ainsi, une analyse détaillée de l'espace chimique de 400K BB commercialement disponibles a été effectuée. L'espace chimique n'était pas défini par les BB eux-mêmes, mais plutôt par les synthons correspondants qui sont des incréments introduits dans la molécule finale lors de la réaction. Pour cela, une boîte à outils de réactions basée sur la connaissance, appelée Synthons Interpreter (SynthI), a été développée pour l'analyse et la conception de la chimiothèque. Elle se compose de quatre modules: SynthI-Classifier (classifie les BB), SynthI-BB (génère des synthons à partir des BB), SynthI-Fragmentation (fragmente des molécules plus grosses vers des synthons) et SynthI-Enumeration (combine plusieurs synthons en molécules plus grosses). Les synthons sont des incréments du BB qui seront ajoutés au composé final lors d'une réaction chimique particulière. Dans SynthI, les synthons sont utilisés comme représentation unifiée des BB et des fragments - ils sont générés non seulement à partir de réactifs, mais sont également le résultat de la fragmentation pseudo-rétrosynthétique de plus grandes molécules d'intérêt. Leur caractéristique distinctive est la

présence de marques spéciales à l'ancienne position des groupes partants. Le type de marque définit le type de centre de réaction - électrophile, nucléophile, radical, etc.

Dans la **Figure 6** on peut voir des exemples de classification et de synthonisation de BBs. Certaines des classes de BBs, par exemple les amines secondaires, ne produisent qu'un seul synthon par BB (**Figure 6 A**). D'autres, comme les cétones, peuvent donner lieu à de nombreux synthons en fonction des conditions de réaction (**Figure 6 C**). Un exemple de synthonisation d'aminoesters avec l'option keepPG est montré dans la **Figure 6 E**.

10 - Electrophilic RC   20 - Nucleophilic RC   30 - Bivalent electrophilic RC   40 - Bivalent nucleophilic RC
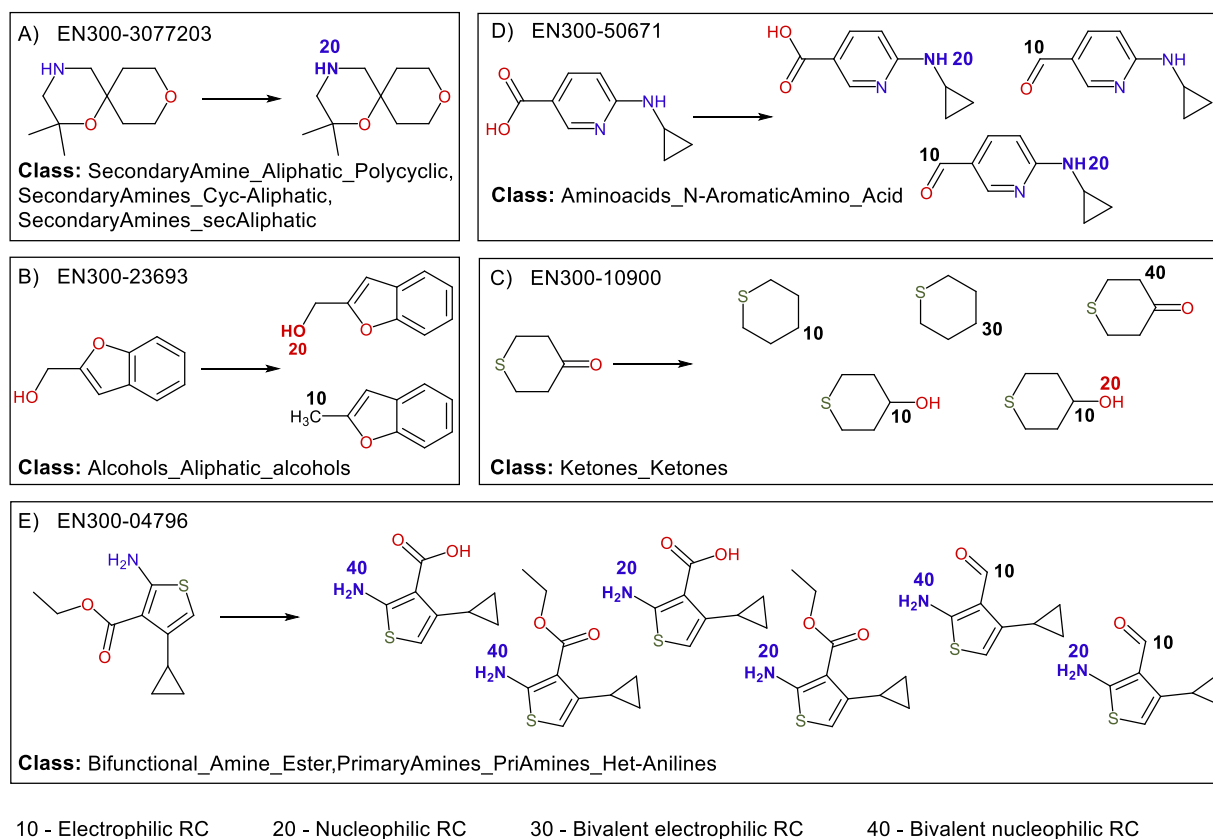
Figure 6. Exemples de classification et de synthonisation de BBs. Les étiquettes sur les synthons définissent la nature du centre de réaction (RC).

Les principales classes de BB ont été analysées en termes de disponibilité, de qualité définie par la règle de deux[16] et de diversité. La capacité des BB à faire face aux besoins de chimie médicinale a été évaluée par leur comparaison avec un ensemble de référence de synthons biologiquement pertinents, dérivés de la fragmentation ChEMBL avec l'aide de SynthI (**Figure 7**). Cette comparaison a été réalisée à l'aide d'une GTM universelle nouvellement construite sur l'espace chimique des synthons, qui permet de visualiser les deux chimiothèques en même temps et d'analyser leur chevauchement, ainsi que les régions spécifiques à la chimiothèque.
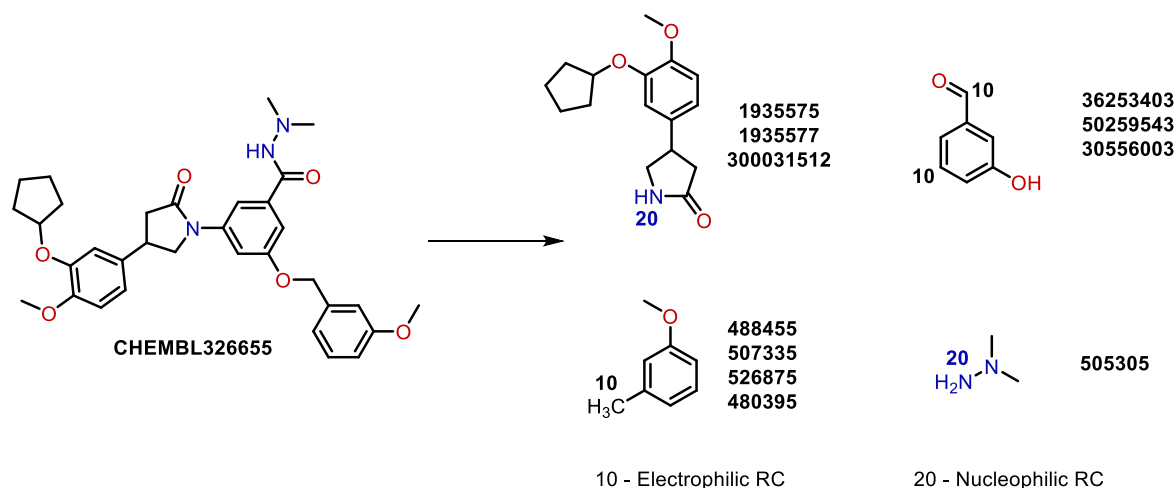
Figure 7. Exemple de fragmentation d'une molécule CHEMBL vers des synthons disponibles dans le commerce (les identifiants eMolecules des BBs correspondants sont fournis).

Dans la **Figure 8** on peut voir 16 paysages comparatifs ChEMBL vs BBs achetables pour différents groupes de synthons. Cette comparaison a permis d'identifier que seulement dans le cas de quatre classes de synthons, les synthons PBB couvrent largement l'espace chimique des synthons dérivés de ChEMBL : synthons pour la métathèse, agents d'acylation, O- et N-nucléophiles (**Figure 8 (a)**). Pour les autres groupes, même pour ceux ayant un fort excès de synthons PBB (**Figure 8 (b)**), il existe de nombreuses zones d'espace chimique spécifiques à ChEMBL sans aucun analogue achetable. La plupart de ces zones correspondent aux BBs polyfonctionnels sous-représentés sur le marché.

a. Synthons classes with the highest libraries overlap

Synthons for Metathesis

ChEMBL – 2 288
PBB – 7 325

O-Nucleophiles

ChEMBL – 28 807
PBB – 77 539

Acylation Agents

ChEMBL – 25 976
PBB – 115 805

N-Nucleophiles

ChEMBL – 89 583
PBB – 198 664

b. Synthons classes with the highest excess of PBB but the lowest libraries overlap

Bivalent C-electrophiles

ChEMBL – 13 059
PBB – 27 321

Bivalent C-nucleophiles

ChEMBL – 10 042
PBB – 18 679

Csp2Ar-Electrophiles

ChEMBL – 98 540
PBB – 122 753

Bivalent N-nucleophiles

ChEMBL – 3 708
PBB – 65 445

C-Alk-Electrophiles

ChEMBL – 57 372
PBB – 161 856

Electrophilic radicals

ChEMBL – 2 901
PBB – 108 794

c. Synthons classes with the highest excess of ChEMBL BB

C-Alk-Nucleophiles

ChEMBL – 63 814
PBB – 13 251

Csp2Ar-Nucleophiles

ChEMBL – 31 181
PBB – 1 794

Nucleophilic radicals

ChEMBL – 4 715
PBB – 701

S-Nucleophiles

ChEMBL – 4 893
PBB – 1 890

C-Nucleophiles Boronics

ChEMBL – 27 271
PBB – 12 774

d. Synthons class with balanced libraries representation

Sulfonylation Agents

ChEMBL – 4 433
PBB – 4 588

ChEMBL synthons 0 10 20 30 40 50 60 70 80 90 100% PBB synthons
%100 90 80 70 60 50 40 30 20 10 0

Figure 8. Comparaison des synthons PBBs (zones noires) et des synthons dérivés de ChEMBL (régions rouges) basée sur les classes de synthons.

1.3.4 Produits naturels

Étant conçues par évolution sur des millions d'années pour se lier aux macromolécules biologiques, les NP sont restées une source d'inspiration importante pour les chimistes médicinaux. Ainsi, l'espace chimique des NPs de la chimiothèque COCONUT et des composés NP-like de ZINC et ChEMBL ont également été analysés[17]. Plus de 200 hGTM basées sur la nouvelle carte universelle (NP-Umap - **Figure 9**) ont été construites. Il a été montré que l'ensemble de ces cartes fournit une séparation significative des chémotypes, qui peut être utilisée pour l'analyse structurale des NP et dans une recherche d'analogues naturels ou synthétiques d'une molécule d'intérêt. La comparaison des NP de COCONUT et des sous-ensembles de ZINC de type NP a abouti à près de 20 000 chémotypes uniques, spécifiques à une seule chimiothèque (**Figure 10**). 90% des familles de composés spécifiques du ZINC contiennent des N-hétérocycles. Concernant les composés spécifiques des NPs, la majorité d'entre eux correspondent aux glucides ou oxohétérocycles complexes avec des chaines latérales contenant de l'oxygène. Ceci illustre le fait bien connu que les composés contenant de l'azote sont mieux explorés par la chimie de synthèse que les NP contenant de l'oxygène complexes.
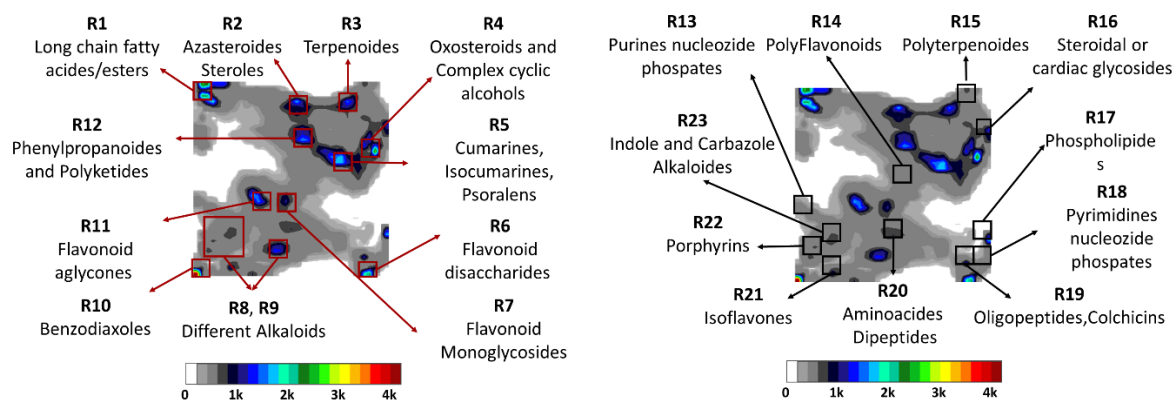
Figure 9. Paysage de densité des NP de COCONUT. A gauche - chémotypes pour les régions fortement peuplées, à droite - pour les régions peu peuplées. Les zones multicolores correspondent aux régions très peuplées, tandis que la couleur grise définit les zones moyennement occupées. Les zones blanches sont vides.
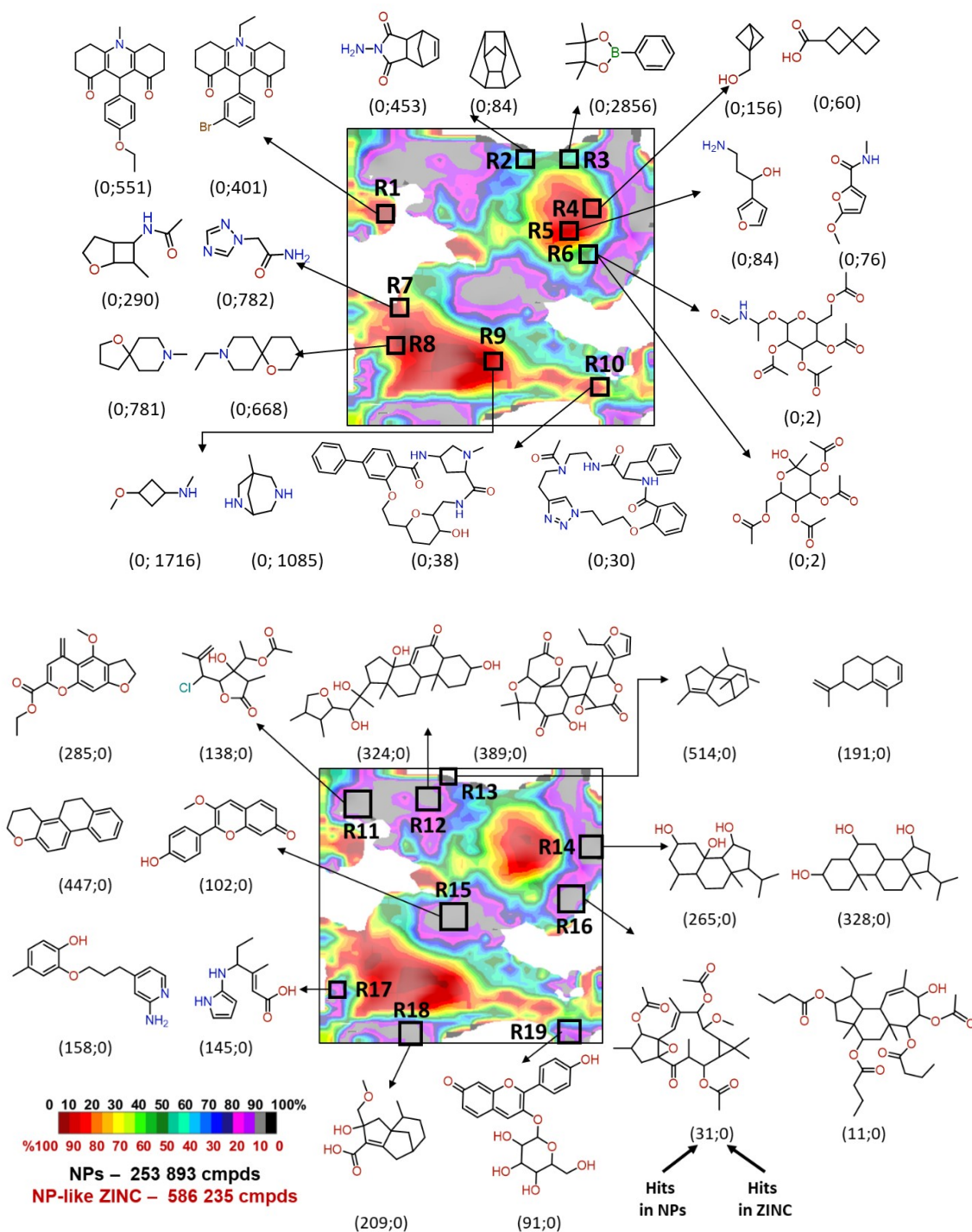
Figure 10. Paysage de classes comparant les produits naturels de COCONUT (noir) aux composés de ZINC NP-like (rouge). Le schéma en haut fournit des exemples de MCS spécifiques à ZINC, tandis que celui en bas démontre les MCS spécifiques à NP. Le premier chiffre entre parenthèses indique le nombre de hits dans c-COCONUT, le deuxième - dans NP-like ZINC.

1.4 ChemSpace Atlas - un outil pour l'exploration efficace de l'espace chimique

ChemSpace Atlas est un outil polyvalent intuitif pour l'exploration efficace de l'espace chimique ultra-large et son analyse par rapport aux problèmes de chimie médicinale. Il est basé sur des dizaines de milliers de GTM, construites dans des projets précédemment décrits et peut être séparé en plusieurs chapitres en fonction des sous-espaces chimiques mis au point: criblage de composés (fragment-like, lead-like, drug-like et PPI-like), DEL, NP et synthons de BB. Les GTM organisées hiérarchiquement permettent à un utilisateur de naviguer facilement parmi les centaines de millions de composés, d'une vue globale à vol d'oiseau à la détection de motifs structurels. Afin de faciliter la navigation, un petit ensemble de composés, jouant un rôle de «balises» peut être fourni par l'utilisateur. Ces molécules seront projetées sur les GTM, apparaissant sous forme de points sur les paysages sélectionnés. Ces points aideront à choisir les zones de l'espace chimique à explorer dans le contexte des besoins de l'utilisateur.

La **Figure 11** montre la page de résultats principale contenant l'un des paysages sélectionnés. Le fond coloré de la carte correspond à la ou aux bibliothèques qui ont été sélectionnées comme base du paysage (dans l'exemple fourni - ZINC (régions rouges) et ChEMBL (régions noires) ; toutes les couleurs intermédiaires correspondent aux zones occupées par les deux bibliothèques). Les composés définis par l'utilisateur sont affichés sous forme de points noirs (**Figure 11 (5)**). Après avoir cliqué sur l'un de ces points, le composé correspondant s'affiche du côté droit de la carte (**Figure 11 (7)**). Sous la structure chimique, deux barres illustrent la proportion de composés NP et NP-like de ZINC trouvés dans l'environs les plus proches du "tracker" sélectionné (**Figure 11 (8)**). Dès que les barres sont jaunes, les composés correspondants ne peuvent pas être affichés, car ils sont trop nombreux. Dans ce cas, le bouton "Zoom" (**Figure 11 (9)**) devrait être présent, permettant de visualiser la carte zoomée - le niveau de navigation suivant.
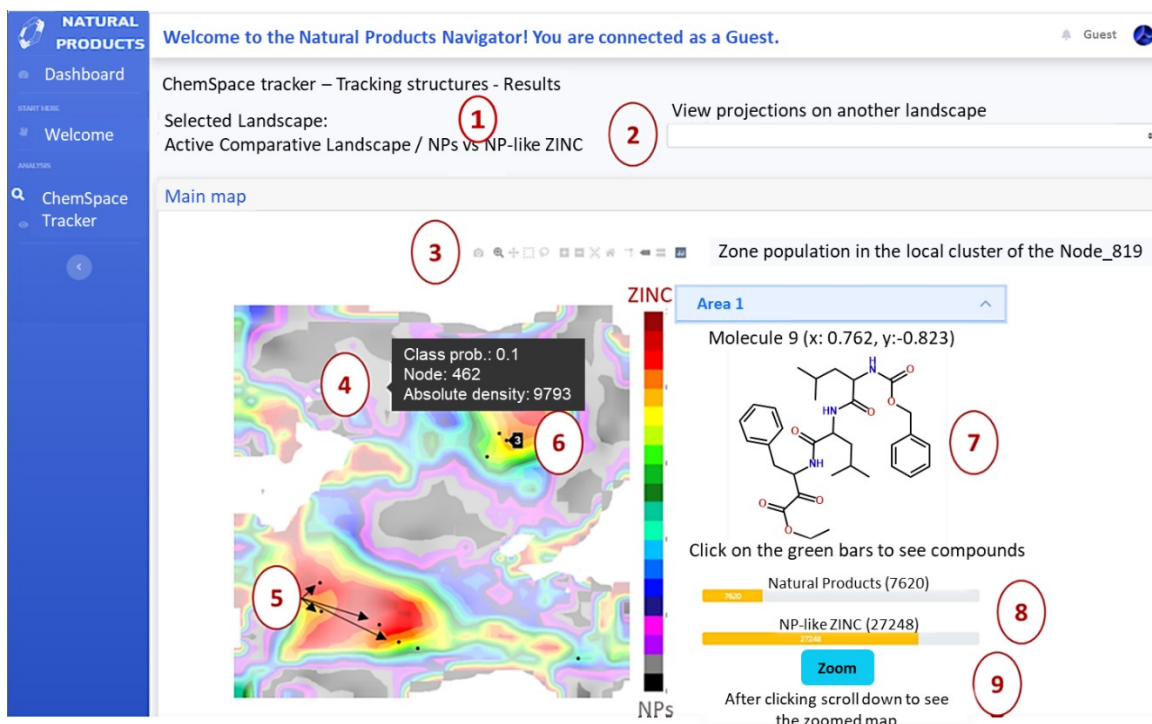
Figure 11. Visualisation du paysage au niveau principal : 1) type de paysage affiché ; 2) menu déroulant permettant de changer le paysage affiché ; 3) barre d'outils Plotly permettant différents types de navigation dans le graphique ; 4) information « hoover-activated » sur la composition du nœud (la densité absolue correspond approximativement au nombre de composés résidant dans le nœud, et la probabilité de classe indique la proportion de composés NP(0) et ZINC(1)) ; 5) les points noirs représentent les molécules définies par l'utilisateur – trackers ChemSpace ; 6) information « hoover-activated » sur le tracker ChemSpace (numéro d'index du composé dans la liste fournie) ; 7) composé dit « tracking » sélectionné ; 8) nombre d'analogues les plus proches du composé sélectionné à ce niveau du HGTM (si elles sont vertes, les barres deviennent cliquables et les composés correspondants peuvent être affichés) ; 9) bouton de zoom permettant d'afficher le niveau de navigation suivant en se concentrant sur la zone sélectionnée de l'espace chimique.

Lorsque les barres deviennent vertes, les plus proches voisins du composé dit « tracking » sélectionné peuvent être affichés. Les identifiants de source fournis pour chaque molécule sont hyperliés à l'interface web de la bibliothèque correspondante, ce qui permet un accès direct aux informations sur le composé. Un composé peut avoir plusieurs identifiants si la bibliothèque source contient plusieurs stéréoisomères. Pour des raisons de simplicité, la stéréochimie a été omise dans l'analyse des bibliothèques ultra-larges. Par conséquent, tous les identifiants de stéréoisomères ont été attribués à une seule structure chimique appauvrie en stéréochimie. Au dernier niveau de zoom, l'analyse des MCS est disponible. Les utilisateurs peuvent récupérer les MCS spécifiques à la bibliothèque et les MCS communs caractérisant la zone sélectionnée.

Outre la navigation simple, ChemSpace Atlas peut être utilisé pour une analyse efficace des chimiothèques sous-jacentes - distribution des chémotypes, propriétés physico-chimiques, activité biologique (rapportée et / ou prévue) et disponibilité commerciale. De plus, une prédiction d'activité basée sur le modèle consensuel de sept cartes universelles est également disponible.

1.5 Conclusions

La cartographie topographique générative (GTM) a été utilisée avec succès pour l'analyse de très grands espaces chimiques (jusqu'à près de 2,5 milliards de composés) pertinents pour la chimie médicinale. Plusieurs sous-espaces ont été analysés dans cette thèse : criblage de composés (fragment-like, lead-like, drug-like and PPI-like subsets), NP et composés de type NP, chimiothèques codées par ADN et blocs de construction. Pour ces sous-ensembles, les composés disponibles dans le commerce ont été comparés à des molécules testées biologiquement. Les chémotypes qui en résultent incitent à enrichir les catalogues commerciaux ou à explorer de nouvelles voies en chimie médicinale.

La hiérarchie des GTM, comportant différents sous-espaces composés, a été combinée dans un nouvel outil en ligne polyvalent disponible gratuitement - ChemSpace Atlas (https://chematlas.chimie.unistra.fr). Il permet une navigation interactive de la vue globale à vol d'oiseau à une vue rapprochée avec une analyse structurelle des composés de petites régions de l'espace chimique. ChemSpace Atlas peut être utilisé pour l'analyse structurelle et des propriétés, la comparaison de chimiothèques, la recherche d'analogues et même la prédiction de propriétés.

1.6 Liste des presentations

Zabolotna Y., Casciuc Iuri, Horvath D., Marcou G., Bajorath J., Varnek A. Generative Topographic Mapping in Virtual Screening: why ensemble of maps is needed? *Strasbourg Summer School in Chemoinformatics* (26 Juin 2018) **Poster**

Zabolotna Y., Casciuc Iuri, Horvath D., Marcou G., Bajorath J., Varnek A. Generative Topographic Mapping in Virtual Screening: why ensemble of maps is needed? *7th French-Japanese Workshop on Computational Methods in Chemistry* (2-3 Juillet 2018) **Poster**

Zabolotna Y., Horvath D., Varnek A. Zoom – a closer look on chemical space. *Journée Scientifique Doctorant et Master UMR7140* (7 Mai 2019) **Oral**

Zabolotna Y., Ertl P., Bonachera F., Horvath D., Marcou G., Varnek NP Navigator: a new look at the Natural Products Chemical Space. *3rd ICReDD (Institute for Chemical Reaction Design and Discovery, Hokkaido University) International Symposium* (22-24 Février 2021). **Poster**

Zabolotna Y., Ertl P., Bonachera F., Horvath D., Marcou G., Varnek NP Navigator: a new look at the Natural Products Chemical Space. *Journée Scientifique Doctorant et Master UMR7140 (4 Mai 2021).* **Oral**

Zabolotna Y., Ertl P., Bonachera F., Horvath D., Marcou G., Varnek NP Navigator: a new look at the Natural Products Chemical Space. *Natural products & food informatics Symposium at the American Chemical Society Conference (23 Août 2021).* **Oral**

1.7   Liste des publications

Casciuc I., **Zabolotna Y.,** Horvath D., Marcou G., Bajorath J., Varnek A. (2018) Virtual Screening with Generative Topographic Maps: How many maps are Required? *Journal of Chemical Information and Modeling,* 59, 564-572. https://doi.org/10.1021/acs.jcim.8b00650.

Volochnyuk D. M.; Ryabukhin S. V.; Moroz Y. S.; Savych O.; Chuprina A.; Horvath D.; **Zabolotna Y**.; Varnek A.; Judd, D. B. (2019) Evolution of commercially available compounds for HTS. *Drug Discovery Today*, 24, 390-402. https://doi.org/10.1016/j.drudis.2018.10.016.

Zabolotna Y., Lin A., Horvath D., Marcou G., Volochnyuk D.M., and Varnek A. (2021) Chemography: Searching for Hidden Treasures. *Journal of Chemical Information and Modeling,* 61 (1), 179-188. https://doi.org/10.1021/acs.jcim.0c00936.

Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. (2021) NP Navigator: a New Look at the Natural Product Chemical Space. *Molecular Informatics,* https://doi.org/10.26434/chemrxiv.14236457.v1.

Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A., (2021) SynthI: a new open-source tool for synthon-based library design *ChemRxiv. Cambridge: Cambridge Open Engage;* https://doi.org/10.33774/chemrxiv-2021-v53hl-v2. This content is a preprint and has not been peer-reviewed.

Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Horvath, D.; Gavrylenko, K.; Marcou, G.; Moroz Y.S.; Oksiuta, O. Varnek, A., (2021) A close-up look at the chemical space of commercially available building blocks for medicinal chemistry. *ChemRxiv. Cambridge: Cambridge Open Engage;* https://doi.org/10.33774/chemrxiv-2021-clq4h. This content is a preprint and has not been peer-reviewed.

Zabolotna, Y.; Pikalyova R.; Volochnyuk, D.; Horvath, D.; Marcou, G.; Varnek, A., (2021) Exploration of the chemical space of the DNA-encoded libraries. *ChemRxiv. Cambridge: Cambridge Open Engage;* https://doi.org/10.33774/chemrxiv-2021-dpbdx. This content is a preprint and has not been peer-reviewed.

2 Introduction

2.1 General Introduction

At the dawn of medicinal chemistry, its focus was mainly on molecules extracted from natural sources or discovered by serendipity[18]. In the middle of the twentieth century, decades of notable discoveries in medicinal chemistry could be summarized in a few pages (http://www3.uah.es/farmamol/The Pharmaceutical Century/). Thus experts had the leisure to get acquainted with every remarkable new drug or drug candidate. After the development of advanced physicochemical methods, such as X-Ray crystallography[19, 20], NMR spectroscopy[21-23], and cryo-electron microscopy[24], medicinal chemists started to understand the nature of molecular activity against a given biological target. The revolution in informatics and robotics led to parallel and automated synthesis[25], combinatorial chemistry[26, 27,] and high throughput screening (HTS)[28, 29], causing an enormous growth of chemical collections. In parallel, the breakthrough in molecular biology and genomics unveiled an extremely diverse panel of enzymes and receptors – some more relevant for disease control than others, some more "druggable" than others[30, 31]. New information started to be produced at a higher speed than it could be analyzed and interpreted by drug design experts. In 2014, Lusher et al. presented the first concerns about medicinal chemistry entering the Big Data era and challenges it begets[1].

One of the most significant contributors to the expansion of chemical data is combinatorial chemistry. However, many of the early combinatorial libraries are now considered far from the optimal chemical space appropriate to initiate a successful drug discovery project[32]. The realization that unbiased library synthesis and screening cannot revolutionize the drug discovery process and overshadow natural products led to the "fall" of combinatorial chemistry.[33] In response, medicinal chemists turned to virtual (also called tangible) compound libraries in a search for higher diversity, quality, and novel chemotypes[34]. It became the state of the art to use virtual libraries for virtual screening (VS)

in order to obtain a more extensive and diverse pool of primary hits, out of which a smaller subset would be selected for synthesis and experimental testing. This trend encouraged the creation of numerous virtual libraries, with each new one being significantly larger than the previous ones, leading to the current moment when it became hardly possible to comprehend the whole scope of all available compounds.

Even leaving aside purely theoretical libraries resulted from exhaustive enumeration of all possible organic molecules regardless of their synthesability (like GDB libraries[35]), there are still a lot of tangible compound collections up to date. They consist of already enumerated compounds or a set of reaction rules and respective reagents for their generation. Some of them are public, like SCUBIDOO[36] (21M), SAVI[37] (283M), and CH/PMUNK[38] (95M). Others consist of tangible compounds that are not just synthesizable, but in theory, can be purchased from respective chemical suppliers with a success rate of around 80% - WuXi Virtual library(100M), Enamine REAL database[39] (1.3B), and Enamine REAL space[40] (29B). Similar tangible libraries from other suppliers are significantly smaller, but they are still included in PubChem[41] (100M) and ZINC[42] (nearly 2B) databases, which became a golden standard of VS. In addition, multiple Big Pharma companies developed their own proprietary virtual libraries, adapted to their in-house building block (BB) collections and reactions. Among them, there are PLC - Proximal Lilly Collection by Eli Lilly[43] (10^{10}), BICLAIM by Boehringer Ingelheim[44] (10^{11}), Pfizer Global Virtual Library[45] (10^{14}) etc.

However, apart from the initial publications, focused mostly on one library at a time and reporting an easy statistical analysis of some property distributions, there are almost no comprehensive investigations concerning these libraries' potential value for drug discovery. Some of the listed virtual collections come with an online interface and specifically designed search engines, limited, however, only to a quick similarity search without the possibility of detailed analysis of particular regions of chemical space or the collection as a whole.

The main reason for the absence of detailed studies is the high computational challenge for ultra-large library analysis and comparison. Up to date, the largest chemical space to be visualized and closely analyzed consists of around 20 M compounds[13, 46]. This large number, even though being a small portion of available now compound libraries, can be considered as a current upper limit of contemporary chemical space analysis techniques. Thus there is a need for an efficient computational approach for expanding this limit in order to move to ultra-large chemical space navigation and exploration. Considering the main trends in drug discovery, such an approach cannot be limited to a simple similarity search. Physicochemical

properties distribution, synthetic accessibility, experimental and/or predicted biological activity, ADME-Tox properties, and scaffold analysis should also be available. Moreover, all of these must be "Big Data"- compatible in order to cope with hundreds of millions of compounds.

Chemography, by analogy with geography, as an "art of navigating in chemical space" is one of the most efficient approaches suitable to tackle described challenge[35]. Generated by means of different dimensionality reduction methods, 2D chemography maps are comprehensive and easy-to-use representations of the complex chemical space. As a chemical neighborhood is the fundamental basis of this endeavor, the selection of the appropriate descriptors for representing molecules in N-dimensional chemical space and an efficient method for its dimensionality reduction is crucial for successful visualization and analysis.

Even though there are plenty of different approaches to translate compounds from the initial descriptor space to a 2D latent space, Generative Topographic Mapping[2], or GTM, outperform most of them thanks to its non-linearity, probabilistic basis, and log-likelihood objective function enabling meaningful training of the GTM manifold. In contrast to Self-Organizing Maps[47], GTM distributes molecule projections over the map with node-specific probabilities (responsibilities) instead of unambiguously assigning each compound to only one point on the map. This smoothness enables creation of GTM landscapes – maps, colored by average values of different properties, e. g. density, property, biological activity, assigned class, etc. These maps can be turned into potent quantitative structure−activity relationship/quantitative structure−property relationship (QSAR/QSPR) models.[5, 48-50] Although a 2D map may be limited in the number of compounds it can accommodate, a hierarchical zooming approach[3, 4] allows solving this problem and capture details of the chemical population at any point of the global map. This technique consists in a new map training based on a set of compounds extracted from a given zone on the parent map in order to ensure a locally optimal mapping. The hierarchical pile-up of GTMs, from the "universal" overview map to detailed maps of local clusters, makes this strategy "Big Data"-compatible. Moreover, as new information emerges every day, it is a significant advantage that new data points can be easily projected into the existing map without retraining GTM.

Thus, this thesis is dedicated to the detailed GTM-based analysis of the currently available chemical space and the development of a new intuitive web-based tool, called ChemSpace Atlas (https://chematlas.chimie.unistra.fr/) . It enables efficient exploration of the ultra-large chemical space and its detailed analysis in terms of chemotype distribution,

physicochemical properties, (reported and/or predicted) biological activity, and commercial availability.

2.2 Publicly available sources of chemical information

The search for potential ligands of thousands of therapeutic targets via the experimental screening of large compound collections is complex and expensive. Chemoinformatics assists this process allowing to analyze and compare compound collections, predict various properties, and rationally design libraries for more successful experimental screening. In addition to efficient computational techniques, the availability of high-quality chemical and biological data is essential in this domain.

With the advancement in synthesis and biological screening, the amount of annually produced information has increased significantly – more than 20K-30K of new compounds are published every year in the leading medicinal chemistry journals in a form that does not allow an automated search and retrieval.[51] The development of the computer and internet technologies enabled storing all medicinal chemistry relevant data electronically with a convenient way of access and search. In the last two decades, several dozens of publicly accessible libraries were established.[52] They differ in their primary focus:

- sequences and 3D structures of biological macromolecules (Protein Data Bank[53, 54], GenBank[55], UniProt[56], etc.)
- experimental measurements of the biological effect of ligands of important biological targets (BindingDB[57], ChEMBL[58, 59], PubChem[41], DrugBank[60], etc.);
- commercially availability of screening in-stock or tangible libraries (ZINC15[61] and ZINC20[42], eMolecules[62], etc.).

These web applications and/or databases help experimentalists and computational experts quickly integrate different data types and advanced drug design tools in their everyday research tasks.[63] Here, the review of some publicly accessible, chemistry-oriented databases used in the current thesis is provided.

2.2.1 ChEMBL

ChEMBL is a large-scale collection of bioactivity data from binding, functional and ADMET assays[58]. It is maintained and curated by the European Bioinformatics Institute, an outpost of the European Molecular Biology Laboratory in the UK. Most of the ChEMBL

records are manually extracted from the medicinal chemistry scientific literature, but it also includes bioactivity data from deposited datasets. For example, it contains confirmatory assays with dose-response endpoints from PubChem and bioactivity data extracted by BindingDB from patent documents.[52] The data in ChEMBL is regularly updated at least once a year. Therefore, in this work, five different versions of ChEMBL database were used (**Table 1**)

Table 1. The main characteristics of ChEMBL database versions used in this thesis.

Version Release	Compounds	Activities	Assays	Targets	Source docs
V.23[64] 19.05.2017	1 735 442	14 675 320	1 302 147	11 538	67 722
V.24[65] 31.05.2018	1 828 820	15 207 914	1 060 283	12 091	69 861
V.25[66] 28.03.2019	1 879 206	15 504 603	1 125 387	12 482	72 271
V.26[67] 3.03.2020	1 950 765	15 996 368	1 221 311	13 377	76 076
V.28[68] 17.02.2021	2 086 898	17 276 334	1 358 549	14 347	80 480

As a result of the thorough curation process, ChEMBL bioactivity data became a golden standard in VS for QSAR model training. Moreover, on account of the extensive panel of bioactivities and a large number of compounds covered by this database, ChEMBL can be perceived as a chronicle of choices made by medicinal chemists in various drug discovery projects. A wide range of compounds previously selected to be tested in various dose-response assays may serve as the most reliable representation of the biologically relevant chemical space.

2.2.2 PubChem

PubChem is a public repository established by the National Center for Biotechnology Information (NCBI) at the U.S. National Institutes of Health (NIH)[41]. Similar to ChEMBL, PubChem is a freely available database that contains bioactivity data for small molecules.

However, while ChEMBL mainly focuses on the results of multiconcentration dose-response studies, PubChem data primarily originated from HTS experiments. In the latter case, each compound is listed simply as *"Active"* or *"Inactive"* at a given concentration. Serving as a central repository for extensive primary screening campaigns, PubChem has grown to contain the most significant amount of publicly available screening data.[52]

2.2.3 Directory of useful decoys (DUD)

Apart from the high-quality training data, benchmarking datasets are needed to evaluate the performance of the QSAR/QSPR and docking models. One of the most popular datasets for this purpose is a Directory of useful decoys or DUD[69]. DUD decoys were chosen to resemble experimentally validated ligands in terms of physico-chemical properties (molecular weight, clogP, etc.), but be topologically dissimilar to minimize the probability of binding to the target. DUD contains 2 950 annotated actives for 40 different targets. For each ligand, 36 decoy molecules were selected, leading to a database of 98 266 compounds.

2.2.4 ZINC

ZINC is a publicly available database that collects commercially available compounds from various chemical vendors and annotated compounds from libraries such as PubChem and ChEMBL[70]. ZINC has grown from fewer than 1 million compounds[71] in 2005 to nearly 2 billion now[42]. Each molecule in ZINC is annotated with purchasability information (vendors and estimated delivery time) and calculated physico-chemical properties. It is available for download in 2D and 3D versions. It concerns the whole library and predefined subsets, such as target-focused, natural products, metabolites, lead-like, fragment-like, etc. Moreover, an online interface enables fast substructure and similarity search, searches by biological activity, physical property, vendor, compound name, and CAS number.

Commercially available compounds are grouped into several purchasability categories[61]:

- in stock - delivery in under two weeks, 95% typical acquisition success rate;
- procurement agent - in stock, delivery in 2 weeks, 95% typical acquisition success rate;
- make-on-demand - delivery typically within 8 to 10 weeks, 70% typical acquisition success rate;
- boutique - where the cost may be high but still likely cheaper than making it from scratch, 70% typical acquisition success rate.

In this work, the first two groups combined created in-stock commercially available subset. All the rest formed the tangible one.

2.2.5 eMolecules BBs library

eMolecules Inc.[62] is the most efficient BBs aggregator in the industry. It is only partially available to the public – free downloads include only compound structures and their internal identifiers. At the same time, price, availability, and suppliers' details are accessible only under Full Plus License. In total, eMolecules contains around 1.5M BBs from over 130 vendors. Most of them are BBs that can be synthesized on-demand, and only 450K compounds are readily available waiting on the shelves. This dataset, provided by eMolecules under a non-disclosure agreement, has been used to analyze the chemical space of BBs.

2.2.6 COCONUT

The COlleCtion of Open Natural prodUcTs (COCONUT) is the most complete up-to-date dataset of natural products (NPs), containing 406 076 unique compounds with no stereochemistry[72, 73]. They were extracted from 53 various data sources, like Traditional Chinese Medicine database[74], Marine Natural Products[75], Collective molecular activities of useful plants[76], Super Natural II[77], etc. All compounds were curated, registered, and annotated with various pre-computed molecular properties. In addition, information about the literature sources, producer taxonomy, and their geography was included whenever possible without extensive manual curation. The web interface supports different search modes: by chemical structure, by compound name, and by molecular features. Moreover, the entire content of COCONUT is available for download in multiple formats.

2.3 Chemical space concept

With the rapid growth of the abovementioned libraries (especially tangible ones), the drug discovery campaigns more and more resemble the search for the needle in the haystack. In such conditions, a deep understanding of the data currently available for medicinal chemists is of the highest importance. Chemical space is one of the most valuable concepts that allows one to study all existing and chemically feasible compounds at once.

Numerous fuzzy definitions of this term were published over the years. [78] Here, we will define it as follows: a *chemical space* is an abstract space in which points represent compounds and in which neighborhood relationships are clearly defined. It can be represented in the form of a vector space, based on the vector of molecular descriptors serving to position each compound, where the associated metric (dissimilarity score) defines the neighborhood behavior (NB). Reversely, compound locations can be defined by specifying the complete matrix of inter-compound distances (the "kernel" – based on which a set of implicit molecular descriptors could in principle be derived by "embedding"). In both cases, such spaces are eligible for dimensionality reduction[79] approaches, i.e. they can be mapped.

Last but not least, a less information-rich neighborhood specification consists in providing information only about the nearest neighbors (NNs) of each compound. Formally these outlined NNs can be viewed as "connected" compound pairs, making it possible to

Main terminology

Virtual screening (VS) - a computational technique used in drug discovery in order to identify potential binders of a drug target.

Chemical space – is an abstract space in which points represent compounds and in which neighborhood relationships are clearly defined.

Graph – a set of objects organized in a structure, in which some pairs of the objects are in some sense "related".The objects correspond to mathematical abstractions called nodes and each of the related pairs of nodes is connected by an edge.

Scaffold - is the union of ring systems and linkers in a molecule with no side chains included.

Descriptors – numerical values that encode the structural and/or physicochemical properties of molecules combined in a vector.

Dimensionality reduction – transformation of data from a high- into a low-dimensional space so that the obtained low-dimensional representation preserves some properties of the original data.

outline that chemical space as a graph of interrelated molecules. Note that any mappable chemical space also allows graph representations (a dissimilarity cutoff to decide which compound pairs are close enough is all that is needed), but the reciprocal does not apply.

A short description of the *graph-based* and *map-based methods* of chemical space visualization, together with several examples, are provided below.

2.3.1 Graph-based methods of chemical space representation

In graph-based methods, the chemical space is represented by mathematical graphs consisting of a set of molecular structures connected by edges representing the relationships between them. One of the most well-known graph-based methods used in chemoinformatics to visualize the large chemical datasets is based on the arrangement of molecular scaffolds in a hierarchical structure called scaffold trees[80]. A scaffold is the union of ring systems and linkers in a molecule with no side chains included[81]. Their hierarchy is obtained by rule-based, repetitive removal of rings starting with more complex scaffolds - «leaf» nodes of the scaffold tree - till one-ring scaffold - the «root» node (**Figure 12**).

Another way to visualize and analyze graph-based chemical space is by using similarity-based chemical space networks (CSNs), where nodes indicate compounds and edges designate pairwise similarity relationships[82]. If two compounds are connected, it means that the similarity value between them satisfies some threshold criterion. The network connectivity pattern is called its network topology and is an essential parameter for characterizing networks both globally and locally. It depends directly on the chosen similarity threshold value: the topology will most likely change if the threshold is altered.

The main advantage of CSNs is their ability to capture both the discrete structure of the analyzed chemical spaces and the similarity relationships between pairs of molecules residing within them. With CSNs, there is no need to construct a coordinate system or use any form of dimensionality reduction. However, the main limitation of such methods is the size of the analyzed libraries. For example, the ChemTreeMap tool, producing a hierarchical tree with branch lengths proportional to molecular similarity, can only visualize up to approximately 10 000 data points[83]. This limitation can be explained by the need to calculate the pairwise similarities for all residents of the chemical space, which can be computationally expensive and time-consuming.
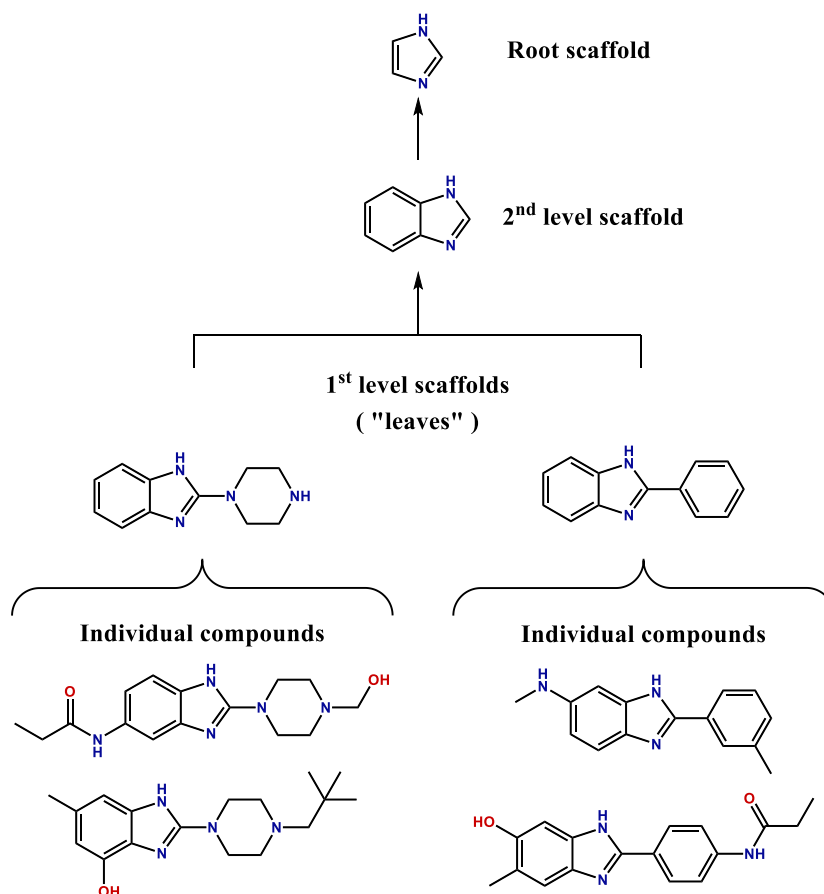
Figure 12. Scaffold tree visualization of the chemical space. From down to top: at the first step, all substitutes are truncated, and first-level scaffolds are obtained; then, according to definite rules, one cycle is cut to obtain a second-level scaffold and further. In the top - one ring "root" scaffold.

In order to overcome this restraint, Probst and Reymond combined locality sensitive hashing (LSH) and graph theory into a new algorithm called Tree Map (TMAP)[46]. At the first stage, fingerprints representing each data point are indexed by the MinHash procedure to create an LSH forest[84] of n trees. This LSH forest is then used to simplify the extraction of the k approximate NNs for each compound to form a graph in which nodes are the structures and edges are the NN relationships weighted by the fingerprint distance. In such a way, the compound similarity is expressed by the proximity of compounds through tree branches, which makes this methodology applicable to large datasets of up to 10^7 compounds. However, with the increase of the size of the analyzed library, the global graphical depiction of any network, including TMAP, became more complex until it can no longer capture the detailed structure of the chemical space. Moreover, the addition of new data points requires the reconstruction of the entire graph.

2.3.2 Map-based methods of chemical space representation

In map-based methods, molecules are represented as data points in multidimensional descriptor space. The dimensionality of such space is defined by the number of molecular descriptors - numerical values representing the structural and/or physicochemical properties of molecules[85, 86]. The descriptor space can be mapped onto the human-readable 2D map using dimensionality reduction methods. The most well-known techniques used for dimensionality reduction are principal component analysis (PCA)[87], self-organizing maps (SOM)[47], t-distributed stochastic neighbor embedding (t-SNE)[88], and generative topographic mapping (GTM)[2].

PCA is a linear dimensionality reduction method used to emphasize variation in the data and recognize patterns in it. From a mathematical point of view, the aim of PCA is to provide a new set of uncorrelated variables, called principal components, which will explain as much variation in the data as possible. Each of the principal components represents a linear combination of the original descriptor vectors. The first principal component - PC_1 - always accounts for maximum variance in data, which means that the data are spread mainly along its axis. A significant property of the principal components is that they are all orthogonal to each other. Their quantity is equal to the number of descriptors initially encoding the dataset.

t-SNE is a widely used non-linear stochastic method of highly-dimensional space visualization. The first step of t-SNE consists in converting the Euclidean distances between two data points in the higher- and lower-dimensional spaces into the conditional probabilities that those two points will be neighbors in a selected space. The difference between these conditional probabilities is then minimized, so the neighbors in the initial descriptor space will be mapped closely into the 2D plot.

SOM is another non-linear stochastic method of dimensionality reduction. It is based on unsupervised, competitive learning. It consists of a single layer of artificial neurons assembled in a two-dimensional array, with each neuron having a fixed number of neighbors. The neuron is represented by the vector of randomly initiated numbers. It has the same dimensionality as the chemical space and thus defines the neuron's position in the multi-dimensional space, in the same way as the descriptor values define coordinates of the compounds. The values of the neuron's vector are adjusted during the training to move them closer or overlap with the training data. After that, each molecule is unambiguously assigned the closest neuron in descriptor space.

All of those methods have their advantages and disadvantages. For instance, as a linear method, PCA can process massive datasets only if they have linearly dependent features. SOM and t-SNE are non-linear methods and thus overcome this drawback. However, both of them, in their classical implementations, are stochastic algorithms. Therefore, different runs would result in different 2D plots, which raises the problem of reproducibility and inability to compare different maps trained on the same data. In addition, due to the necessity to store a distance matrix for the whole dataset, t-SNE is limited in its application to relatively small datasets. The standard solution, in this case, would be to train the model using a representative subset and project the remaining data onto the 2D map. However, it is not applicable to t-SNE due to the inability to project new data onto the previously built map. In contrast, the great advantage of SOM is that the new data can be projected without its reconstruction. From the other side, classical SOM forces molecules to be assigned to only one "winning" neuron without considering neuron-specific probabilities, which increases the amount of information lost upon dimensionality reduction.

GTM (often seen as a probabilistic extension of SOM) overcomes all the disadvantages mentioned above and provides additional benefits for data analysis. PCA-based initialization of the manifold ensures reproducibility of the resulting GTM maps. Besides, the log-likelihood objective function enables meaningful optimization of the manifold coordinates in high-dimensional space in order to describe chosen training dataset in the best way. In addition, the ability to project new data without map reconstruction opens the possibility to analyze larger datasets using a map, trained on the small representative subsets. Opposed to SOM, GTM distributes molecule projection over the map with node-specific probabilities. This smoothness enables the creation of GTM landscapes that can be used not only for visualization but also as quite accurate predictive models.

2.3.3 ISIDA descriptors

No chemical space is invariant to the descriptors used to encode molecular information: different representations would lead to different spaces and chemical neighborhood relationships may or may not be maintained among these representations.[82] Therefore, not only the dimensionality reduction method but also the descriptors type should be chosen wisely.

In this work, various ISIDA property-labeled fragment descriptors[89] are used. They encode molecular structures as counts of occurrences of specific subgraphs in each compound. Nodes of these subgraphs correspond to atoms and can be labeled by element type or by some local property/feature: pH-dependent pharmacophore type, electrostatic potential, force field type etc. Edges of the subgraphs correspond to the bonds (the bond type information can be either present or omitted). ISIDA fragments could be classical atom pairs, linear sequences, augmented atoms (central atoms with their environment), or multiplets. In such a way, a user can choose between hundreds of ISIDA fragmentation schemes with different levels of resolution of the chemical information extracted into the descriptors.

2.4 Freely available web tools for the interactive chemical space visualization

Over the last decades, multiple standalone software based on the methods described above have been developed.[83, 90-93] They provide a wide range of functionalities for chemical space visualization and analysis. However, they can be difficult to install and maintain.[94] Their usage may require technical coding or scripting skills, making them available mostly to chemoinformatics professionals.[95] Therefore, online resources can be a more convenient choice as soon as they usually are intuitive and relatively easy to use.

At the moment, there are almost a dozen of freely available online servers that allow navigation and analysis of the chemical spaces defined by different MedChem relevant libraries (**Table 2**). Most of them rely on the map-based chemical space representation methods - PCA and t-SNE - and only tMap server features CSNs-like representation. As mentioned above, PCA allows processing massive datasets and visualize them both in 2D and 3D. However, it is a linear method and may thus miss non-linear relations among the input molecules, making resulting maps less informative.

All of those tools visualize precomputed libraries, and some servers even allow users to project a limited set of user-defined compounds. However, the latter usually takes a long time and sometimes forces websites to crush. The size of the precomputed datasets varies from 10^2 to 10^7, which is the current limit not only of web tools capabilities but chemical space visualization techniques in general. Moreover, with an increase in the number of data points, the available functionality decreases. Indeed, two implementations that enable navigation among up to 10M compounds – tMap and Faerun – provide only simple visualization of physicochemical properties without the possibility to project new data for analysis. tMap also allows some activity visualization (e.g., biological target classes). In addition, interpretability and convenience of navigation expectedly drop for the largest chemical spaces, as soon as all existing models provide only global level of chemical space detalization.

Table 2.Comparison of web-implementation for 2D and 3D chemical space visualization, sorted by the size of the largest analyzed chemical space. Tools marked with asterisk have been developed and reported in the course of current thesis

Name of the server	Descrip-tors	Dimensio nality reduction method	Chem space size	Analyzed libraries	New data project ion	Activity visualiza -tion	Phys Chem visuali-zation	QSAR/ QSPR	Structural analysis	Intuitivity and simplicity of web interface
2D visualization										
PUMA[94] *(Year:2017 Citations: 20)*	ECFP4	PCA	10^2-10^3	Pathogen Box; Epigenetic_focu sed; FDA-oncology	✓	✓	✓	✗	✗ *but various diversity plots are availble*	✓ *3D visualization is also available*
*Chemical Space project[96] *(Year:2019 Citations: 12)*	ECFP6 fingerprin ts	Parametric t-SNE	10^3	PubChem (TAAR1 ligands) DUDe (nuclear receptors' ligands)	✗	✓ *only two activities*	✗	✗	✗	✗ *inconvenient compound selection and display*
*iBioProVis[97] *(Year:2020 Citations: 1)*	ECFP4 fingerprin ts	PCA + t-SNE	10^5	ChEMBL (v25)	✓	✓	✗	✗	✗	✓
AtlasCBS[98] *(Year:2012 Citations: 5)*	Ligand Efficiency Indices	-	10^6	BindingDB *(19/05/2012)*, PDBBind *(v2011)* and ChEMBL *(v13)*	✓	✓	✓	✗	✗	✗ *Complex descriptors reduce interpretability*

Name	Descriptor	Method	Size	Database						Notes
*tMAP[99] (Year:2020 Citations: 26)	ECFP4 and ECFP4 fingerprints	TMAP	10^7	ChEMBL, FDB17, the Natural Products Atlas, DSSTox	✗	✓	✓	✗	✗	✓ *Hard to analyze the largest datasets*
3D visualization										
webDrugCS[100] (Year:2016 Citations: 14)	Various fingerprints	PCA	10^3	DrugBank	✓	✗	✓	✗	✗	✓
ChemMaps[95] (Year:2018 Citations: 8)	648 1D/2D RDKit descriptors +502 3D descriptors	PCA	10^4-10^5	DrugBank (v5.1.2) DSSTox (2019-3-09)	✓	✗	✓	✓ *Toxicity*	✗	✓
ChemGPS-NP[101] (Year:2007 Citations: 150)	35 PhysChem descriptors	PCA	10^5	Dictionary of Natural Products (October 2004)	✓	✗	✓	✗	✗	✗ *website visualization was unsuccessful*
Faerun[102] (Year:2018 Citations: 19)	Various fingerprints	PCA	10^7	ChEMBL, sureChEMBL, FDB17, GDBChEMBL, GDBMedChem, PubChem, Peptide CS	✗	✗	✓	✗	✗	✓ *hard to analyze the largest datasets*

At the same time, smaller navigators like PUMA and ChemMaps provide users with broader functionalities allowing to project new datasets, compare them with pre-computed libraries. In the case of PUMA, diversity analysis (scaffold and fingerprint diversity plots etc.) is also available, and ChemMaps has an option of toxicity prediction.

None of the existing web implementations support activity profiling, even though the activity maps can be displayed. Another significant shortcoming of existing tools is the availability of only one global view on the chemical space, without the possibility to analyze the local features of smaller clusters containing close analogs. It also explains the absence of structural functionality, like scaffold and MCS analysis.

Mentioned limitations are mainly caused by the weakness of the underlying chemical space representation techniques. Thus, in order to design a powerful polyfunctional online navigator of the chemical space, different methodology should be selected. As discussed in the previous chapter, GTM is a highly efficient dimensionality reduction method that possesses numerous advantages and overcomes many drawbacks of other approaches. Apart from the ability to turn the activity maps into predictive models, GTM in its hierarchical extension becomes BigData compatible. It provides intuitive, easy-to-use, and highly interpretable global and local outlooks of the chemical space and enables structural analysis of the selected zones. All of that makes GTM one of the best choices for developing ChemSpace Atlas – a new chemical space visualization tool with extended functionality.

2.5 Generative Topographic Mapping overview

Generative Topographic Mapping (GTM) is a dimensionality reduction method introduced in 1998 by Bishop et al.[2]. GTM can be understood as a probabilistic extension of SOM and PCA. As a dimensionality reduction technique, the algorithm performs a non-linear projection from the initial N-dimensional space (descriptor chemical space) onto a 2D latent space. The latter is called a *manifold* and is a finite-size surface defined using a linear combination of Gaussian Radial Basis Functions (RBFs). It is embedded in the descriptor space and sampled using a grid of points (nodes). It can have a complicated shape, with turns, twists, bends, and can cross itself. As the GTM trains to model the data distribution, the manifold itself is inserted in the densest regions of the *frameset* (the pool of molecules used to probe the chemical space of interest). Compounds are projected on the manifold, which is, in a second stage, unfolded into convenient for interpretation form of a 2D map.

The degrees of association of each compound to all nodes of the grid are called *responsibilities*. Incorporated into the responsibility vector, they define compound's position on the map (**Figure 13**). Based on such vectors for all molecules, different types of *landscapes* can be created, where each node is colored using the properties of the compounds projected there. Using those landscapes, GTM can be applied for chemical space analysis, libraries comparison, or VS.

> **Main terminology**
>
> **Manifold** – 2D latent space, described as a square grid of nodes on a flexible hypersurface.
>
> **Manifold training** – defining optimal nodes' coordinates in the initial space to approximate the shape of the frameset.
>
> **Frameset** – training dataset used to probe the chemical space during unsupervised training of the manifold.
>
> **Responsibility** – probability of compound to be assigned to a particular node.
>
> **Log likelihood** – logarithm of the probability with which the data could be associated with the manifold.
>
> **Color set** – dataset annotated with specific property that can be used to create a landscape.
>
> **Landscape** – a "colored" map obtained by adding a property as a third dimension of the 2D map.
>
> **Scoring set** – annotated dataset used to assign a particular score to each manifold during optimization.
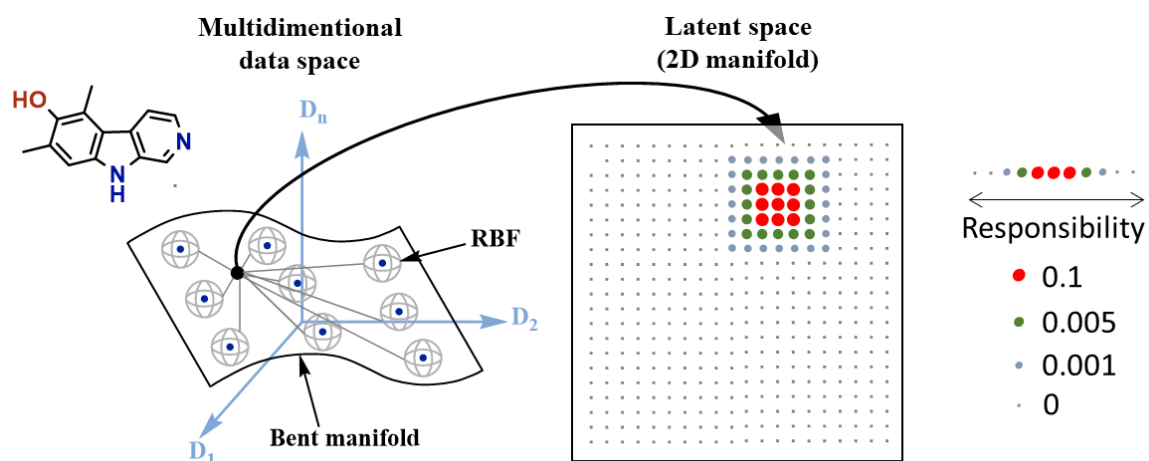
Figure 13. The general concept of GTM. The data point representing a molecule in the multi-dimensional space is projected to the 2D latent space with node-specific probabilities, called responsibilities. For every object, the responsibility is normalized over the grid of nodes; therefore, the sum of responsibilities for a given object is 1.

2.5.1 GTM algorithm

The surface of the manifold is defined by M points that serve as fixed centers for Gaussian RBFs. The linear combination of the latter forms continuous probability distribution, which for computational reasons is sampled using a grid of K nodes. Both K and M are user-defined tunable parameters that influence the complexity of the manifold and map resolution.

The RBF evaluation on a particular node (ϕ_{mk}) is defined as a function whose value depends on the distance between node coordinate x_k and fixed RBF center μ_m:

$$\phi_{mk} = \exp(\frac{\|x_k - \mu_m\|^2}{2\omega^2}) \qquad (2.1)$$

Matrix $\boldsymbol{\Phi}$ contains M×K evaluations of each RBF on each node of the manifold. In equation (2.1), ω controls the width of the Gaussian and by default, is the average squared Euclidian distance between two RBF centers.

The $\boldsymbol{\Phi}$ matrix always remains constant for a given K, M, and ω. The manifold "bending" is described by trainable matrix $\mathbf{W}$ of M×D dimensions that store the weights defining the manifold in the initial high-dimensional space. Changing the manifold will affect how the objects will be mapped from the D-dimensional into the 2D space – the closer the nodes will situate to the data points of the frameset, the better the resulting 2D map will describe them.

The mapping function Y (equation (2.2)), which is the inner product of the $\mathbf{\Phi}$ matrix with the $\mathbf{W}$ matrix, computes the coordinates of the nodes in the initial data space (**Figure 14**):

$$Y = \mathbf{\Phi}W \qquad (2.2)$$
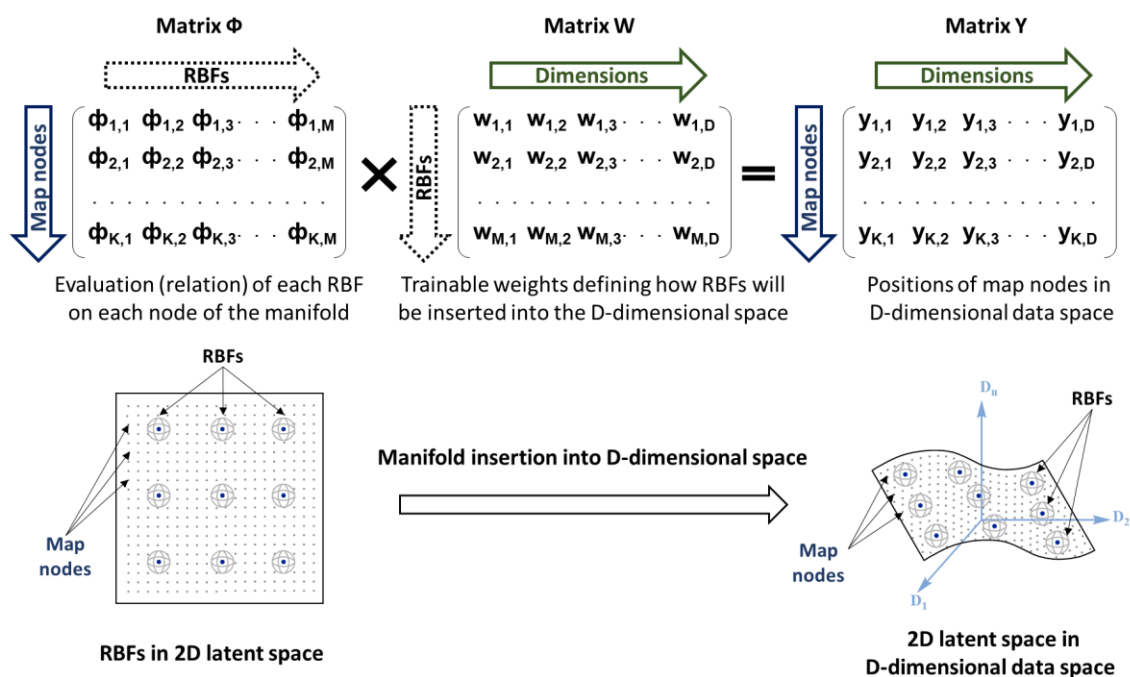
Figure 14. Matrix representation of the mapping process in GTM (equation (2.2)). Matrix Φ represent 2D latent space (manifold); matrix W – trainable weights used to insert manifold into D-dimensional chemical space; matrix Y (the result of the mapping function Y) – fitted 2D latent space in the initial space.

Unsupervised manifold fitting

The manifold fitting consists in finding its optimal shape to approximate the data distribution. The latter is defined by N compounds from the *frameset*, used to probe the chemical space of interest. The first step of the GTM training process is the initialization of the parameter matrix $\mathbf{W}$. In other words, we need to specify the starting coordinates of the manifold in the D-dimensional space. It is consistently done by application of PCA, where only the first two principal components are used. The manifold in its plane rectangular form is inserted in the two first principal components encompassing the corresponding dataset scores. The coordinates of the nodes are stored in the matrix $\mathbf{X}$, while the loadings of the first two PCs - in a matrix $\mathbf{U}$. Therefore, the initialized manifold is defined by the matrix $\mathbf{W}$:

$$W = \Phi^{-1}(XU) \tag{2.3}$$
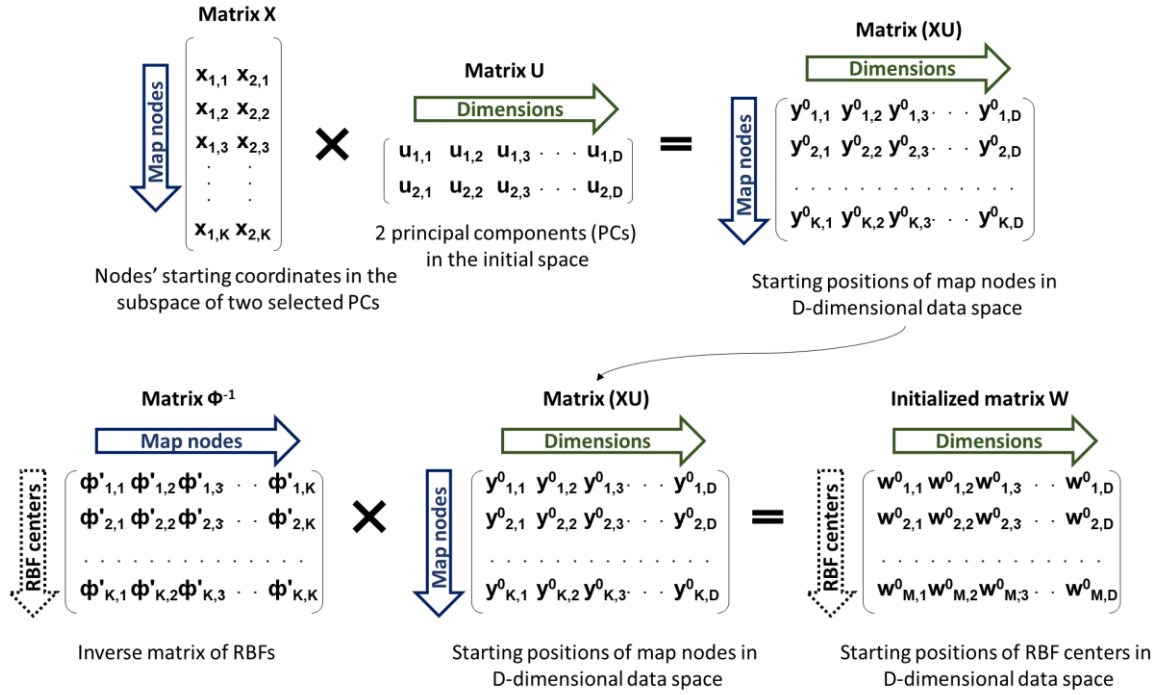
Figure 15. Matrix representation of parameter matrix W initialization (equation (2.3)). The first two principal components are selected by PCA (matrix **U**), and map nodes are represented in their basis (matrix **X**). X×U multiplication result gives starting coordinates of the manifold nodes in the initial space. Multiplication of inverse matrix $\mathbf{\Phi^{-1}}$ by a resulting matrix **(XU)** yields starting positions of RBF centers in D-dimensional data space.

Here, **U** is a matrix 2×D defining two eigenvectors resulted from PCA, and **X** is a K×2 matrix of nodes' coordinates. The result of their multiplication – matrix **(XU)** contains starting positions of map nodes in the D-dimensional data space (**Figure 15**). Now, in order to obtain starting coordinates of RBFs in the initial space, the equation (2.2) should be reversed.

The initialized manifold is then inserted into the data space, followed by frameset compounds projection. The probability density of a compound with coordinates t_n in the initial space to be associated with the node k with position x_k in the latent space is calculated with the following equation:

$$p(t_n|x_k, W, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{-D}{2}} exp\left(-\frac{\beta}{2}\|y_k - t_n\|^2\right) \tag{2.4}$$

Here, y_k defines coordinates of node k in multidimensional space and is obtained using equation (2.2) and β is inverse variance of the distribution. Its value is fitted to the data during training and initialized based on the 3rd eigenvalue of the PCA.

Integrating over the manifold allows obtaining the probability density (or likelihood) of a compound n to be projected into the manifold:

$$p(t_n|\, \boldsymbol{W}, \beta) = \frac{1}{K} \sum_{k=1}^{K} p(t_n|x_k, \boldsymbol{W}, \beta) \qquad (2.5)$$

In other words, this probability density measures the goodness of fit of the manifold to this particular data point. For mathematical convenience, its natural logarithm, known as *log-likelihood (LLh),* is preferred to characterize the quality of the projection of each compound:

$$LLh_n(\boldsymbol{W}, \beta) = \ln\big(p(t_n|\, \boldsymbol{W}, \beta)\big) \qquad (2.6)$$

The *LLh* for the whole frameset serves as an objective function for optimizing **W** (finding the optimal shape of the manifold) - the higher this value is, the better the manifold represents the data:

$$LLh(\boldsymbol{W}, \beta) = \sum_{n=1}^{N} LLh_n(\boldsymbol{W}, \beta) \qquad (2.7)$$

The manifold fitting to the data of the frameset is then performed via the Expectation-Maximization algorithm that searches the matrix **W** and the distribution width β^{-1} which maximize the $LLh(\boldsymbol{W}, \beta)$ of the training data. On the E-step, the algorithm computes a matrix **R** (a K×N matrix), containing for each compound n the list of its responsibilities to be associated with each map node k (r_{kn}). The latter is calculated using the Bayes formula and normalizing over the grid of K nodes (equation (2.8)). The second matrix computed on the E-step is diagonal K×K matrix **G,** containing the sum of responsibilities of all frameset compounds associated with a particular node (g_{kk}) that defines its population (equation (2.9)).

$$\mathbf{R} = (r_{kn})$$

$$r_{kn} \propto \frac{p(t_n|x_k, \mathbf{W}, \beta)}{\sum_{k'=1}^{K} p(t_n|x_{k'}, \mathbf{W}, \beta)} \tag{2.8}$$

E-step

$$\mathbf{G} = (g_{kk})$$

$$g_{kk} = \sum_{n=1}^{N} r_{kn} \tag{2.9}$$

On the M-step, the parameter matrix **W** is updated (equation (2.10)) using calculated on the previous step matrixes G and R, constant N×D matrix T (describes N frameset data points in the initial D-dimensional descriptor space), regularization coefficient λ and M×M unit matrix I. Based on the $\mathbf{W}'$, the algorithm computes new values of β' according to the equation (2.11). The new width and weight matrix are used as input for the next optimization step, starting with an expectation calculation. The algorithm continues until convergence that is measured based on relative loglikelihood: $\left(LLh(\boldsymbol{W}', \beta') - LLh(\boldsymbol{W}, \beta)\right)/LLh(\boldsymbol{W}, \beta) \leq 0.001$.

$$\mathbf{W}' = (\mathbf{\Phi^T G \Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi^T R T} \tag{2.10}$$

M-step

$$\frac{1}{\beta'} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{kn}\|y(x_k, \mathbf{W}') - t_n\|^2 \tag{2.11}$$

Supervised manifold selection

Manifold training described above is unsupervised, i.e., independent of molecular properties of FS members. However, the type of descriptors, the composition of the frameset, and several parameters of GTM (grid size, number of RBFs, RBF width, and regularization coefficient), can be selected in a supervised manner. For that, multiple manifolds can be constructed based on different parameters and then evaluated by a user-selected scoring function. The manifold with the best score can then be selected as an optimal choice. Two approaches can be applied in order to generate a pool of candidate manifolds:

i)   brute force grid search that generates all possible combination of optimized parameters

ii) genetic algorithm (GA) – stochastic approach allowing evolution toward better solutions.

In this work, all GTMs have been optimized using the second approach. Application of GA to GTM optimization has been previously described in detail in several publications[13, 103, 104]. Briefly speaking, in GA, each "candidate manifold" is described by a chromosome – vector of values of the abovementioned parameters that need to be optimized. In the beginning, the algorithm randomly generates a set of starting chromosomes, and respective manifolds are constructed. Each manifold is evaluated based on the"goodness" or fitness score (FSc). Higher scored chromosomes will be allowed to generate "children" using cross-overs and mutations, which might result in potentially better FSc. The GA stops in two cases: either no FSc improvement has been observed during the last two generations, or the maximal number of attempts has been achieved.

Users are free to define FSc to reflect user expectations of the map given its context- and project-specific intended applications. For example, the goodness of a map serving as a diversity selection tool resides in its ability to ensure a homogeneous spread of library compounds (have high Shannon entropy). By contrast, a map serving as QSAR predictor should have its fitness score set to some cross-validated statistical criterion reflecting the predictive power of the activity landscape it hosts. Eventually, "universal" maps are optimal if they may achieve the best mean "compromise" quality of predictive powers over an entire profile of various bioactivity QSAR challenges.

All maps in this work were optimized with respect to their success in various classification tasks – three-fold cross-validation was performed, and the mean Balanced Accuracy (BA) was computed. BA takes the rate of correct predictions of both classes in equal proportions, and varies from 1 (ideal case) to 0.5 (random predictions). Setting FSc=BA allows the correct evaluation of the predictive performance of the model while using unbalanced datasets:

$$BA = \frac{1}{2}\left(\frac{TruePositive}{TruePositive + FalseNegative} + \frac{TrueNegative}{FalsePositive + TrueNegative}\right) \quad (2.12)$$

2.5.2 Pretrained manifold application for various chemoinformatics tasks

New data projection

As soon as the manifold is trained, new data can be projected on it. Each projected point is described on the 2D map by its *LLh* value and a vector of responsibilities, calculated by equations (2.7) and (2.8). The former is used to determine whether the given manifold passes close enough to the position of the compound in the highly-dimensional descriptor space to meaningfully map it on the 2D latent space. In order to avoid ambiguous mapping, only compounds situated within a specific range around the manifold can be projected into it (likelihood-based applicability domain (AD) of GTM). For that, the *LLh* threshold is determined based on the *LLh* distribution for the frameset compounds. There are two approaches for that:

i) the cutoff can be set at n% of data points (usually 5%) having the smallest *LLh*[48];

ii) the Gaussian can be fitted to the frameset compounds distribution, minimizing the root mean square error, and the threshold will be determined as *LLh* value with the highest population (peak) minus three Gaussian widths ("3σ" rule)[4].

The responsibility vector determines the fuzzy position of the compound on the map. It makes compounds appear on the map as spots rather than points (**Figure 13**). Such an approach decreases information loss upon dimensionality reduction and reduces the probability of two compounds to take the exact same place on the map.

GTM landscapes for chemical space visualization and properties prediction

Summing responsibilities of all compounds for each node of the map allows creation of fuzzy density landscapes where color code demonstrates the population of each node (**Figure 16 (I)**). Such landscapes allow to easily spot over-or underpopulated areas of the chemical space and thus analyze compound distribution and possible disbalance towards particular chemotypes in the visualized libraries.

I) Continuous property landscapes

Density   **Molecular weight**

II) Fuzzy class landscapes

Activity class   **Library comparison**

0   20   40   60   80   200   300   400   500   600   Inactive   Active   Library 1   Library 2
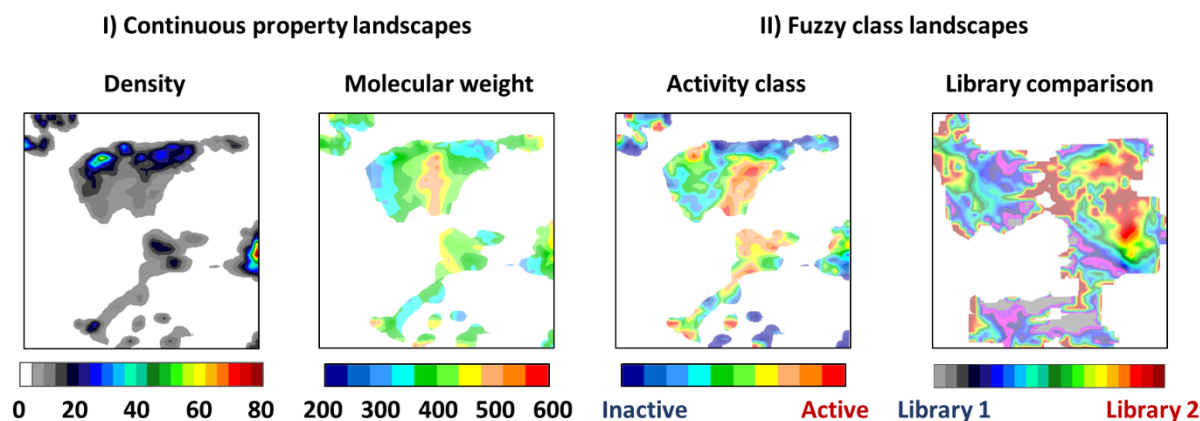
Figure 16. Different types of GTM landscapes: I) Continuous property landscapes (density and molecular weight landscapes) and II) fuzzy class landscapes (activity - activity labels as classes) and library comparison landscapes (source libraries as classes).

If the population of the nodes is complemented by the property values of compounds residing there, the property landscape can be obtained (**Figure 16 (I)**). The values defining the color of each node in such landscapes are calculated according to the equation (2.13):

$$\langle pp \rangle_k = \frac{\sum_{n=1}^{N} pp_n * r_{kn}}{\sum_{n=1}^{N} r_{kn}} \tag{2.13}$$

where pp_n is the property value for the compound n, and $\langle pp \rangle_k$ is the mean property value for node k. Such landscapes represent a distribution of the analyzed property in the latent space. Thanks to the probabilistic nature of GTM, they can be used not only for visualization but also as regression models. As soon as the property landscape is created using the annotated training set (also called a *color set*), a new compound q (assuming it is in the likelihood-based AD of the map) can be projected. The prediction of the analyzed property for q is based on the mean property values p_k for those nodes k, where compound q was projected with probabilities r_{kq}:

$$pp_q = \sum_{k=1}^{K} \langle pp \rangle_k * r_{kq} \tag{2.14}$$

If the initial dataset is split into several classes, each node can be characterized by the probability to find there a member of a particular class:

$$P(c_i | \mathbf{x}_k) = \frac{P(\mathbf{x}_k | c_i) * P(c_i)}{\sum_j P(\mathbf{x}_k | c_j) * P(c_j)} \tag{2.15}$$

where $\qquad P(x_k | c_i) = \frac{\sum_{n=1}^{N} r_{nk}(c_i)}{N_{c_i}}$

58

$$P(c_i) = \frac{N_{c_i}}{N_{tot}}$$

$r_{nk}(c_i)$ is the responsibilities of the members of the class c_i from the node k, N_{c_i} is the number of items for the class c_i and N_{tot} is the total number of training items.

Such maps can be used as a predictive classification model or to compare the distribution of classes in the chemical space it (**Figure 16 (II)**). The class value for the new compound q can be predicted similarly to the property prediction:

$$P(c_i|q) = \sum_{k=1}^{K} P(c_i|\boldsymbol{x}_k) * r_{kq} \qquad (2.16)$$

For both property and class landscapes, the population of the nodes can be visualized via the transparency of the colored regions of the map. In addition to the abovementioned likelihood-based AD, GTM-based predictive models also have AD, dependent on the density of the coloring set in a particular node. The class or property of the new compound q cannot be predicted if this compound is associated with sparsely populated nodes on the GTM landscape, where the cumulative responsibility is below a certain threshold.[48]

Normalized landscapes

With the increase of the size of analyzed libraries, the chances to face a problem of unbalanced library comparison rise. For example, let us consider a case of one library being 1000 times larger than the other. By default, in a chemical space zone that is equally well represented in both libraries, one would expect exactly the same 1000:1 ratio (which, in absolute numbers, would map as an absolute dominance of the bigger library). Therefore, in normalized plots, the cumulated responsibility of the larger library is first scaled back to values that would have been expected if the larger library would be of comparable size to the first one. Thus, 1000 of the actual cumulated density of a larger collection scales back to 1 of normalized density, and the 1000:1 imbalance is reset to 1:1, expressing equal propensities to reside in the considered zone (**Figure 17**).
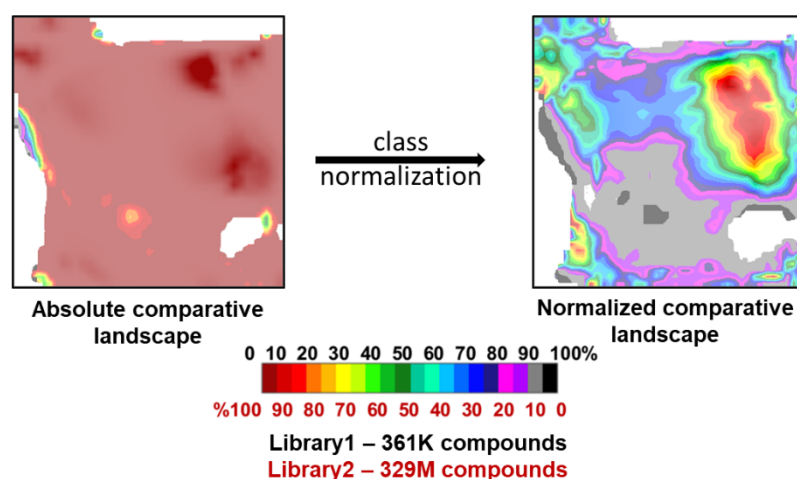
Figure 17. Landscape normalization - facilitating unbalanced libraries comparison.

Structural analysis of the map residents

Residents of the selected nodes or groups of nodes (zones) of the GTM landscape can be extracted and subjected to structural analysis, like scaffold analysis[105, 106] or Maximum Common Substructure (MCS) detection[4] etc. Due to the variety of possible landscapes that can be constructed for the same dataset, one can easily identify desired regions of the map to explore, for example:

i) the zones that have the highest population on the density landscape;

ii) areas associated with compounds with the lowest molecular weight on the property landscape;

iii) regions containing mostly active ligands against the biological target of interest on the activity landscape;

iv) nodes, exclusively populated by compounds from one of the analyzed libraries on the comparative landscape.

Apart from the simple visual detection of the areas with desired properties, one can also focus on compounds that project similarly on the map, defined by Responsibility Patterns (RPs)[106]. RPs are discretized responsibility vectors, with all values less than 0,01 being reassigned to zero and all others - to a number from 1 to 10 according to the formula below:

$$rp_{kn} = [10 * r_{kn} + 0.9] \qquad (2.17)$$

According to it, if $0.01 \leq r_{kn} \leq 0.1$, then rp_{kn} would be equal to 1, if $0.11 \leq r_{kn} \leq 0.2$, then $rp_{kn} = 2$, etc (**Figure 18 (I)**). Compounds with different responsibility vectors that correspond to the same RP are considered to be grouped in the same cell of the chemical space and thus have many structural similarities (**Figure 18 (II)**). Some RPs are characterized by a common scaffold, while the others are even more specific, being described by the common substituted scaffold (MCS) or family of scaffolds (e.g., like N heterocycles with varying positions of N atoms)[107]. Similar to the ''privileged scaffolds'', it is possible to define privileged RPs inherent to compounds with desired activities (e.g., compounds with defined activity). Such an approach is more open-minded as soon as for each particular case a different feature - scaffold, a family of scaffolds, MCS, etc. - may provide the best description of the local clusters.
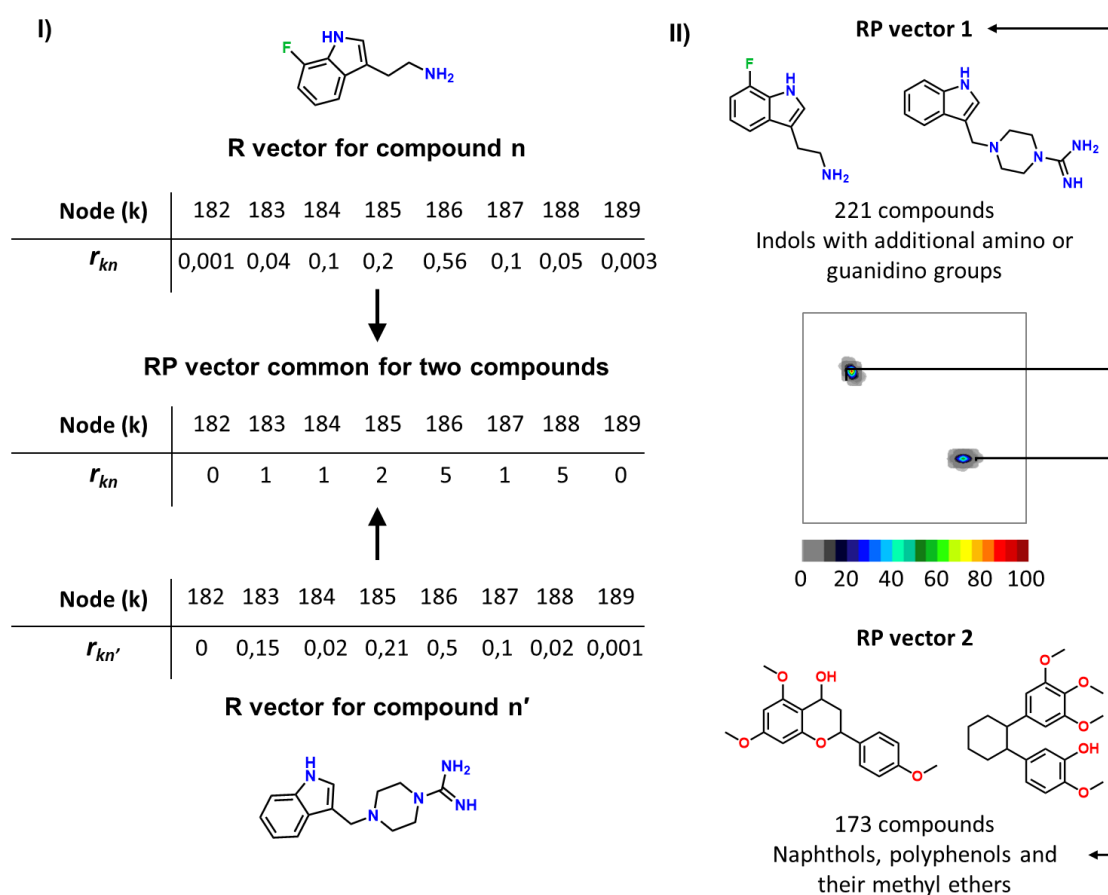

Figure 18. Structural analysis of map residents using responsibility patterns (RP) vectors. I) example of 2 compounds with different responsibility vectors (R) but the same RP vector. II) Density landscape, where each of the two spots is populated by compounds having the same RP vector.

Library comparison

There are several ways to compare libraries with the help of GTM, depending on the type of landscape used for that. Density landscapes allow estimating homogeneity of the chemical space coverage and positions of the highly populated areas for each library on the map. Coupled with a preferred structural analysis technique, the latter allows the detection of the type of compounds dominant in each library. Moreover, the cumulative responsibility vector of a compound library, used to create density landscape, can be considered as a k-dimensional descriptor of the whole library (k – number of map nodes). The similarity score for two libraries can be calculated based on such vectors, allowing quantitative estimation of the overlap between them in the latent space.[108] The comparative landscapes (**Figure 16 (II)**) enable map-based visualization of such overlap. They can be considered a special case of class landscapes, where the class assigned to each compound is a library of its origin. Such landscapes allow fast identification of library-specific areas of the chemical space as well as regions common for both libraries.

Property landscapes of compared libraries provide an overview of the desired properties distribution over the chemical space, allowing to generalize property-related characteristics of each library (e.g., predominance of the low/high molecular weight compounds of one library with respect to another, lack of non-flat molecules with a high fraction of saturated carbons (Fsp3) in one of the libraries, etc.).

2.5.3 GTM and Big Data challenge

Incremental GTM

In GTM training, the frameset compounds positions in the descriptor space are defined by the N×D matrix **T,** while their projections on the latent space are stored in a K×N responsibility matrix **R.** The sizes of both of those matrixes depend on the number of compounds N. In case of large datasets (more than 50K) these matrixes cannot be entirely stored in the computer memory, which limits the application of classical GTM to smaller libraries. One of the ways to create GTM for visualization of large chemical spaces is to use only a subset of the analyzed datasets as a frame for manifold fitting. It was shown by A.Lin

et. al[109], that the chemical space of millions of compounds could be easily represented by several thousands of randomly selected molecules as soon as GTM does not require the chemical space to be dense to train the manifold. However, ultra-large libraries can hardly be described by a few thousand molecules.

Therefore, a special strategy of larger framesets processing - an incremental GTM (iGTM)[110] – is often used. This modified algorithm divides the initial dataset into several blocks of selected size instead of using the whole data matrix. The manifold is then trained sequentially on one block at a time. The algorithm moves to the next block only after achieving convergence of the $LLh(\textbf{\textit{W}}, \beta)$ for the current block. Considering the size of the analyzed datasets in this work, only iGTM was used for the maps construction.

Hierarchical GTM

While analyzing ultra-large compound libraries, the number of compounds mapped on the GTM may be arbitrarily large, while the size of the map is constant. Therefore, the number of molecules associated with each node on the map eventually becomes too large to allow any meaningful separation by chemotypes or class. To solve this problem, hierarchical GTM (hGTM), otherwise known as "Zooming," was developed by Tino et al. in 2002.[3] The main idea of hGTM lies in fitting an additional "zoomed" manifold to the locally clustered data extracted from a small zone on the parent map. Moreover, each zoomed manifold can be further zoomed, producing several levels in a hierarchy of GTMs. Each next level of maps is more detailed and focused on a smaller area of the chemical space. They provide a higher resolution and better class/chemotypes separation.

In the study of Lin et al.[4], this methodology was combined with an automated MCS extraction protocol to develop "AutoZoom" – a tool for efficient structural comparison of large databases. In **Figure 19**, one can see the schematic representation of the AutoZoom application. The procedure consists in dividing the landscape into multiple zones 3*3 nodes and analyzing the population of such regions. Areas that contain more than 1 000 molecules are selected to construct new GTM manifolds using only local compounds from these zones as a frameset. The compound is associated with a particular zone only if the sum of responsibilities for this compound to reside there is higher than a predefined threshold (0.85 by default). If the number of area residents is lower than 1 000, the compounds from this zone are subjected to automated MCS detection instead of zooming. In AutoZoom, MCS is defined as the largest common structural fragment, containing no less than 30% of each

molecule it represents. Analysis of the extracted MCS allows profiling each area of the map with types of compounds that populate them.
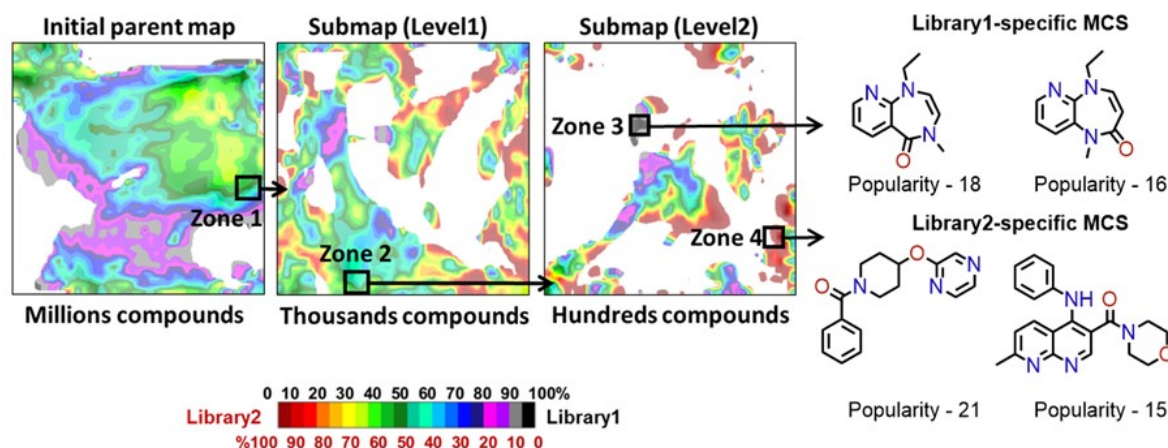
Figure 19. Example of the hierarchical navigation through densely-populated zones of the chemical space using AutoZoom.

2.5.4 Success stories of GTM application in drug discovery

As shown in the previous chapters, GTM is a powerful method encompassing a wide range of chemoinformatics functionalities - from simple data visualization to property predictions. Therefore, it has been widely applied for solving different drug discovery problems: chemical libraries analysis and comparison, VS, de-novo compounds generation with desired properties, etc.

Library analysis and comparison

The probabilistic nature of GTM and the possibility of creating maps featuring the same compounds but colored differently (various types of landscapes explained above) allow analyzing chemical libraries from different perspectives. For example, in the work of H.Gaspar et al.[108], 2M drug-like compounds gathered from 36 commercial libraries were visualized and analyzed with the help of around 15 property landscapes (molecular weight, aqueous solubility, LogP, etc.), providing different views of the dataset. The superposition of these views helped to identify the regions in the chemical space populated by compounds with desirable physicochemical profiles and the suppliers providing them.

The chemical space of antiviral compounds from ChEMBL was analyzed using RPs. The privileged locations of antiviral classes were analyzed in order to highlight underlying

privileged common structural motifs. [106, 111] It was shown that the privileged structural motif detection based on GTM RPs has a significant advantage over the classical privileged scaffolds. The former allows to automatically capture the nature ("resolution detail"—scaffold, detailed substructure, pharmacophore pattern, etc.) of the relevant structural motifs.

Almost all of the approaches of library comparison described in Chapter 2.5.2 were used to compare >10 M real ''fragment-like'' compounds (of 17 heavy atoms) from public databases to a subset of 10 M fragment-like structures extracted from 166 billion GBD-17 library[112, 113] of feasible compounds[13]. The public databases bias in favor of aromatic ring-rich molecules and against chiral compounds was easily derived from property landscapes. In addition, hGTM was used for the detailed structural comparison of the abovementioned libraries, resulting in FDB-17-specific structures identifications. They represent novel theoretical compounds that have not yet been synthesized. The diversity holes of FDB-17, caused by the systematic exclusion of particular chemotypes during FDB-17 generation, were also reported. This work featured an analysis of the largest libraries and set up the current upper limit of library analysis tools capabilities, which was extended in the present work.

The hGTM approach was used for diversifying the in-house drug-like compounds of the Boehringer Ingelheim pharmaceutical company by comparing it with a commercial catalog of more than 8M compounds from Aldrich-Market Select[4]. As a result, it was discovered that 45.5K substructures were absent in the Boehringer database. The compounds containing the identified substructures were then assessed for their drug-likeness and potential biological activity (VS). 1.2K of them were predicted active against different biological targets and thus recommended to BI as a useful dataset in diversifying their in-house collection.

GTM-driven virtual screening

As it was mentioned above, fuzzy class and property landscapes can be transformed into predictive models, useful in ligand-based VS. Even though manifold construction is an unsupervised process, it was explained that GA optimization allows to find optimal GTM parameters and the descriptor space for maximizing the predictive performance of GTM-based QSAR/QSPR models.

There were many projects reporting the application of GTM to VS. Several of them that had experimental validation are discussed here. For example, in the work of Casciuc et al., the VS funnel involving classification SVM and GTM models and ligand-based

pharmacophores was implemented to select potential binders of Bromodomain BRD4. The models were trained on publicly accessible SAR data on BRD4 IC50/pKi from Reaxis and ChEMBL. Using them, 3K compounds were selected out of 2M in-stock Enamine compounds to be tested against Bromodomain BRD4 using the Thermal Shift Assay method. Twenty-nine confirmed hits were detected, representing a 2.6 fold increase in hit rate relative to random screening.

Another more successful application of GTM in VS was based on previously described RPs assuming that molecules of the same RP have similar properties[111]. Even though being weaker than actual NB compliance in a full descriptor space, this hypothesis still allows to exclude the most dissimilar candidates quickly. RPs that mostly correspond to anti- (flavo- and entero-)viral compounds were highlighted, and commercial compounds within the privileged RPs were similarity-scored against reference antivirals within the same RP. Selected compounds were tested in cell-based assays against tick-borne encephalitis virus (TBEV) and a panel of enteroviruses. This approach allowed the identification of 23 new compounds (out of 44 tested molecules) showing anti-TBEV activity with EC50 values in the micromolar and submicromolar range.

A single GTM manifold is not limited to host only one predictive model – hundreds of properties/activities can be predicted simultaneously using correctly optimized universal GTM. The concept of Universal GTM (uGTM) was introduced by Sidorov et al.[114] as a general-purpose map that can accommodate ligands of diverse biological targets on the same GTM manifold. For its construction, the GA optimization was used to choose the best descriptors set and GTM operational parameters (number of nodes and RBFs, manifold flexibility controls, etc.) so as to maximize the mean predictive performance over hundreds of biological activities from ChEMBL. Unlike local GTMs, focusing on only one activity at a time, uGTM featured ligands of more than 400 biological targets from ChEMBL database (v20). This allowed the creation of more than 400 activity landscapes that can be used to perform polypharmacological profiling of new compounds.

Lately, Lin et al.[50] have compared the performance of universal and local GTMs with other popular machine-learning methods in VS. According to this study, GTM models demonstrate the predictive performance comparable to other popular VS techniques while providing the advantage of the visualization support.

2.6 Summary and thesis outline

Over the last 20 years since its first introduction, various GTM adaptations have been developed to make this method more suitable for chemoinformatics problems. Among them, there are the described above tools for:

- efficient GTM parameters optimization;
- library analysis and comparison with the help of various property landscapes;
- predictive QSAR models creation;
- Big Data GTM application.

Such extensive functionality of a single methodology makes GTM a highly competitive chemoinformatics instrument.

With an everyday increase of publicly available chemical information, tracking features or properties of molecules in ultra-large highly-dimensional chemical spaces becomes a crucial problem in the field. Right now, researchers have very little access to navigation tools for these ultra-large chemical spaces. Moreover, the existing tools lack both depth (insufficient information visualized) and breadth (e.g., limited to one vendor, unable to handle larger libraries etc.). Therefore, openly available tools that allow users to view and analyze chemical information on a large scale and at a high level of destabilization would be extremely beneficial, and GTM is one of the few methodologies that can enable that.

Therefore, the main goal of this thesis is to create an intuitive publicly available tool - ChemSpace Atlas - that would allow to use the full GTM functionality for the chemical space navigation and analysis, properties and activity predictions, libraries comparison etc.

The main novelty and contribution of this thesis can be summarized in three statements:

- creation of a GTM-based framework of unprecedented size (tens of thousands of hierarchically organized GTMs) that can be used to analyze ultra-large compound libraries frequently used in drug discovery. This framework will allow increasing the size of the projected data sets significantly and move up the current limit for the chemical space visualization by two degrees of magnitude (from approximately 20 Million to almost 2,5 Billion).

- exhaustive structural and property analysis of the various chemical spaces relevant for medicinal chemistry. It includes comparing commercially available catalogs to the reference libraries of compounds tested in biological essays before. Such analysis and comparison provide a deeper understanding of the chemical space and potential directions for its enhancement.

- development of the universal web interface that can accommodate multiple GTM hierarchies (separate for each MedChem relevant subspace) and provide users with access to the results of the analysis performed on the previous step.

This thesis is organized in the following manner. At first, the main framework of the ChemSpace Atlas – the universal maps built on the ChEMBL (v23.) data – is introduced together with an evaluation of their predictive performance in polypharmacological profiling (Chapter 3). Chapter 4 reports the usage of uGTM and hGTM concepts for the analysis of different important in medicinal chemistry compound subsets. The very last chapter describes the development of the web interface of ChemSpace Atlas (https://chematlas.chimie.unistra.fr/).

3 Universal Maps of Biologically Relevant Chemical Space

Introduction

The success of a ChemSpace Atlas as a chemical space visualization and analysis tool depends on a wise selection of the descriptor space and high-quality "framework" map, covering the biologically relevant chemical space. In addition, this main map should support activity prediction for the wide range of biological targets. Thus, the universal GTM (uGTM) is the best option to be a basis for ChemSpace Atlas. Indeed, uGTM provides 2D representations of chemical space, able to simultaneously represent meaningful activity and property landscapes, associated with many distinct targets and properties.

In this work, eight new universal maps of the biologically relevant chemical space were constructed using ChEMBL database (v23). A total number of 1.5M compounds with known activities on 618 targets (**Table 3**) have been extracted from the ChEMBL using the target-specific ligand series extraction protocol described in the work of Sidorov et al.[114] According to it, each compound has been assigned "active" or "inactive" class for biological targets it was tested against based on the ChEMBL-reported activity values and a chosen activity threshold (AT). A set of rules has been employed for that:

> ## Main terminology
>
> **Universal GTM** - a general-purpose map that can simultaneously accommodate several predictive landscapes manifesting satisfactory performance in different classification/regression tasks.
>
> **Cross-validation** - a statistical method that estimates how accurately the given machine learning model will predict independent data. For that, the training data is split iteratively into two subsets - one for learning and another for performance evaluation. Those two sets cross over in each iteration so that each data point will occur in each of them. The average predictive accuracy overall iterations estimates how the model will work on external data.
>
> **Consensus prediction** - a prediction based on the combination of outputs of the ensemble of predictive models. Those models can be based on different algorithms, various model parameters, or simply differ in input data representation.

1. Only a few activity types were taken into consideration inhibition (%)and dose-response activity measures (Ki, IC50, EC50, and "potency").

2. Ligands with the reported percentage of inhibition less than 50% were considered "inactive".

3. The optimal cutoff for dose-response activity values was selected separately for each target in a way to preserve a reasonable balance of "actives"/"inactives" in the target-specific ligands dataset (target should have at least 100 classified ligands, and at least 20 of them should be "actives"; percentage of "inactives" should always exceed 50% of the dataset).

4. The possible ATs are 1 000, 500, 100, and 50 nM.

5. Compounds with reported dose-response concentration lower than the AT were labeled "active", the ones with that value higher than the ten-fold AT were considered "inactives". All molecules with values in between were ignored in order to facilitate the separation problem.

6. Compounds with multiple entries leading to contradictory activity class assignments were ignored.

The type of ISIDA descriptors and GTM parameters were optimized with GA using the predictive performance of the resulting maps as a scoring function. Ligand series of 236 targets, including GPCRs, kinases, nuclear receptors etc., have been used for 3-fold cross-validation[115]. Nine target-specific compound sets extracted from the Directory of Useful Decoys (DUD) were used for external validation of the polypharmacological predictive performance of each uMap separately, and all of them combined in a single consensus model.

Table 3. 618 ChEMBL(v.23) targets used for universal maps training and validation.

CHEMBL1075104	CHEMBL1293266	CHEMBL1790	CHEMBL1859	CHEMBL4633
CHEMBL1075145	CHEMBL1293267	CHEMBL1795139	CHEMBL1860	CHEMBL4641
CHEMBL1075167	CHEMBL1293289	CHEMBL1795186	CHEMBL1862	CHEMBL4644
CHEMBL1075189	CHEMBL1293293	CHEMBL1801	CHEMBL1864	CHEMBL4657
CHEMBL1075322	CHEMBL1615381	CHEMBL1804	CHEMBL1865	CHEMBL4660
CHEMBL1163101	CHEMBL1741176	CHEMBL1808	CHEMBL1867	CHEMBL5084
CHEMBL1163125	CHEMBL1741186	CHEMBL1811	CHEMBL1868	CHEMBL5103
CHEMBL1255126	CHEMBL1741207	CHEMBL1821	CHEMBL1871	CHEMBL5113
CHEMBL1275212	CHEMBL1741215	CHEMBL1822	CHEMBL1873	CHEMBL5122
CHEMBL1287628	CHEMBL1781	CHEMBL1824	CHEMBL1878	CHEMBL5137
CHEMBL1293222	CHEMBL1782	CHEMBL1825	CHEMBL1881	CHEMBL5141
CHEMBL1293224	CHEMBL1785	CHEMBL1827	CHEMBL1889	CHEMBL5147
CHEMBL1293255	CHEMBL1787	CHEMBL1829	CHEMBL1892	CHEMBL5776
CHEMBL1833	CHEMBL1900	CHEMBL1947	CHEMBL1899	CHEMBL5794
CHEMBL1835	CHEMBL1901	CHEMBL1949	CHEMBL2003	CHEMBL5804
CHEMBL1836	CHEMBL1902	CHEMBL1951	CHEMBL2007	CHEMBL5600
CHEMBL1844	CHEMBL1903	CHEMBL1952	CHEMBL2007625	CHEMBL5608
CHEMBL1850	CHEMBL1904	CHEMBL1957	CHEMBL2008	CHEMBL5627
CHEMBL1853	CHEMBL1906	CHEMBL1908	CHEMBL2016	CHEMBL5646
CHEMBL1856	CHEMBL1907	CHEMBL1913	CHEMBL202	CHEMBL5650
CHEMBL1968	CHEMBL1966	CHEMBL1914	CHEMBL2028	CHEMBL5658
CHEMBL1916	CHEMBL203	CHEMBL1974	CHEMBL2243	CHEMBL5678
CHEMBL1917	CHEMBL2035	CHEMBL1977	CHEMBL225	CHEMBL5697
CHEMBL1918	CHEMBL2039	CHEMBL1978	CHEMBL2250	CHEMBL4767
CHEMBL1921	CHEMBL204	CHEMBL1980	CHEMBL226	CHEMBL4769
CHEMBL1929	CHEMBL2041	CHEMBL1981	CHEMBL2265	CHEMBL4777
CHEMBL1936	CHEMBL2047	CHEMBL1985	CHEMBL227	CHEMBL4789
CHEMBL1937	CHEMBL2055	CHEMBL1987	CHEMBL2276	CHEMBL4791
CHEMBL1940	CHEMBL2056	CHEMBL1991	CHEMBL2285	CHEMBL4792
CHEMBL1941	CHEMBL206	CHEMBL1994	CHEMBL2288	CHEMBL4793
CHEMBL1942	CHEMBL2061	CHEMBL1995	CHEMBL2292	CHEMBL4796
CHEMBL1944	CHEMBL2068	CHEMBL1997	CHEMBL230	CHEMBL5409
CHEMBL208	CHEMBL2069	CHEMBL2000	CHEMBL231	CHEMBL5443
CHEMBL2083	CHEMBL2073	CHEMBL2001	CHEMBL2318	CHEMBL5455
CHEMBL2085	CHEMBL2074	CHEMBL2002	CHEMBL2319	CHEMBL5469
CHEMBL209	CHEMBL232	CHEMBL220	CHEMBL2553	CHEMBL5485
CHEMBL210	CHEMBL2326	CHEMBL2208	CHEMBL256	CHEMBL5491
CHEMBL2107	CHEMBL233	CHEMBL221	CHEMBL2563	CHEMBL5493
CHEMBL211	CHEMBL2334	CHEMBL2216739	CHEMBL2568	CHEMBL6101
CHEMBL2219	CHEMBL2337	CHEMBL2123	CHEMBL258	CHEMBL6115
CHEMBL222	CHEMBL2343	CHEMBL213	CHEMBL2581	CHEMBL6120
CHEMBL2231	CHEMBL2345	CHEMBL2146302	CHEMBL259	CHEMBL6136
CHEMBL2147	CHEMBL2349	CHEMBL248	CHEMBL2593	CHEMBL5818
CHEMBL2148	CHEMBL235	CHEMBL2487	CHEMBL2595	CHEMBL5819
CHEMBL215	CHEMBL236	CHEMBL2492	CHEMBL2598	CHEMBL5847

CHEMBL216	CHEMBL237	CHEMBL250	CHEMBL2599	CHEMBL5855
CHEMBL2163176	CHEMBL2373	CHEMBL2508	CHEMBL260	CHEMBL4900
CHEMBL2169736	CHEMBL238	CHEMBL251	CHEMBL261	CHEMBL4973
CHEMBL217	CHEMBL2386	CHEMBL2514	CHEMBL2611	CHEMBL4977
CHEMBL2179	CHEMBL239	CHEMBL2525	CHEMBL2617	CHEMBL5024
CHEMBL218	CHEMBL2390810	CHEMBL2527	CHEMBL262	CHEMBL5027
CHEMBL2185	CHEMBL240	CHEMBL253	CHEMBL2635	CHEMBL5028
CHEMBL2189110	CHEMBL241	CHEMBL2534	CHEMBL2637	CHEMBL5038
CHEMBL2424	CHEMBL2413	CHEMBL2535	CHEMBL2652	CHEMBL5073
CHEMBL2426	CHEMBL2414	CHEMBL2543	CHEMBL2664	CHEMBL5703
CHEMBL2431	CHEMBL242	CHEMBL255	CHEMBL267	CHEMBL5719
CHEMBL2434	CHEMBL268	CHEMBL2820	CHEMBL2996	CHEMBL5742
CHEMBL2439	CHEMBL2689	CHEMBL2828	CHEMBL3004	CHEMBL5747
CHEMBL2468	CHEMBL2693	CHEMBL283	CHEMBL3009	CHEMBL5203
CHEMBL2474	CHEMBL2695	CHEMBL2850	CHEMBL301	CHEMBL5247
CHEMBL3553	CHEMBL2716	CHEMBL288	CHEMBL3012	CHEMBL5251
CHEMBL3559	CHEMBL2717	CHEMBL2888	CHEMBL3023	CHEMBL5857
CHEMBL3568	CHEMBL2730	CHEMBL2889	CHEMBL3024	CHEMBL5879
CHEMBL2731	CHEMBL289	CHEMBL3025	CHEMBL3231	CHEMBL5896
CHEMBL2736	CHEMBL2896	CHEMBL3032	CHEMBL3234	CHEMBL5903
CHEMBL2742	CHEMBL290	CHEMBL3045	CHEMBL3238	CHEMBL5936
CHEMBL275	CHEMBL2903	CHEMBL3055	CHEMBL3243	CHEMBL5938
CHEMBL2778	CHEMBL2916	CHEMBL3060	CHEMBL325	CHEMBL5971
CHEMBL2781	CHEMBL2938	CHEMBL3070	CHEMBL3250	CHEMBL5979
CHEMBL2782	CHEMBL2939	CHEMBL308	CHEMBL3267	CHEMBL5366
CHEMBL2789	CHEMBL2955	CHEMBL3094	CHEMBL3268	CHEMBL5378
CHEMBL279	CHEMBL2959	CHEMBL3106	CHEMBL3272	CHEMBL5393
CHEMBL2793	CHEMBL2964	CHEMBL3116	CHEMBL3286	CHEMBL5407
CHEMBL2801	CHEMBL2971	CHEMBL3130	CHEMBL3308	CHEMBL5408
CHEMBL2803	CHEMBL2973	CHEMBL3142	CHEMBL331	CHEMBL6009
CHEMBL2808	CHEMBL298	CHEMBL3145	CHEMBL3310	CHEMBL6014
CHEMBL2815	CHEMBL299	CHEMBL3180	CHEMBL332	CHEMBL6030
CHEMBL3181	CHEMBL333	CHEMBL3522	CHEMBL3710	CHEMBL6032
CHEMBL3192	CHEMBL3338	CHEMBL3524	CHEMBL3714130	CHEMBL5518
CHEMBL3201	CHEMBL335	CHEMBL3529	CHEMBL3717	CHEMBL5522
CHEMBL3202	CHEMBL3351	CHEMBL3535	CHEMBL3721	CHEMBL5524
CHEMBL321	CHEMBL3356	CHEMBL3864	CHEMBL3729	CHEMBL5543
CHEMBL3227	CHEMBL3357	CHEMBL3869	CHEMBL3746	CHEMBL5545
CHEMBL3230	CHEMBL3359	CHEMBL3880	CHEMBL3759	CHEMBL5568
CHEMBL3385	CHEMBL3589	CHEMBL3764	CHEMBL3886	CHEMBL6003
CHEMBL3397	CHEMBL3590	CHEMBL3772	CHEMBL3890	CHEMBL6007
CHEMBL3399910	CHEMBL3616	CHEMBL3776	CHEMBL3891	CHEMBL6154
CHEMBL340	CHEMBL3622	CHEMBL3778	CHEMBL3892	CHEMBL4895
CHEMBL3401	CHEMBL3629	CHEMBL3785	CHEMBL3898	CHEMBL4896
CHEMBL3426	CHEMBL3636	CHEMBL3788	CHEMBL3902	CHEMBL4897
CHEMBL3437	CHEMBL3650	CHEMBL3795	CHEMBL3905	CHEMBL4898
CHEMBL3438	CHEMBL3663	CHEMBL3807	CHEMBL3906	CHEMBL4899

CHEMBL3468	CHEMBL3683	CHEMBL3816	CHEMBL3911	CHEMBL4444
CHEMBL3474	CHEMBL3687	CHEMBL3819	CHEMBL3913	CHEMBL4461
CHEMBL3475	CHEMBL3691	CHEMBL3820	CHEMBL3920	CHEMBL4462
CHEMBL3476	CHEMBL3961	CHEMBL3829	CHEMBL3922	CHEMBL4465
CHEMBL3510	CHEMBL3965	CHEMBL3831	CHEMBL3935	CHEMBL4478
CHEMBL3514	CHEMBL3969	CHEMBL3835	CHEMBL3959	CHEMBL4481
CHEMBL3836	CHEMBL3972	CHEMBL4051	CHEMBL4203	CHEMBL4482
CHEMBL3837	CHEMBL3973	CHEMBL4068	CHEMBL4204	CHEMBL4501
CHEMBL3861	CHEMBL3974	CHEMBL4071	CHEMBL4223	CHEMBL4506
CHEMBL3863	CHEMBL3975	CHEMBL4072	CHEMBL4224	CHEMBL4801
CHEMBL3572	CHEMBL3976	CHEMBL4073	CHEMBL4225	CHEMBL4803
CHEMBL3582	CHEMBL3979	CHEMBL4079	CHEMBL4227	CHEMBL4804
CHEMBL3587	CHEMBL3982	CHEMBL4080	CHEMBL4234	CHEMBL4816
CHEMBL3983	CHEMBL4081	CHEMBL4237	CHEMBL4422	CHEMBL4581
CHEMBL3991	CHEMBL4093	CHEMBL4247	CHEMBL4426	CHEMBL4599
CHEMBL4005	CHEMBL4101	CHEMBL4261	CHEMBL4427	CHEMBL4600
CHEMBL4015	CHEMBL4123	CHEMBL4270	CHEMBL4439	CHEMBL5261
CHEMBL4016	CHEMBL4128	CHEMBL4273	CHEMBL4441	CHEMBL5282
CHEMBL4018	CHEMBL4142	CHEMBL4282	CHEMBL4714	CHEMBL5285
CHEMBL4026	CHEMBL4145	CHEMBL4296	CHEMBL4718	CHEMBL5314
CHEMBL4029	CHEMBL4147	CHEMBL4302	CHEMBL4722	CHEMBL5330
CHEMBL4036	CHEMBL4158	CHEMBL4303	CHEMBL4761	CHEMBL5331
CHEMBL4040	CHEMBL4176	CHEMBL4306	CHEMBL4766	CHEMBL6164
CHEMBL4045	CHEMBL4179	CHEMBL4315	CHEMBL4608	CHEMBL6166
CHEMBL4374	CHEMBL4191	CHEMBL4338	CHEMBL4617	CHEMBL6175
CHEMBL4375	CHEMBL4198	CHEMBL4361	CHEMBL4618	CHEMBL4698
CHEMBL4376	CHEMBL4202	CHEMBL4367	CHEMBL4625	CHEMBL4699
CHEMBL4393	CHEMBL4508	CHEMBL4662	CHEMBL4630	CHEMBL4852
CHEMBL4394	CHEMBL4516	CHEMBL4674	CHEMBL4576	CHEMBL4829
CHEMBL4398	CHEMBL4523	CHEMBL4681	CHEMBL4578	CHEMBL4835
CHEMBL4408	CHEMBL4525	CHEMBL4683	CHEMBL4708	CHEMBL4601
CHEMBL4822	CHEMBL4575	CHEMBL4685		

# Virtual Screening with Generative Topographic Maps: How Many Maps Are Required?

Iuri Casciuc,[†] Yuliana Zabolotna,[†] Dragos Horvath,[‡] Gilles Marcou,[†] Jürgen Bajorath,[‡] and Alexandre Varnek*,[†]

[†]Laboratoire de Chémoinformatique UMR 7140 CNRS, Institut LeBel 4, rue B. Pascal 67081 Strasbourg, France
[‡]B-IT, Limes, Unit Chem. Biol. & Med. Chem., University of Bonn, 53115 Bonn, Germany

S *Supporting Information*

ABSTRACT: Universal generative topographic maps (GTMs) provide two-dimensional representations of chemical space selected for their "polypharmacological competence", that is, the ability to simultaneously represent meaningful activity and property landscapes, associated with many distinct targets and properties. Several such GTMs can be generated, each based on a different initial descriptor vector, encoding distinct structural features. While their average polypharmacological competence may indeed be equivalent, they nevertheless significantly diverge with respect to the quality of each property-specific landscape. In this work, we show that distinct universal maps represent complementary and strongly synergistic views of biologically relevant chemical space. Eight universal GTMs were employed as support for predictive classification landscapes, using more than 600 active/inactive ligand series associated with as many targets from the ChEMBL database (v.23). For nine of these targets, it was possible to extract, from the Directory of Useful Decoys (DUD), truly external sets featuring sufficient "actives" and "decoys" not present in the landscape-defining ChEMBL ligand sets. For each such molecule, projected on every class landscape of a particular universal map, a probability of activity was estimated, in analogy to a virtual screening (VS) experiment. Cross-validated (CV) balanced accuracy on landscape-defining ChEMBL data was unable to predict the success of that landscape in VS. Thus, the universal map with best CV results for a given property should not be prioritized as the implicitly best predictor. For a given map, predictions for many DUD compounds are not trustworthy, according to applicability domain considerations. By contrast, simultaneous application of all universal maps, and rating of the likelihood of activity as the mean returned by all applicable maps, significantly improved prediction results. Performance measures in consensus VS using multiple maps were always superior or similar to those of the best individual map.
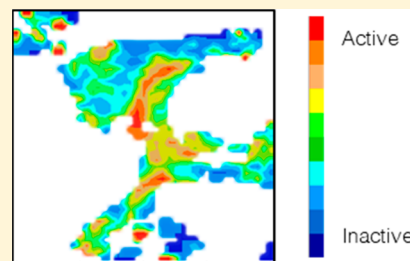
INTRODUCTION

We are currently facing a growing problem with "big data" in many areas, and chemistry is not an exception. Currently, an ensemble of academic, commercial, and propriety databases records more than 100 million compounds.[1] An estimation of the drug-like chemical space size gives us around 10^{33} virtual compounds.[1] Hence, selection of potential drug molecules from vast collections of candidate compounds is a real challenge for medicinal chemists.

Chemical information is intrinsically multidimensional, as it may alternatively focus on, for example, connectivity, electronic cloud densities, shape, or pharmacophore patterns, and each aspect may prove to be very important for understanding chemical properties and biological activities. These various properties can be encoded by specific molecular descriptors, that is, specific vectors of N numbers derived from chemical structure, thus representing a molecule as a point in N-dimensional descriptor space. In principle and at arbitrarily high N, this conceptual space may contain almost all known information about molecules, which, in theory, should allow researchers to predict any desired properties using already obtained experimental values as a training input. However, it is impossible to handle such amounts of information without advanced data mining techniques. Even though a variety of methods exist,[2,3] the main difficulty is striving for a balance between the accuracy of the results and the computational cost of the required calculations.

One of the techniques that is well-suited to reach this balance is generative topographic mapping[4] (GTM), a nonlinear mapping method that is widely used as a visualization tool for analysis of a multidimensional space. GTM landscapes have already been used as quantitative structure−activity relationship (QSAR) models,[5−7] and their predictive performance in virtual screening (VS) tends to increase with the size and diversity of the data set used to "color" the landscape. GTM was successfully used for structure−activity analysis of an antiviral compound set[8] and also of an antimalarial mode of action database.[9] Recently, it has also been successfully applied to visualize large public chemical databases such as PubChem, ChEMBL,[10] and FDB-.[11] Sidorov et al.[12] applied GTM to create "universal" maps of chemical space that easily distinguished active and inactive compounds for more than 400 ChEMBL targets,

yielding an averaged balanced accuracy (BA) higher than 0.6 for all targets, indicating high potential of this method for such applications.

The advantage of universal GTM models over classical QSAR approaches is that the most relevant descriptor space that guarantees polypharmacological competence and preferred operational parameter settings defining the manifold are "learned" only once, at the map construction stage. At this stage, large random collections of relevant (drug-like) compounds are used to span biologically relevant chemical space, serving as a "frame set" for unsupervised GTM manifold fitting, while a large and diverse ensemble of structure–activity sets are employed as "selection sets". Their role is to score the quality of the current manifold for its ability to host predictive landscapes corresponding to each selection set activity. Top manifolds scoring well at this stage are selected as the final "universal" maps, with the expectation that they will also be able to support predictive landscapes for other, distinct properties beyond those present in the selection set. This expectation was well met by more than 400 structure–activity sets consisting of novel compounds associated with completely unrelated targets and properties by Sidorov et al.[12] Certainly, dedicated models that might be built for a given property could exceed the predictive power of universal GTM-based property landscapes—if sufficient training data are available. By contrast, universal GTM manifolds act like "default", zero-parameter models that can even be employed to explore scarcely studied properties with little experimental data. Therefore, they are both the best strategy to use with incipient, small structure–activity series and an economic, rapid, fitting-free approach to model building for large and diverse series.

GTM-based property prediction is unavoidably penalized by the dimensionality reduction step and the inescapable loss of information it implies. Projecting the multidimensional items (molecules for which high-dimensional descriptor elements each capture specific structural features) onto a two-dimensional (2D) latent grid is expected to mechanically reduce the predictive power, compared to any ideal machine learning method that operates in the original descriptor space. Nevertheless, previous studies[9,10,12–15] have typically shown that GTM-driven classification or regression models are on par or only slightly less predictive than equivalent support vector machine or random forest approaches.

However, "universal" GTMs like the ones advocated here were conceived to cover the entire drug-like chemical space. Like any global maps, their resolution is expectedly lower than the one that could be achieved by dedicated GTMs, focusing on specific series of compounds. The key question addressed in this work is whether such global maps, primarily conceived to serve as a rather coarse-grained "atlas" of the various structural motifs explored in to-date medicinal chemistry,[10,12,14] may nevertheless be successfully exploited as an accurate virtual screening and property prediction tool. This is envisaged by means of a consensus predictor using several universal maps, built on distinct initial descriptor spaces capturing distinct chemical information. Therefore, information lost on a given map may still be preserved by the others. If so, a strong synergetic (consensus) effect of their individual predictions might compensate all the above-mentioned drawbacks of "universal" GTM-driven virtual screening.

In this work, we assess the predictive performance of eight newly constructed universal GTM models in VS of nine target-specific compound sets extracted from Directory of Useful Decoys (DUD).[16] These GTMs have been constructed on the basis of ChEMBL[17] (v.23) structure–activity data for the respective targets; each is based on a different initial descriptor vector, encoding distinct structural features. Their average polypharmacological competence is (roughly) equivalent; they are all members of the top-ranked population produced by the evolutionary map-building process. Nevertheless, they significantly differ in the quality of each property-specific landscape. We show that distinct universal maps represent complementary and strongly synergistic view of chemical space. The predictive power of any classification landscape built for ChEMBL data can be internally assessed by the cross-validated balanced accuracy (BA_{CV}) criterion in an "aggressive" 3-fold cross-validation experiment repeated five times, with data scrambling. However, the BA_{CV} indices were shown to be unable to predict the success of that landscape in VS. Thus, it would be an error to prefer the universal map with best CV results for a given property as the implicitly best predictor. For a given map, predictions for many DUD compounds are not trustworthy, according to applicability domain (AD) considerations. By contrast, simultaneous application of all universal maps, and rating of the likelihood of activity as the mean returned by all applicable maps, significantly improved prediction results. On the basis of a different measure, the performance of consensus maps in VS was consistently better than that of individual maps.

■ METHODS

Data. The target-specific compound series extraction protocol by Sidorov[12] has been applied to release 23 of the ChEMBL database. A total of 618 data sets containing ligands of different ChEMBL human targets have been extracted. The same structure standardization procedure (vide infra) has been applied to DUD database, followed by removal of molecules that were present in ChEMBL to create orthogonal external data sets. For most of the targets shared by ChEMBL and DUD, this required elimination of all the actives from the DUD series. However, in nine cases the DUD target-specific series contained sufficiently numerous original actives and were used for VS. Table 1 summarizes the composition of selected compound data sets.

Table 1. Description of Target-Specific Subsets Used for Model Training (ChEMBL) and VS (DUD)

| ChEMBL ID	target name	DUD data set		ChEMBL data set	
		active	inactive	active	inactive
1827	phosphodiesterase 5A	170	25 334	691	1515
1952	thymidylate synthase	63	6 113	124	455
251	adenosine A2a receptor	79	28 001	1303	3618
260	MAP kinase p38 alpha	100	32 925	1453	2567
279	vascular endothelial growth factor receptor 2	94	22 595	2047	4663
301	cyclin-dependent kinase 2	189	25 675	638	2305
4282	serine/threonine-protein kinase AKT	52	14 228	725	2619
4338	purine nucleoside phosphorylase	102	6 334	100	111
4439	TGF-beta receptor type I	82	8 013	282	385
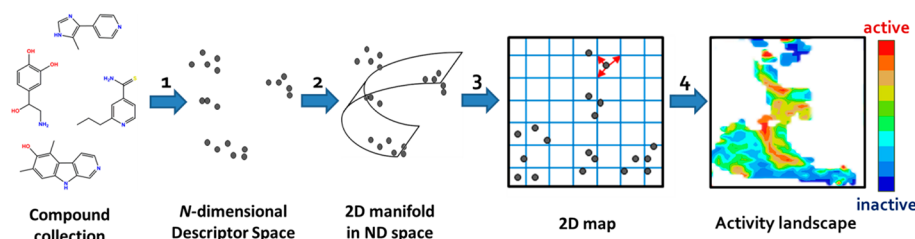
Figure 1. A frame set of compounds is represented in the *N*-dimensional descriptor space. A flexible 2D manifold, which is a square grid of nodes, is injected into that space and is fitted to the data. The molecules are nonlinearly projected onto it, and when the manifold is unbent, a 2D map is obtained. Each node can be colored according to the activities of molecules residing there, producing "activity landscapes", where red zones are populated only by active molecules and blue by inactive; all colors in between correspond to the regions occupied by compounds of both classes in different proportions. White zones are empty.

Note that in Table 1, the "actives" in the ChEMBL data set represent the topmost potent compounds accounted for in the ChEMBL database, according to their specific activity measure(s), IC_{50} or K_i values. As mentioned in the original paper by Sidorov, the cutoff value required to qualify as active was chosen, for each series, from three possible options: 50 nM, 100 nM, or 1 μM. The retained, series-specific thresholds were the ones leading to the best balance of actives versus inactives in ChEMBL sets, optimally including 20% of actives and 80% of inactives. Recall that inactives, in this context, are compounds with activities weaker than the 10-fold of the threshold, while "gray zone" compounds between were ignored. For DUD compounds, the definition of "actives" is the one proposed by the original authors of these sets, while inactives are, presumably inactive, decoy molecules.

Workflow. The following workflow was applied:

(1) Standardization of ChEMBL and DUD data sets followed by descriptor generation;

(2) Coloring the manifolds of universal maps by each of nine target-specific class landscapes using ChEMBL subsets;

(3) 3-fold cross-validation of predictive landscapes within the ChEMBL data sets;

(4) Application of these landscapes for the VS of the corresponding DUD subsets

For some of these steps a dedicated section is presented below.

Data Preparation and Descriptor Generation. Structures from both databases ChEMBL (version 23) and DUD were standardized according to the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics in the University of Strasbourg (infochim.u-strasbg.fr/webserv/VSEngine.html) using the ChemAxon Standardizer:[18]

- Dearomatization and final aromatization according to the "basic" setup of the ChemAxon procedure (heterocycles like pyridone are not aromatized)
- Dealkalization
- Conversion to canonical SMILES
- Removal of salts and mixtures
- Neutralization of all species, except nitrogen(IV)
- Generation of the major tautomer according to ChemAxon

After the standardization, 1 540 615 compounds from ChEMBL and 914 379 compounds from DUD remained.

The descriptors used here were ISIDA descriptors computed by ISIDA Fragmentor.[19−21] More than 100 different types of descriptors sets were generated. They include sequences, atom pairs, circular fragments, and triplet counts of different length, colored by formal charges, pharmacophore features, or force field types. These fragmentation schemes were selected for the relatively low number of fragments they generate.

Generative Topographic Mapping. Generative Topographic Mapping (GTM) is a nonlinear mapping method used for data visualization originally described by Bishop. In GTM, 2D latent space (called manifold) is embedded into the descriptor space. The points that are close in the latent space remain neighbors in the data space. The manifold represents a grid of $k \times k$ nodes; each node is mapped in the initial descriptor space using the mapping function $y(x, \mathbf{W})$. The mapping function is given as a grid of $m \times m$ radial basis functions (RBFs). To build a GTM-based QSAR model, the weighted average of properties of all molecules associated with any particular node is used to "color" the manifold according to that property. Here, the projected property is activity class membership, resulting in a fuzzy activity landscape (Figure 1). Molecule "responsibilities" are used as weights. Red and blue zones are populated by only active and inactive compounds, respectively; all colors in between correspond to the regions occupied by compounds of both classes in different proportions. White zones represent unpopulated areas.

GTM supports several applicability domain (AD)[6] definitions, but only the density-based AD is applied here. Compounds projected onto a "white zone" of the map (accumulating no responsibilities of "training" compounds used to build the landscape) are out of the AD.

Note, however, that the AD considerations in VS may differ from those in predictive QSAR. In the latter case, compounds outside of the AD should be ignored; no prediction of their property should be attempted. In VS, however, the inability to obtain a trustworthy prediction for out-of-AD compounds practically implies that those compounds will be never selected for synthesis and testing as if they were predicted to be inactive. Therefore, external compounds falling within the blank spots of the employed class landscapes were assigned zero probability to be active, placing them at the bottom of rankings.

Global manifolds (universal maps) were derived following the procedure in ref 12 but employing updated compound data sets. They are based on frame sets of maximal diversity (aimed at spanning the entire drug-like chemical space) and employed 236 (randomly picked) of the above-mentioned 618 compound series for map selection. As in any global mapping approach, they are not meant to capture the detailed SAR of every target-specific set but allow analysis of several activities at the same time. Note that global activity landscapes are relying on a common manifold, itself derived from a selected

Table 2. Description of Eight Universal Maps, Their Descriptor Types, and the Descriptor Space Dimensionality

| map	abbreviation	definition	descriptor space dimensionality
1	IA-FF-FC-AP-2-3	sequences of atoms with a length of 2−3 atoms labeled by force field types and formal charge status, using all paths.	5161
2	IIRAB-FF-1-2	atom-centered fragments of restricted atom and bonds of 1−2 atoms labeled by force field types	3172
3	IAB-PH-FC-AP-2-4	sequences of atoms and bonds of a length 2−4 atoms labeled by pharmacophoric atom types and formal charges using all paths	4245
4	IA-2-7	sequences of 2−7 atoms.	6520
5	IAB-FC-AP-FC-2-4	sequences of atoms and bonds of 2−4 atoms labeled by formal charge, using all paths	3437
6	IA-FF-P-2-6	sequences of atom pairs with a length of 2−6 intercalated bonds, labeled by Force Field type	2901
7	III-PH-3-6	atom triplets labeled by pharmacophoric atom types with topological distance from 3 to 6 bonds	4846
8	III-FF-3-4	atom triplets labeled by force field types, with topological distance from 3 to 4 bonds	8953
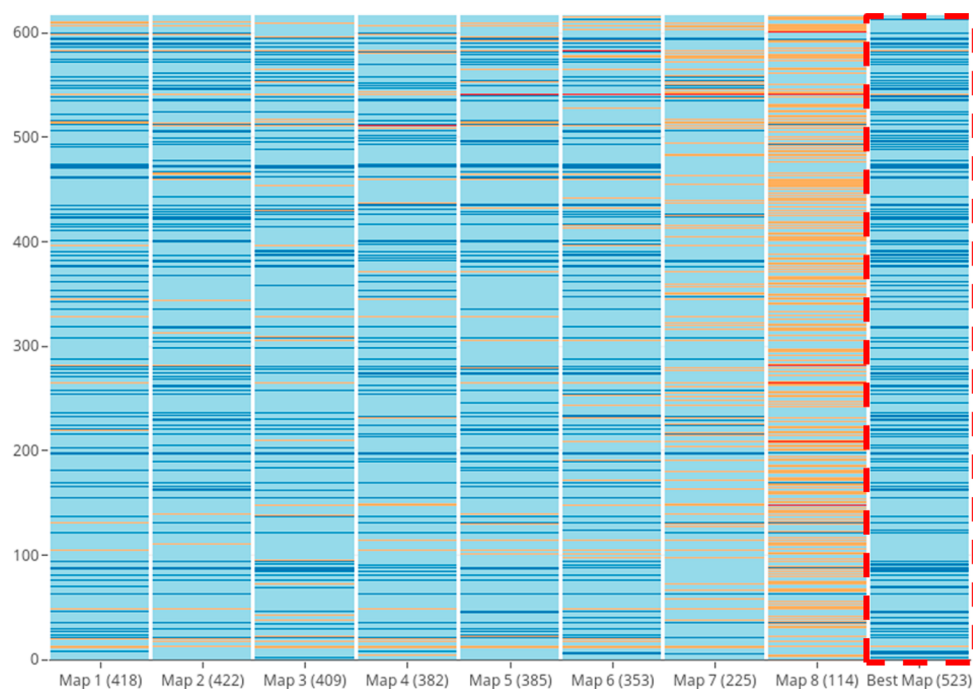
Figure 2. Heatmap showing the performance of universal maps on 618 selected series. Color-codes: dark blue, BA > 0.85; light blue, 0.65 < BA ≤ 0.85; orange, 0.5 < BA ≤ 0.65; and red, BA ≤ 0.5. Between parentheses is shown the number of target-specific classification problems for which a map scores BA > 0.75.

descriptor space in order to maximize the mean predictive power of all these landscapes. It is obvious that global manifolds represent a best compromise to describe biological activity in general, based on some "consensus" descriptor space. Interestingly, several such descriptor spaces were identified, each focusing on different aspects of chemical structures. Eight global (universal) maps based on eight distinct ISIDA fragment descriptor spaces were selected (Table 2). On average, their mean predictive power over all the 618 considered activity sets is similar, while corresponding predictions for each activity series fluctuate.

Performance Evaluation. Model performance was evaluated using BA in 3-fold CV and VS, receiver operating characteristic area under curve (ROC AUC) in VS, and enrichment factor (EF) in VS. BA has been mainly used during cross-validation. BA serves to assess the ability of landscapes to predict the correct activity class of candidates not used for landscape construction, that is, both in "internal" cross-validation and "external" VS. Note that reported BA scores for individual maps, both in CV and in VS applications, are always calculated on the entire concerned sets, including

species projected into empty map zones (out of applicability domain) and which were considered, by default, inactives.

However, ROC AUC is a more natural VS evaluation criterion than BA, because the latter requires a formal prediction, active versus inactive, for each external compound. In VS, however, the key element is the relative ranking of candidates; a significant prioritization of the actives with respect to the inactives is sufficient to guarantee VS success. Ranking was performed according to the GTM landscape-predicted probability of each compound to be active. The compounds falling outside the applicability domain were assigned zero probability of activity; thus, they were placed at the bottom of the ranking list.

To complement ROC AUC values, the EF of actives ranked within the 100 top compounds was also monitored. EF for the top 100 ranked molecules was calculated according to the equation

$$EF_{100} = \frac{\text{Actives}_{100}/100}{\text{Actives}_{\text{total}}/N_{\text{total}}}$$

where Actives$_{100}$ is the number of true positives in the top 100 compounds, Actives$_{total}$ the total number of active compounds in the data set, and N_{total} the total number of compounds in the data set.

However, selection of the top 100 compounds may be considered only if there is a significant gap between the probabilities to be active of the 100th selected compound and that of the 101st not-selected candidate. In practice, several candidate compounds will have the same predicted probability to be active (reported with a precision of 0.01), and therefore, all those that are equiprobable to the 100th selected compound would be equally deserving to enter the selection. In order to force selection of a top 100 compounds, a random subset of these equiprobable must be picked in completion of the better ranked candidates. In this a posteriori study, three scenarios are considered to compute the EF:

(1) Pessimistic: out of candidates that are equiprobable to the 100th selected compound, inactives are selected first, and then the remaining places in the pessimistic top 100 are completed by actives.
(2) Optimistic: the opposite strategy (actives are filled in first, remaining places taken by inactives).
(3) Stochastic pick out of candidates that are equiprobable to the 100th selected compound.

Scenarios 1 and 2 are deterministic. The values obtained are termed pessimistic enrichment factor (PEF) and optimistic enrichment factor (OEF), respectively. Scenario 3 is not deterministic, and repeated random drawing/averaging would be required to converge to expectation values. Yet, it is possible to estimate an average value, termed interpolated enrichment factor (IEF) using the following equation:

$$IEF = \lambda \times PEF + (1 - \lambda) \times OEF$$

$$\lambda = \frac{n}{N}$$

where IEF is the interpolated enrichment factor; OEF the optimistic enrichment factor; PEF the pessimistic enrichment factor; and λ the ratio n/N, with N being the size of set including all the candidates that are equiprobable to the 100th selected compound and n the number of these latter candidates. For instance, if the set including all four candidates that are equiprobable to the 100th selected compound contains 102 hits, then $N = 4$ and $n = 2$ such that $\lambda = 0.5$.

■ RESULTS

Cross-Validation of ChEMBL Activity Class Landscapes. Three-fold CV of the BA was repeated five times for each of the ChEMBL series. For the 236 randomly picked "selection" series, this was part of the GTM manifold scoring process, where the fitness score reflects the mean of each BA$_{CV}$ value. For the eight selected manifolds, the same CV procedure was applied to the remaining 618 − 236 "external" series, thus obtaining the complete matrix of the predictive power of every map for each of the 618 (Figure 2). Unsurprisingly, not every property is equally well predicted by each map, although the average BA$_{CV}$ value may not differ much from map to map. Each map was examined in order to identify the number of targets for which it is able to solve the active/inactive classification problem at BA$_{CV}$ above a given threshold.

Figure 3 shows that for 617 of 618 targets, BA$_{CV}$ scores of 0.6 or better are achieved by at least one of the maps. The exception (CHEMBL5678) represents a set with too few
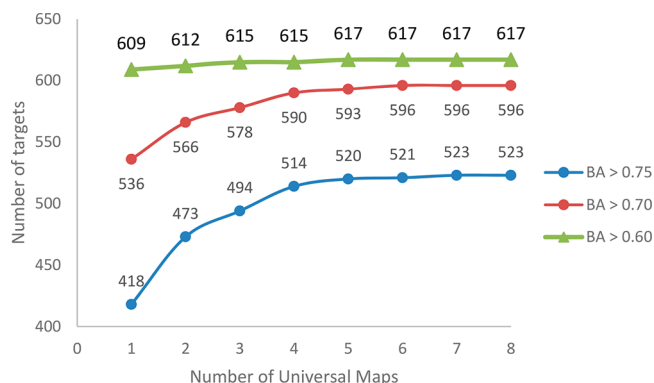

Figure 3. Cumulated performance of universal maps: number of predicted target-specific series vs number of used maps.

compounds. Note that maps are ranked according to their original fitness score (mean BA$_{CV}$ scores over the 236 selection SAR series), and it can be seen from Figure 3 that the first map is strongly predictive (BA$_{CV}$ > 0.75) for 418 distinct series. Note that part of these 418 are selection series but include a significant number of external series nevertheless. It is also noteworthy that every single map is able to provide significantly better-than-random separation of actives and inactives (BA$_{CV}$ > 0.6) for virtually all (609/618, in the case of map 1) SAR sets, which fully justifies the label of "universal" maps. However, no single map is expected to flawlessly model all series; no single descriptor space (fragmentation scheme) on which a map is built could capture all the relevant chemical information that might impact so many different structure–activity relationships. The eight selected maps are highly complementary: series less well explained by one map will work better on another manifold, exploiting specific information from its distinct descriptor space to host a strongly predictive model. Cumulated prediction performance increases with the number of considered maps (Figure 3), which clearly demonstrates map complementarity: Seven universal maps based on as many distinct descriptor spaces are sufficient to provide at least one satisfactory result for more than 85% of used targets even at the very stringent BA$_{CV}$ > 0.75. Thus, for further analysis, only seven universal maps were used.

Is BA$_{CV}$ a Reliable Indicator of VS Success? Next, the question how to identify the best universal map for a particular activity was addressed. It may be expected that the model that shows highest predictive CV performance in target-specific ligand classification would be the best model in VS. To test this hypothesis, correlation between landscape performance in CV and VS was evaluated for each of the 63 QSAR models (activity landscapes for nine targets on seven universal maps). Figure 4 compares, for the specific activity landscapes of target CHEMBL260 hosted on each map, the "internal" estimation predictive power (BA$_{CV}$) on one hand and the observed predictive power in "external" VS of the DUD subset on the other hand.

The Pearson correlation coefficient of BA$_{CV}$ versus BA$_{VS}$ over the seven maps was calculated for all nine sets; they vary in the range of 0.02−0.63, which means that a map can hardly be chosen on the basis of its CV performance. Unfortunately, but not unexpectedly,[22] high BA$_{CV}$ is a necessary but not sufficient guarantee of model success in VS. The success in a
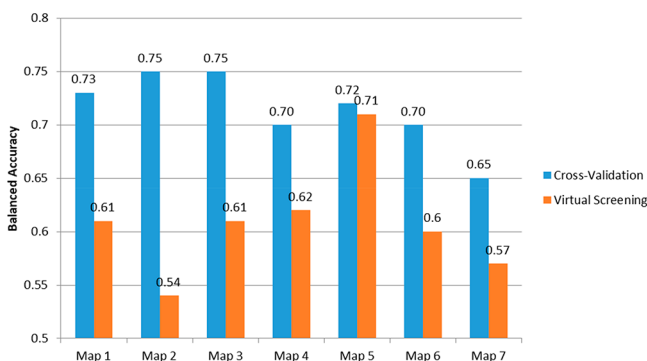
Figure 4. BA values obtained in CV and VS of the CHEMBL260 data set.

predictive challenge depends on the peculiar composition of the test set.

Consensus of Universal Maps. Given the genuine complementarity of the seven maps, consensus predictions by averaging results of these complementary views of chemical space might be a promising strategy. Here, for each compound from the external test set, averaging was applied to the predicted probability of being "active" over the seven landscapes, *excluding*, however, landscapes in which the compound is projected into an "empty" zone (Figure 5). In this study, the density-based AD criterion as implemented by default in ISIDA GTM was applied.[6] Compounds that fell
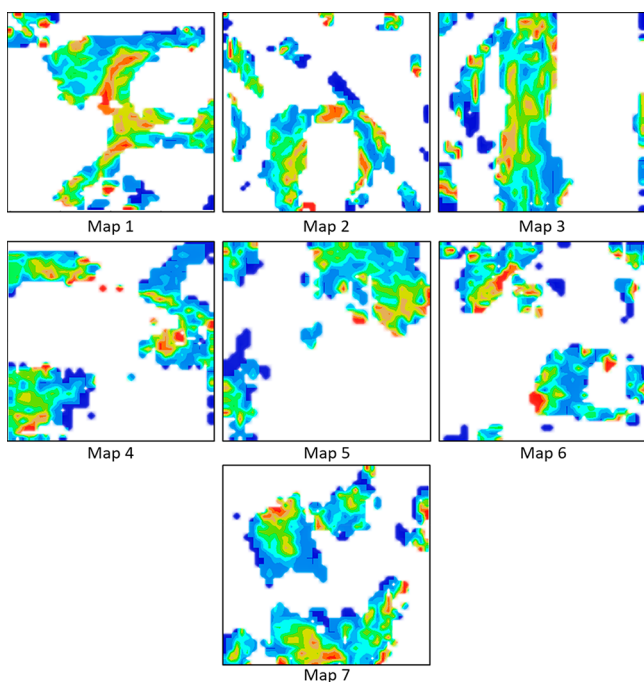

Figure 5. Activity class landscapes of the CHEMBL260 data set in seven universal maps numbered according to Table 2. Because the seven latent spaces are independent projections of distinct initial chemical spaces, these activity class landscapes cannot be "overlaid" to produce a "consensus" landscape. Instead, consensus predictions are obtained by placing the item to predict on each of these activity class landscapes and estimating, if its projection falls within a densely populated region (with the AD), its probability to be "active", then taking the average of these estimated probabilities.

outside the AD on all the maps were considered, by default, as inactives.

Apart from the fact that consensus allows making predictions without choosing a priori one best map, it has another important advantage: data coverage increase (percentage of the compounds that are considered to be in AD). For example, none of the maps of the CHEMBL260 subset provided 100% data coverage achieved by the consensus. Similar observations were made for the remaining eight data sets. Only for two was coverage less than 100% (CHEMBL4338, 79.8%; CHEMBL4439, 97.5%). Recall that in a VS context, compounds out of AD are not "discarded" but given a probability of zero to be active, which implicitly ranks them at the bottom of the list. Thus, data coverage in this context does not impact the size of the screened compound set (BA, EF, and ROC AUC values are reported with respect to the full DUD sets, respectively). Data coverage, however, impacts the reliability of results because increasing data coverage reduces compounds with zero probability of activity.

Figure 6 shows that consensus BA values generally exceed the majority of BA scores achieved by individual universal maps. Only universal map 5 outperformed the consensus model for CHEMBL260 in terms of balanced accuracy, but not with respect to ROC AUC or EF.

In terms of EF, no individual model except universal map 4 was able to rank any of the active compounds from DUD into the top 100. For the universal map 4, EF = 2.87 corresponded to a single active compound in the top 100. However, the EF for the consensus model reached 11, which resulted from five true actives in the top 100.

The results for all nine data sets are shown in Table 3. The consensus model performed better than any individual map on the basis of EF.

To understand the strengths and limitations of GTM-driven prediction, please recall that GTM activity class landscapes are obtained by "transferring" the knowledge about the most likely class to be encountered in a given chemical space neighborhood onto the latent grid nodes "representing" that neighborhood. Conversely, prediction implies locating the candidate into one of these "standard" neighborhoods represented by nodes, therefrom learning the class to which it should be assigned. GTM-driven predictors quintessentially behave like nearest-neighbor-based predictors, including support for identification of candidates outside of its applicability domain, that is, species which do not sufficiently resemble to any of the reference compounds, in order to allow an extrapolation of their properties by virtue of the similarity principle. The complementarity of the seven universal maps largely reflects the complementarity of the similarity principle focused on distinct and different structural aspects. Candidates discarded as not similar enough (out of AD) with respect to some structural aspects were correctly recognized as significantly similar with respect to some different aspects. Note that some of the maps are built on hand of descriptors (detailed atom-centered fragments) capturing connectivity information, while other rely on fuzzier atom pair counts and last but not least on topological pharmacophore descriptors. If reference compounds of ChEMBL are obviously related to the active DUD examples (they are members of a same series, with roughly the same scaffold and same pharmacophore pattern), then several universal maps will provide a robust "detection" of the related DUD actives within the, in terms of generic chemotype, very distinct decoys. If, however, DUD actives are
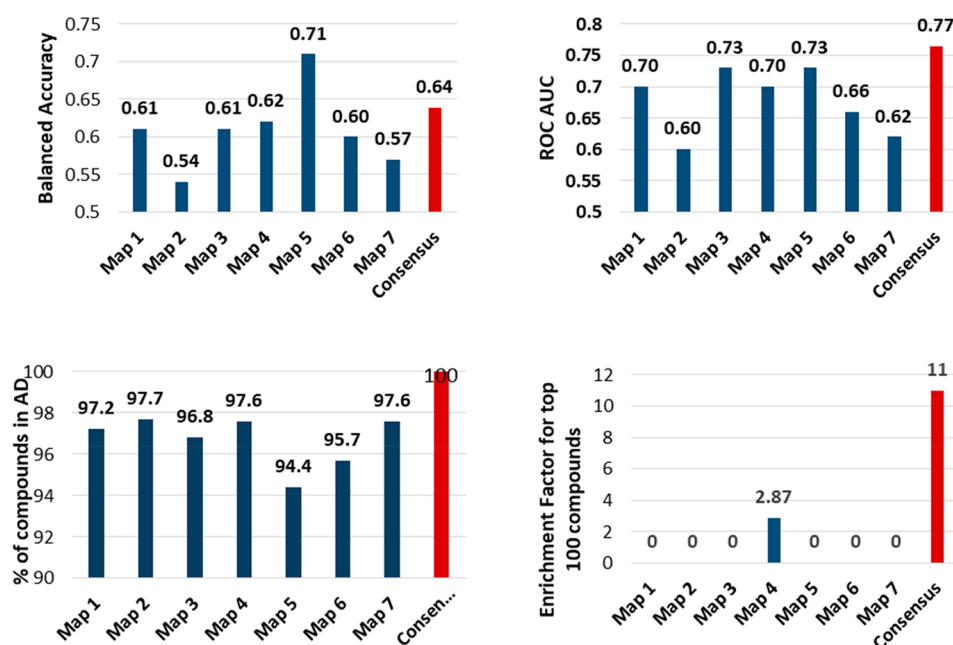
Figure 6. Performance of VS on DUD with the models developed for the CHEMBL260 data set assessed on the basis of BA (top left), ROC AUC (top right), data coverage (bottom left), and EF calculated for top 100 compounds (bottom right)

Table 3. Performance in CV and VS for Individual Universal Maps Compared to Consensus Models

	cross-validation		virtual screening				consensus model		
target	best map:	BA	best map:	BA	ROC AUC	EF	BA	ROC AUC	EF
CHEMBL1827	4	0.82	7	0.70	0.73	0.00	0.67	0.74	1.5
CHEMBL1952	4	0.83	5	0.82	0.85	0.13	0.82	0.86	14.7
CHEMBL251	2	0.77	3	0.77	0.84	1.56	0.80	0.88	17.8
CHEMBL260	2	0.75	5	0.71	0.73	0.00	0.64	0.77	11.00
CHEMBL279	2	0.73	4	0.71	0.78	0.00	0.66	0.82	4.83
CHEMBL301	3	0.80	5	0.74	0.80	0.60	0.81	0.87	5.47
CHEMBL4282	5	0.81	3	0.81	0.87	17.39	0.83	0.92	52.18
CHEMBL4338	5	0.83	3	0.71	0.73	0.00	0.54	0.66	0.00
CHEMBL4439	5	0.81	5	0.75	0.88	1.97	0.67	0.88	4.94

only partially related to the ChEMBL reference molecules, then only the maps able to recognize the specific underlying similarity will be competent solvers of the challenge. At one extreme, candidates may be scaffold-hopping analogues of reference compounds, typically not perceived as similar by the human eye. In this case, maps focusing on connectivity-based similarity criteria would also exclude the candidates (as well as the decoys) from their AD. Pharmacophore descriptor-based maps will, by contrast, successfully distinguish them from the random pharmacophore patterns of decoys. However, a fuzzier definition of neighborhood increases the risk of fortuitously co-opting decoys into the active neighborhood of the maps. Last but not least, it is important to highlight that similar activity of two compounds does not imply any underlying structural similarity: two actives may have both distinct topologies and distinct pharmacophores, because they bind to different (sub)pockets of the active site. Such examples of radical "binding paradigm shifts" cannot be foreseen by machine-learned models, in general.

In light of the numerous factors impacting the predictive power of GTM landscapes, it may be very difficult to highlight a detailed explanation for the specific prediction successes and failures observed here. In the following, the predictive behavior for target CHEMBL4338 (purine nucleoside phosphorylase,

the one exception for which no conclusive synergy effect of the individual maps was observed) has been analyzed in more detail. The herein used DUD set features 102 purine-like actives and 6334 decoy compounds.

Among the latter, a rather large subfamily of 580 phenylsulfonamides and -anilides was specifically scrutinized, as representing the "typical" set of decoys medicinal chemists would easily agree that clearly differ from the purine-like reference representatives of the ChEMBL data set. Their predicted status has been monitored (Table 4) on each map is reported next to map-specific CV and VS statistical parameters.

The ChEMBL series used to build the activity landscape mainly contained fused aromatic heterocycles such as hypoxanthine, pyrrolopyrimidine, and benzimidazole-4,7-quinone (Figure 7). In the DUD series, the majority of compounds that were correctly predicted contained a purine moiety similar to training set molecules.

A first intriguing observation is that maps 5 and 6, with better-than-random but rather deceiving VS results in terms of balanced accuracy, record outstanding VS results according to the ROC AUC criterion. This is no contradiction, merely a reminder that no single statistical criterion may claim the status of absolute measure of model quality. BA scores contribute to accurate prediction of activity class. However, this parameter

Table 4. Detailed Statistical Parameters of the Seven Universal GTM Models for Target CHEMBL4338

map number	cross-validation		virtual screening		prediction of the 580 phenylsulfonamide decoys		
	BA	ROC AUC	BA	ROC AUC	out of AD	inactive	active
1	0.75	0.81	0.62	0.86	579	0	1
2	0.71	0.79	0.61	0.73	333	245	2
3	0.72	0.81	0.71	0.73	567	9	4
4	0.70	0.79	0.64	0.74	475	98	7
5	0.83	0.87	0.68	0.96	578	2	0
6	0.72	0.78	0.66	0.90	568	12	0
7	0.75	0.82	0.42	0.50	32	497	51

suffers from the binarization artifact of the continuous likelihood to be active, which is not the case for ROC AUC. Also, note that active/inactive classification is intrinsically empirical: a compound that counts as "active" (low μM) in an incipient phase of a drug discovery project will be later discarded as "inactive", in contrast to the lately optimized low nanomolar binders. In this work, training set (ChEMBL) compounds were labeled as active/inactive in a context-dependent way, according to a threshold that was raised for the series rich in strong binders. The test compounds of DUD are assigned "active" status according to different standards and by contrast to, presumedly, inactive decoys. The fact that VS is able to prioritize these, in spite of potential incoherence in activity class flagging strategies, is per se a nontrivial observation, highlighting the robustness of classification models.

Furthermore, Table 4 outlines a clear negative correlation between the number of wrongly predicted "active" phenyl sulfonamides and the ROC AUC score in VS. This is, of course, not only due to the cited compounds being misplaced on the ROC curves but illustrates the above-discussed effect of the different "perceptions" of neighborhood provided by each map. As mentioned, phenyl sulfonamides appear as clearly distinct from the CHEMBL purine-like reference compounds, actives, or inactives alike. From the medicinal chemist's point of view, these are expected to fall in blank zones of a landscape colored by the completely unrelated purines, hypoxantines, pyrrolopyrimidnes, etc. Maps 1, 5, and 6 fully comply with this

point of view. Maps 2 and 4 demonstrate slightly "fuzzier" definitions of molecular similarity: a few phenyl sulfonamides are now being placed within the CHEMBL reference compounds, whereas map 7 based on scaffold-hop-supporting pharmacophore triplet counts actually assumes that most of the phenyl sulfonamides reside in the purine nucleoside phosphorylase-relevant chemical space zone. An overwhelming majority of these in-zone residing decoys are correctly recognized as inactives; however, even a "minority" of false positives may represent a very large number compared to the much rarer actives in the highly imbalanced DUD set. This is the reason for the predictive failure of map 7, which could not be understood in terms of its cross-validation results. When cross-validating, the map is exclusively confronted with purine nucleoside phosphorylase-relevant chemicals, where there are no "exotic" chemotypes to be spuriously co-opted into relevant chemical space by a—for this predictive challenge—"too permissive" perception of molecular similarity.

■ CONCLUSION

A new series of "universal" chemical space maps from data sets in the ChEMBL23 database was built using the GTM dimensionality reduction algorithm and following a previously reported evolutionary procedure to select preferred descriptor spaces and GTM parameter strings. These maps were able to provide better than random separation ($BA_{CV} > 0.6$) of actives and inactives in 609 of 618 ChEMBL sets, irrespective of whether series were used for map selection or not. However, consistently accurate predictions for each activity class could not be achieved by any individual map. However, these maps, which were each based on a different descriptor space, were highly complementary. For 617 of 618 activity classes, at least one out of the seven top universal maps represented a highly discriminatory activity landscape.

Because there is no correlation between performance in CV and external predictive power of individual activity landscapes, the one possible solution is to use a consensus approach. Thus, all landscapes with favorable density distributions of VS candidates make positive contributions to the consensus model. The most important advantages of a consensus map are (1) 100% data coverage in most of the cases; (2) significant increase in EF for the 100 top-ranked compounds; and (3) high performance of the consensus model compared to
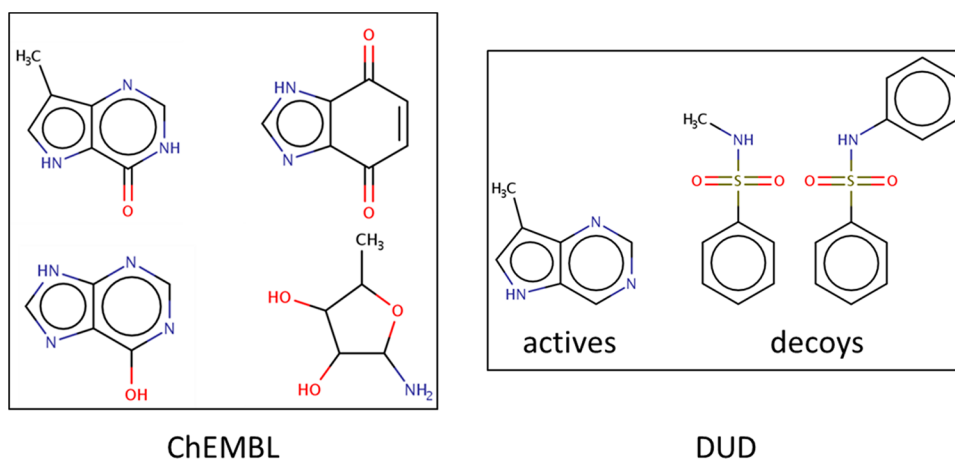

Figure 7. Representative substructures of compound subsets of the purine nucleoside phosphorylase receptor in the CHEMBL4338 data set and DUD.

individual models on the basis of ROC AUC. Thus, while any single universal map displays moderate predictive power, the combination of complementary maps results in a strong consensus effect in VS. Seven universal maps were sufficient to generate complementary views of biologically relevant chemical space that resulted in further increased VS performance.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00650.

> Activity landscapes for all nine DUD subsets used in VS (PDF)

> Archive of files "ChEMBL-target-ID@source.smi_id_-class" containing SMILES, compound CHEMBL ID or DUD ID (if applicable), and activity class label (1-inactive, 2-active) of the nine target-specific series from the two sources (ChEMBL, DUD respectively) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: varnek@unistra.fr.

ORCID Ⓞ
Dragos Horvath: 0000-0003-0173-5714
Jürgen Bajorath: 0000-0002-0557-5714
Alexandre Varnek: 0000-0003-1886-925X

Notes
The authors declare no competing financial interest.
ISIDA GTM software is developed by the Laboratoire de Chemoinformatique Strasbourg and can be obtained upon request (visit http://infochim.u-strasbg.fr/spip.php?rubrique41).

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(2) Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464−1480.

(3) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010−1024.

(4) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.

(5) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301−312.

(6) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34*, 348−356.

(7) Kayastha, S.; Kunimoto, R.; Horvath, D.; Varnek, A.; Bajorath, J. From Bird's Eye Views to Molecular Communities: Two-Layered Visualization of Structure–Activity Relationships in Large Compound Data Sets. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 961−977.

(8) Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure–Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* **2016**, *56*, 1438−1454.

(9) Sidorov, P.; Davioud-Charvet, E.; Marcou, G.; Horvath, D.; Varnek, A. AntiMalarial Mode of Action (AMMA) Database: Data Selection, Verification and Chemical Space Analysis. *Mol. Inf.* **2018**, *37*, 1800021.

(10) Kayastha, S.; Horvath, D.; Gilberg, E.; Gütschow, M.; Bajorath, J.; Varnek, A. Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. *J. Chem. Inf. Model.* **2017**, *57*, 1218−1232.

(11) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13*, 540−554.

(12) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of Drug-like Space: Towards a Polypharmacologically Competent Map of Drug-Relevant Compounds. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087−1108.

(13) Glavatskikh, M.; Madzhidov, T.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Malakhova, D.; Marcou, G.; Varnek, A. Predictive Models for Kinetic Parameters of Cycloaddition Reactions. *Mol. Inf.* **2018**, DOI: 10.1002/minf.201800077.

(14) Sidorov, P.; Viira, B.; Davioud-Charvet, E.; Maran, U.; Marcou, G.; Horvath, D.; Varnek, A. QSAR Modeling and Chemical Space Analysis of Antimalarial Compounds. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 441−451.

(15) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **2013**, *53*, 3318−3325.

(16) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(17) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(18) *ChemAxon, Standardizer, C,* version 5.12; ChemAxon, Ltd: Budapest, Hungary, 2012.

(19) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

(20) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA-Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191.

(21) Varnek, A.; Fourches, D.; Solov'Ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful in Silico Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **2007**, *25*, 433−462.

(22) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

Summary

Eight universal maps of biologically relevant chemical space, defined by the ChEMBL database, have been "evolved" by a GA with map parameter space being key degrees of freedom (including descriptor choice, grid size, manifold flexibility controls, etc,). An average predictive performance over hundreds of biological activities was used as an objective function. Each of the newly constructed uGTM is based on a different set of ISIDA descriptors

These GTMs were proven to successfully serve as hosts for 618 activity landscapes associated with the respective target-specific structure-activity ChEMBL compound series. The average predictive performance of those maps is roughly equivalent. Nevertheless, they significantly differ in the quality of each property-specific landscape.

It appeared that there is no correlation between performance in cross-validation and external predictive power of individual activity landscapes. Thus the one possible solution is to use a consensus approach. The most important advantages of a consensus map are:

- extended AD - the chance that a compound fall into empty or too sparsely populated areas in all the maps is close to zero, so at least some of the uGTMs will be able to return meaningful predictions;

- significant increase in enrichment factor for the 100 top-ranked compounds;

- high performance of the consensus model compared to individual models based on ROC AUC.

The minimum necessary number of uGTMs needed to provide satisfactory predictions for more than 600 biological activities has also been also investigated. It was shown that the 8th uMap is, in fact, redundant, and 4 326 (618*7) activity landscapes on seven first uGTMs are sufficient to enable polypharmacological profiling with reasonable accuracy. Later on, with the release of the v24 of ChEMBL, these activity landscapes were updated with newly added compounds. Almost a thousand new activity landscapes were created for additional 131 biological targets, bringing it up to 749 biological activities in total. Resulting 5K landscapes on seven uGTMs became the basis of the consensus GTM Profiler - a VS tool freely accessible at the Laboratory of Chemoinformatics of Strasbourg website (http://infochim.u-strasbg.fr/webserv/VSEngine.html, under ''QSAR-based Property Predictions'').[107] The consensus GTM Profiler have also became the main predictive instrument of ChemSpace atlas.

Considering that ChEMBL is the most extensive database of biologically tested compounds with dose-response activity values, the distinct universal maps constructed using this library represent complementary and strongly synergistic views of biologically relevant chemical space. They can be used not only as a predictive tool but also as frameworks for the analysis of large chemical libraries in the medicinal chemistry and drug design context. The first universal map was further used to analyze chemical space defined by biologically tested compounds from ChEMBL, commercially available molecules for HTS from ZINC, and DNA-encoded libraries enumerated using purchasable BBs. Thus in the ChemSpace Atlas those respective sections are based on the first uMap. However, due to the fact that there is a limited number of NPs in ChEMBL, a specific NP-uMap was constructed using compounds from the COCONUT collection of NPs. Similarly, a dedicated universal map of synthons was created (without considering the leaving groups in actual reagents). This map was trained on synthons generated both from commercially available reagents and ChEMBL compounds (via their fragmentation).

4 Exploration and Analysis of Ultra-Large Chemical Spaces

The chemical space is vast, but medicinal chemists do not deal with all compounds at once. Depending on the stage and strategy of drug design, the focus moves from one type of compounds to another. Therefore in this thesis, we look at the different parts of the chemical space separately. At first, compounds used in the conventional screening approaches are analyzed – drug-like and lead-like molecules, used in HTS; fragment-like libraries for the fragment-based drug design; PPI-like compounds for the search of protein-protein interaction inhibitors (Chapter 4.2). These approaches are pretty expensive techniques that are out in the field for a few decades already. A new promising technique – DNA-encoded libraries screening – introduces multiple advantages and makes screening available not only to Big Pharma but also for the academic laboratories. The chemistry used for the DEL synthesis is limited, making DEL chemical space somewhat different from the conventional screening libraries. Therefore its analysis is separated and reported in Chapter 4.3.

All of the abovementioned segments of the chemical space are mostly populated by compounds synthesized by means of organic chemistry. Thus, the availability of the reagents used for their synthesis deserves a separate discussion (Chapter 4.4). However, analysis of the BBs poses particular challenges, yet unsolved by the chemoinformatics community (at least in the form of openly available software). Therefore a new tool was developed for the BB analysis, treatment, and library design and described in Chapter 4.4.1

On the other side, natural products (NPs) were the source of medicines for hundreds of years, and they still serve at least as inspiration for drug discovery. The chemical space of NPs and NP-like compounds is analyzed in Chapter 4.5.

4.1 Evolution of commercially available compounds for HTS

Introduction

Numerous chemical suppliers provide a diverse choice of compounds that became the primary source of the potential hits at the early stages of drug discovery. The quality of such compound collections, i.e. their correspondence to the main beliefs of what the optimal screening library should look like, is crucial for drug discovery success and thus significantly influences the client choice for acquisition. Therefore various medicinal chemistry trends, popping up gradually during the last decades, reshaped significantly commercially available chemical space. The most powerful and game-changing among such trends is high-throughput screening (HTS)[116] as a preferred choice for the enlargement of biomedical knowledge. Almost all of the currently existing chemical suppliers propose at least one screening library for their customers. However, the question is how are they differ, and do they correspond to the current needs of medicinal chemists.

In this work, screening libraries of the leading suppliers were analyzed in terms of physicochemical properties, novelty, diversity, and quality as a source of potential hits. The distinctive feature of this work is an overview of the principal changes that commercial chemical space has overcome over the last years and how it evolved to meet the main criteria possed by medicinal chemists. Besides, the possibility to compile an "ideal" diverse dataset for primary screening against a novel target with compounds purchased from different suppliers was investigated for the first time.

Main terminology

High-throughput screening (HTS) – an experimental methodology that uses automated equipment to rapidly test thousands to millions of samples for biological activity at the model organism, cellular, pathway, or molecular level.

Pan-assay interference compounds (PAINS) are chemical compounds that often give false-positive results in high-throughput screens due to the non-specific interactions with numerous biological targets.

Lilly MedChem filters - set of 275 rules, developed over 18 years, used to identify compounds that may interfere with biological assays: reactivity, interference with assay measurements, activities that damage proteins, instability, and lack of drugability.

Reviews • FOUNDATION REVIEW

Teaser An assessment of 16 million commercially available compounds, (properties and quality), comparing vendors offerings and how they have evolved to meet modern physiochemical requirements. A selection of 500,000 lead-like compounds for high throughput screening.

Evolution of commercially available compounds for HTS

Dmitriy M. Volochnyuk[1], Sergey V. Ryabukhin[2], Yurii S. Moroz[3,7], Olena Savych[4], Alexander Chuprina[3], Dragos Horvath[5], Yuliana Zabolotna[5], Alexandre Varnek[5] and Duncan B. Judd[6]

[1] Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine
[2] The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine
[3] ChemBioCenter, Kyiv National Taras Shevchenko University, 61 Chervonotkatska Street, Kyiv 02094, Ukraine
[4] Institute of Bioorganic Chemistry and Petrochemistry, National Academy of Sciences of Ukraine, Kyiv 02094, Ukraine
[5] Laboratoire de Chemoinformatique, 4, rue B. Pascal, Strasbourg 67081, France
[6] Awridian Ltd, Gunnelswood Road, Stevenage SG1 2FX, UK
[7] Chemspace, ilukstes iela 38-5, Riga, LV-1082, Latvia[8]

Over recent years, an industry of compound suppliers has grown to provide drug discovery with screening compounds: it is estimated that there are over 16 million compounds available from these sources. Here, we review the chemical space covered by suppliers' compound libraries (SCL) in terms of compound physicochemical properties, novelty, diversity, and quality. We examine the feasibility of compiling high-quality vendor-based libraries avoiding complicated, expensive compound management activity, and compare the resulting libraries to the ChEMBL data set. We also consider how vendors have responded to the evolving requirements for drug discovery.

Introduction

A growing body of evidence from clinical outcomes, along with scientific and technological advances over the past decades, has resulted in shaping the strategies of early-stage drug discovery [1]. High-throughput screening (HTS) has evolved since its introduction during the early 1990s. Initially, many pharmaceutical companies were screening hundreds of thousands of compounds against hundreds of targets per year. Today, HTS is often complemented with fragment-based lead discovery (FBLD) [2], encoded library technologies [3], and phenotypic approaches [4] to form a comprehensive screening toolbox and an opportunity to combine knowledge from each

Dmitriy Volochnyuk shares his time as head of the Biologically Active Compounds Department at the Institute of Organic Chemistry of the NAS of Ukraine and as a professor in the Institute of High Technology, Kiev National University. He received his PhD in Organic Chemistry in 2005 and his DSc in organic and organomettalic chemistry in 2011. He has 10+ years' experience in managing chemical outsourcing projects having previously worked in contract research organizations. Dr Volochnyuk is an expert in fluoroorganic, organophosphorus, heterocyclic, combinatorial,and medicinal chemistry. He is also an author on over 120 scientific papers.

Sergey Ryabukhin is an associate professor in the Institute of High Technology, Kiev National Taras Shevchenko University. He was awarded his PhD by Kiev National University in 2008. He has 10+ years' experience in managing combinatorial chemistry departments as well as chemical outsourcing projects having previously worked in contract research organizations. Dr Ryabukhin is an expert in combinatorial methods in organic chemistry, organosilicon, and organoboron chemistry. He is an author on over 50 scientific papers.
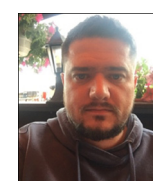
Duncan B. Judd is consultant at Awridian Ltd, currently working with a range of organizations including international companies. He is an accomplished medicinal chemist with extensive outsourcing experience and a 39-year proven track record with a blue-chip pharmaceutical company. Duncan has made significant contributions to numerous drug discovery projects, and is cited on many patents and publications. He has extensive outsourcing experience and has published and presented on open innovation in drug discovery, for which he is a strong advocate.

Corresponding author: Judd, D.B. (duncan.b.judd@awridian.co.uk)
[8] chem-space.com.

approach to successfully identify new lead molecules. Despite these industry-changing 'paradigm shifts', the number of new drugs approved per US$1 billion spent on research and development (R&D) has been halving every 9 years since 1950 [5], and now an estimate of R&D spending per new product exceeds US$2 billion [6].

There has been much speculation in the literature and in the industry around the quality of HTS data derived from random screening, both in terms of sample purity and the physicochemical properties of HTS screening decks. Many consider the classical approaches used by James Whyte Black during the 1960s–1970s [5,7] as being a preferred alternative. However, further studies have clearly shown that HTS is a valuable part of a proven scientific toolkit, and the wide use of the method is essential for the discovery of new chemotypes [8]. Furthermore, the modern HTS is on the 'Plateau of Productivity' phase in the Gartner Hype Cycle, and is now integral in lead discovery along with a combination of different approaches.* Moreover, the content, size, and quality of a compound collection used in HTS campaigns are all fundamental to the success of a project: the most advanced screening technologies and the most physiologically relevant assays were thought to be compromised by the low quality of compound collections [9].

At a time when the HTS technology had achieved its 'Peak of Inflated Expectations' and ultra HTS (uHTS) had evolved, it became apparent that large numbers of screening compounds were required. In response, big pharmaceutical companies ('Big Pharma') started enhancing their compound collections, launching file enrichment programs during the early 2000s. However, many of the early combinatorial libraries are now considered far from the optimal chemical space appropriate to initiate a successful drug discovery project [10]. This activity, as well as mergers and acquisitions (M&A), have led to an increase in the size of their respective corporate libraries some to several million compounds: (Pfizer, 4 million [11]; BHC, 2.7 million [12]; AZ, 1.7 million, own collection* and 4 million, accessed through collaborations [13]; Novartis, 1.7 million [14]; GSK, 2 million (1.8 million diversity set) [15,16]; Sanofi in collaboration with Evotec, 1.7 million [17]; and Roche, 1.2 million [18]). Moreover, AstraZeneca (AZ) and Bayer have made their collections available to one another for specific HTS campaigns. The overlap for the combined AZ-Bayer set is minimal (~3.5% of the combined library size) and that is attributed to compounds being purchased from chemical vendors [19].

During this period, several companies emerged to meet the demand for more compounds. Furthermore, advances in cheminformatics tools have enabled the design of development libraries, such as the elimination of compounds with inappropriate parameters. Starting from the Lipinski Rule of 5 (Ro5) coined in 1997 [20], many related drug-like criteria have been proposed [21]. In 1999, Teague *et al.* [22] observed that, during optimization, the

molecular weight (MW) of the lead molecule increased by 200 Da, whereas logP increased by 0.5–4, which yielded another key concept of lead-likeness. The latter was further developed in 2008 by Pfizer's researchers revealing the Rule of 3/75 (Ro3.75) [23], and the current list of filters is more stringent than the original drug-likeness philosophy. Finally, the Rule of 3 (Ro3) proposed by Congreve *et al.* in 2003 [24] has found a wide application in FBLD.

The aforementioned physicochemical guidelines in combination with the structural filters (reactive compounds [25], REOS [26], PAINS [27], Eli Lilly Rules [28] etc.) and diversity selection methodologies [29,30] have resulted in improvement in the quality of subsequent hits. In addition, the concept of lead-oriented synthesis introduced by Churcher *et al.* in 2012 [31] focused on appropriate chemical space. Despite criticism [32], the current trends in compound set design include filtering of databases before a screening campaign based on chemical structure, calculated properties, rule-based criteria, or the binding efficiency predictions. These filters are routinely combined to form an efficient triage [33] that effectively shrinks chemical space created during the 1990s and early 21st century to make it more appropriate for high-quality HTS. These filtering approaches combined with the synthetic methods have allowed the creation of large drug-like, lead-like, and fragment-like compound collections, which have been aligned with the current paradigm within the industry. Furthermore, it is Big Pharma, with their substantial financial and infrastructure resources, that have developed their collections, which have become 'family jewels' and, therefore, until recently had been inaccessible to those outside the companies, such as academic users and small biotechs.

Despite these challenges, there have been several initiatives to explore HTS outside the pharmaceutical industry [34]. In 2004, the US National Institutes of Health (NIH) and the European Union Innovative Medicines Initiative (EU IMI) both initiated projects to enhance their respective compound collections with the aim of making high-quality compound libraries accessible to the wider scientific community [35]). In the main, these initiatives relied on buying appropriate compounds from chemical vendors. In some cases, pharmaceutical companies have broken new ground by opening their technologies and resources in HTS to selected academics and external institutions [36].

Many outside of Big Pharma have the capabilities to select and order compounds, but the logistics of compound handling tend to get overlooked, such as in the consolidation of libraries from different vendors. Automated production of assay-ready compound plates for screening requires specialized formatting facilities, which could cost US$7 million [37], thus being unaffordable for smaller organizations. There are two approaches to overcome the above-mentioned issues: (i) ordering from companies that specialize in consolidating and formatting libraries; or (ii) purchasing a preformatted library ready to use from a limited number of vendors. To the best of our knowledge, there is only one study evaluating SCL from the user's standpoint, published in 2013 [38]. The main conclusion of that study was that the available screening compounds appeared small and was, at that time, represented by fewer than 350 000 compounds [38].

Despite several analyses of the chemical space covered by SCL published in 2004 [39], 2005 [40], 2006 [41], and 2015 [42]

---

* Mayr, L.M. and Wigglesworth, M. High-Throughput Screening: Challenges & Opportunities 8th ELRIG Drug Discovery Conference Manchester/UK 2014, September 2–3.http://elrig.org/downloads/dd14/20140904_Mayr_ELRIG2014.pdf.

* Mayr, L.M. and Wigglesworth, M. High-Throughput Screening: Challenges & Opportunities 8th ELRIG Drug Discovery Conference Manchester/UK 2014, September 2–3. http://elrig.org/downloads/dd14/20140904_Mayr_ELRIG2014.pdf.

(including our studies in 2011 [43], 2012 [44]) the question remains as to whether the available purchasable chemical space could enable the creation of a high-quality compound library for HTS projects that are comparable to Big Pharma's proprietary repositories. Thus, the goals of present study were: (i) to provide a critical view from a user's standpoint on the existing SCL offerings and to clarify whether they are comparable to Big Pharma's collections in terms of 'compound novelty, diversity, and quality'; (ii) to examine the feasibility of facile compiling a high-quality compound library via a limited number of vendors, hence avoiding complicated and expensive compound management; and (iii) to include in the analysis a comparison of vendor's offerings.

A preferable supplier can be identified using the following criteria: (i) cost effective and timely delivery of quality compounds; (ii) a wide range of compounds with appropriate physical and chemical properties [Ro5, Ro3, with limited undesirable functionality: no 'PAINS', stable, no hot functionality (except covalent libraries)]; (iii) possibility of provision of analogs for hit follow-up in a time- and cost-effective manner (except for NP and metabolites); (iv) the SCL represents numerous and/or original chemotypes, as defined by Bemis-Murcko, Tanimoto, and so on; and (v) the vendor updates the catalog regularly, and is clear about pricing with transparent and prompt communication throughout the purchasing process.

However, a comprehensive analysis of the vendors fulfilling the above-mentioned criteria limited to the information extractable from open sources because most companies prefer not to share their analysis of various vendors. Therefore, we used cheminformatic approaches to compare the SCLs found in open platforms. As an indirect indicator of the vendor's activity in the field, we analyzed the dynamics of the reshaping and growth of their collections over a set time period.

Results and discussion
Collection of the data and characteristics of the data sets
The starting point of the current study was the creation of the chemical space covered by purchasable screening compounds using the ZINC database.[†] To create this space, we performed standardization of SMILES for all the sets involved in our search using RDKit nodes for the KNIME analytics platform.[‡] This space was defined as the union of standardized SMILES strings of all sets prepared, as mentioned earlier. Duplicates were deleted from the newly created large set. After removal of duplicates, the standardized space comprised 16 902 208 unique structures, including stereoisomers (all stereochemical features mentioned by vendors were included). As illustrated by Fig. 1 and Fig. S1 in the Supplementary information online, the impact of the vendors on the space differed significantly by the number of structures as well as by percentage of unique compounds. From 33 sets, eight showed a high fraction of unique compounds (80% and more): Abamachem, AnalytiCon Discovery, BCH Research, Enamine, FCH Group, Intermed, Selenachem, and UORSY; all these sets, except for AnalytiCon Discovery, contained more than 1 million molecules. Eight sets contained a medium number of unique compounds (40–80%), and three of these sets were of 1 million or more

molecules (Asischem, ChemBridge, and ChemDiv). Even though Princeton Biomolecular Research and Vitas-M contained 1.2 million and 1.4 million molecules, respectively, the fraction of unique compounds was <10% for both databases.

Compound-level analysis (for the 16 902 208 set)
For the preliminary evaluation of the quality of the purchasable chemical space as well as the set from each vendor, ten selected molecular properties were chosen: MW, logP, heavy atom (HA) count, number of hydrogen bond donors (HBDs), number of hydrogen bond acceptors (HBAs), polar surface area (PSA), number of rotatable bonds (ROTB), Fsp^3, number of rings, and number of aromatic rings. The mean values of these parameters are detailed in Table 1. We also compared these values with the corresponding data from our previous analysis from 2011 [44]. The data showed that, during the past 7 years, the mean values of the six parameters mentioned in our previous paper significantly shifted from drug-likeness to lead-likeness, which accords with general trends of the screening libraries criteria. The mean MW ($\Delta = -26$), logP ($\Delta = -0.67$), PSA ($\Delta = -22.4$), HBA ($\Delta = -1.57$), and ROTB ($\Delta = -0.47$) significantly decreased whereas mean HBD slightly increased ($\Delta = +0.20$). Given the impact of historical compounds from the collections of the main players in the field, which strongly affected the mean values, we compared the mean value of the compounds appearing from 2010 to 2017[§]; encouragingly, these results were the closest to the lead-oriented synthesis concept.[¶] Comparison of the characteristics of the 'new compounds' set from the SCL 2010–2017 with the European Lead Factory[**] (ELF) library [45] (mean values, calculated on the basis of the data from two publications [46,47] showed that parameters of the SCL 2010–2017[‡‡] set were stricter [mean MW (SCL 2011–2017) = 340, MW (ELF) = 425; logP (SCL 2011–2017) = 2.38, logP (ELF) = 3.1] and closer to DrugBank mean values (MW = 315, logP = 2.4) than were those of ELF (Table 1).

In addition to the mean values, we analyzed the distribution of the aforementioned parameters for all purchasable chemical space as well as for each vendor collection (for exact information on vendors, see mmc3.xlsx in the Supplementary information online). To simplify the visualization of the distributions of each vendor compared with the space, we divided the distributions into several areas. The distributions that were difficult to assign to the areas are marked in the figures as 'outliers'. The representative examples of such simplifications are shown in Fig. S2 in the Supplementary information online.

For example, in reviewing the results for MW, we believe there are three general categories of suppliers: Area 1: ten distribution curves (Abama Chemicals, BCH Research, Intermed Chemicals,
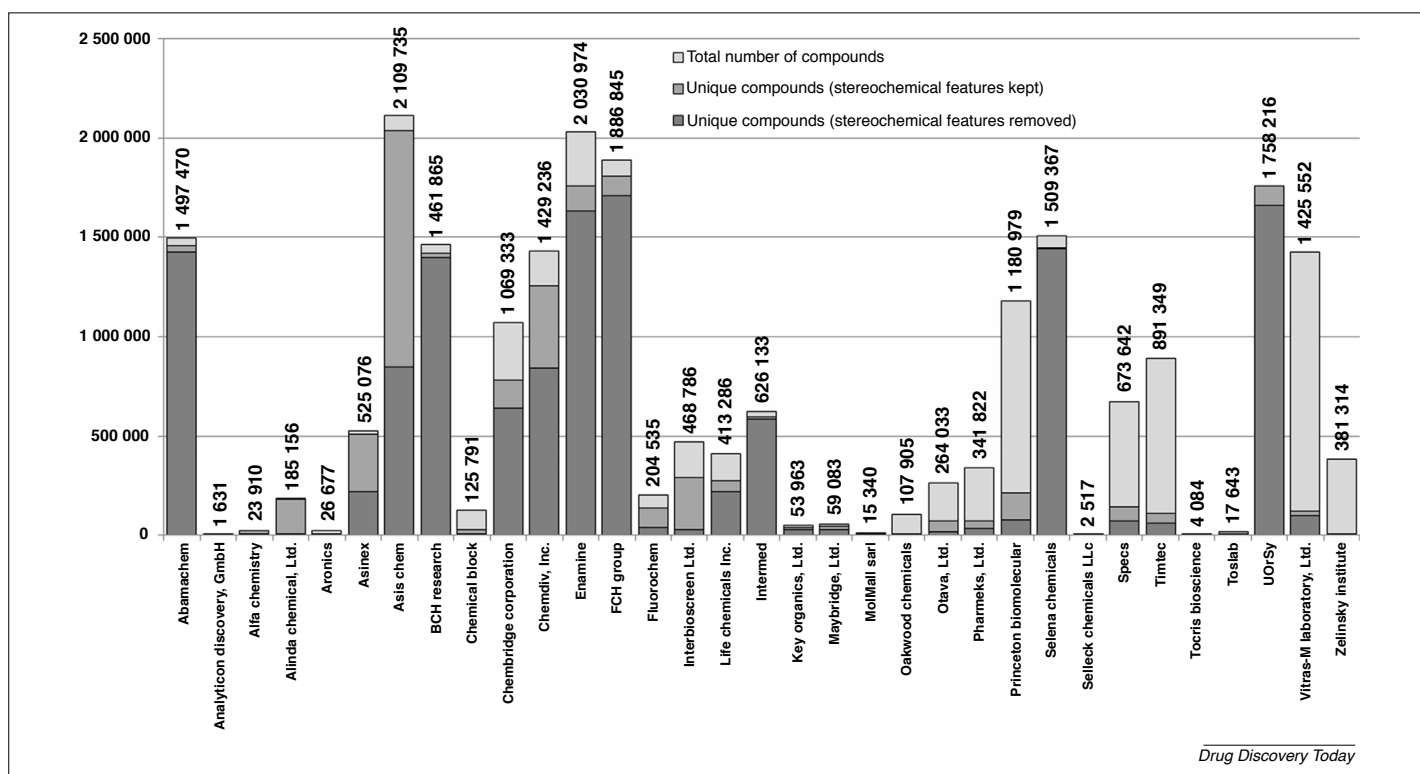

FIGURE 1

The chemical space of purchasable screening compounds represented by vendors.

TABLE 1

Mean values of selected molecular properties of the purchasable chemical space in 2010, 2017, and the ELF library

Parameter (X)	2010	2017	$\Delta\langle X\rangle$ (2010–2017)	$\langle X\rangle\Delta$ (2010–2017)	ELF
MW	388.82	362.49	−26.33	339.59	425
logP	3.64	2.96	−0.67	2.38	3.1
Fsp^3	–	0.40	–	–	0.4
tPSA	94.23	71.84	−22.39	52.38	91
Heavy atoms	–	25.11	–	–	–
HBA	6.18	4.61	−1.57	3.25	–
HBD	0.96	1.16	0.20	1.33	–
ROTB	5.28	4.82	−0.47	4.41	–
Rings	–	3.02	–	–	–
Aromatic rings	–	2.03	–	–	–

Selena Chemicals, ChemBridge, Enamine, FCH Group, Key Organics, Maybridge, and UORSY) have narrow peaks with maxima between 300 and 400 Da; Area 2: 18 distribution curves (Alinda Chemicals, Asinex, ChemDiv, Aronis, Asischem, Chemical Block, InterBioScreen, Life Chemicals, Otava Chemicals, Pharmeks, Princeton Biomolecular Research, Selleck Chemicals, Specs, Timtec, Tocris, Toslab, Vitas-M Laboratory, and Zelinsky Institute) have wide peaks with a vertex at 400 Da. By contrast, five curves (AnalytiCon Discovery, Alfa Chemistry, Fluorochem, MolMall, and Oakwood Chemicals) were left as is and recognized as 'outliers'. Another representative example of simplification is the distribution of HBD number given in Fig. S2 in the Supplementary information online. Using such an approach, distribu-

tions of all above-mentioned parameters were calculated and are shown in Fig. 2.

Among the compound suppliers, AnalytiCon Discovery, Alfa Chemistry, Fluorochem, MolMall, and Oakwood Chemicals were identified as 'frequent outliers'. The main reason for this rests on the main business activity of these companies. AnalytiCon Discovery specializes on natural products and macrocycles; Fluorochem and Oakwood Chemicals are widely known as suppliers of building blocks and reagents; Alfa Chemistry is a contract research organization; and MolMall is a small collection of samples from different sources. All these companies are not 'classical' producers of the compounds for HTS. However, despite differences in the parameter distributions of each vendor, the cumulative distribu-
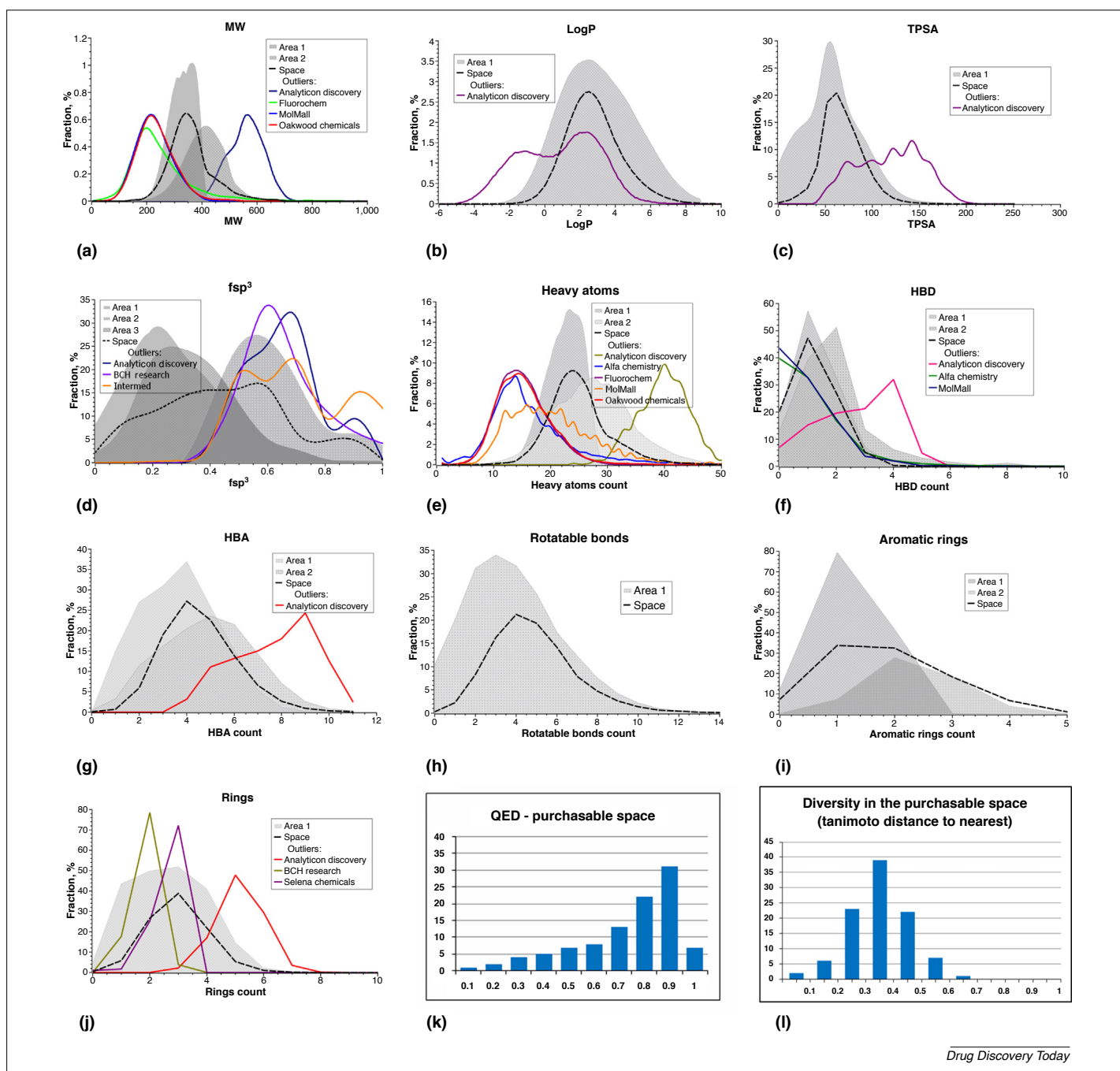
FIGURE 2

Distribution of the selected molecular properties of the purchasable chemical space with 'vendor areas' and outliers together with QED and ECFP4-based Tanimoto similarity profiles for the space. Please see main text for definitions of abbreviations.

tions of the parameters of purchasable space have one peak, which is usual for screening collection. An exception is the Fsp³ distribution, which has a more complex character, unlike the curves of vendors. In this case, old historical collections and the newly synthesized compounds have significantly different Fsp³ parameter values (Fig. S3.01 in the Supplementary information online). Nevertheless, the quantitative estimate of drug-likeness (QED) [48] histogram for the purchasable space revealed the quality of the compounds based on this parameter (see mmc4.xlsx in the Supplementary information online). The maximum QED accounted for 0.8–0.9 (Fig. 2).

The chemical diversity of the space and vendor collections was analyzed by ECFP4-based Tanimoto similarity of each compound with its nearest neighbor (for all vendors, see Figs. S3.01–3.10 in the Supplementary information online). For the purchasable space, the corresponding histogram is shown in Fig. 2. Its profile demonstrates a diverse set with a mean Tanimoto distance to nearest neighbor of 0.3. Notably, Tanimoto diversity for the purchasable space is worse than the data announced for the Joint European Compound Library (JECL): a mean Tanimoto distance of 0.4 to the nearest neighbor [47]. Deeper analysis of the contribution of each supplier to a joint diversity of the space showed that

some sets represent completely different areas of chemical space, whereas others have a significant overlap. As an example, the AnalytiCon set has a low internal diversity but occupies a significantly different space from other vendors (median Tanimoto distance 0.18 within the set, but 0.55 against the full space).By contrast, the Vitas-M set is narrowly distributed (median Tanimoto distance 0.24 in set, and median Tanimoto distance in comparison with the full space 0.29). Selleck set had high internal diversity and differed from other vendors (median Tanimoto distance was 0.56 in the set but median Tanimoto distance in comparison with full space was 0.46). The corresponding histograms are shown in Figs. S4.01–4.33 in the Supplementary information online.

For the 3D-shape analysis of the purchasable space as well as vendor sets, the Plane of Best Fit (PBF) – Principal Moments of Inertia (PMI) approach was used [49]. Generation of coordinates and geometry optimization (mmff94, 100 iterations per molecule) along with subsequent PMI and PBF calculations, were performed using RDKit. Density plots were built in R Statistics using the hexbin package; the plot for the complete space is shown in Fig. 3a.

According to the PBF = 0.6 and NPRsum = 1.1 cut-off filter, the number of 'out-of-plane molecules' in purchasable space was 8 668 016 (51%). The same calculations for each vendor set (Figs. S5.01–5.34 in the Supplementary information online) revealed that the fraction of compounds passed through the filter fell in a range of 36–47%, with exception for AnalytiCon (76%), Alfa Chemistry (20%), Alinda (33%), Aronis (26%), Fluorochem (20%), and Oakwood (21%).

Scaffold level analysis

Bemis–Murcko loose frameworks (scaffolds) analysis [50] was used to evaluate the 2D shape and topology of the compounds in the purchasable space and each vendor collection (Figs. S6.01–6.33 in the supplementary information online). This analysis gave 2 886 942 unique frameworks representing purchasable space. Cumulative scaffold frequency plots (CSFP) [51] were built for the space and vendor collections. As in the case of compound-level analysis, the main 'area' and outliers were identified. This time, UORSY appeared in outliers, the CSFP of which was close to those of Binding DB and DrugBank (Fig. 3b).

Equal distributions of compounds across molecular scaffolds were found in the Selleck and Tocris collections, mainly because of the main profiles of these companies: Selleck and Tocris are worldwide recognized suppliers of reference compounds, which are usually used as standards in different screening assays as well as in biomedical investigations. Our data are in slight disagreement with a recently published analysis of the libraries of the main players [52], but the CSFP curves obtained therein fit the 'area' in Fig. 3b.

SCL changes analysis

An important factor in the choice of compound vendor is the viability of the sample resupply and further opportunity for the hit follow-up support [38]. Another is how vendors have responded to the desire for more lead-like compounds. To address these issues, we focused on companies active in this field. Promotional materials of those companies do not give a true picture; therefore, we evaluated such companies by comparing the results of analyses carried out in 2010 and in the current paper. Initially, differences

in compound numbers in collections were plotted (Fig. 4). Some vendors presented in 2010 (AMRI, ComGenex, Tripos, ART-CHEM, Nanosyn, SALOR, IVK Laboratories, ChemStar, Ufark, and Spectrum) were absent in 2017 in ZINC. Some of these companies had been sold (e.g., ComGenex[§§] or Tripos[¶¶]), whereas others, such as AMRI and Nanosyn, provided integrated MedChem solutions using in-house libraries. Moreover, all these vendors were not active participants in screening compound production. In 2017, 14 new vendors were present: AnalytiCon, Selleck, Tocris, MolMall, Alfa Chemistry, Aronis, Chemical Block, Alinda, Zelinsky Institute, Intermed, BCH research, Abamachem, Selena Chemicals, and FCH Group. The libraries of the latter four contain more than 1 million unique diverse compounds with good PhysChem properties (see mmc2.xlsx in the Supplementary information online), proving their activity on screening compounds market.

The vendors referred to in the analysis of 2010 could be divided into several categories (i) outgoing from the market: TOSLab, Maybridge (−9070 cpds/33% and −10 779 cpds/15%) and Inter-BioScreen (almost no changes in 7 years); (ii) not growing: Key Organics, Asinex (+6307 cpds/13% and +67 234 cpds/15%, respectively: <15% increase of the library size without significant qualitative changes) and Life Chemicals (−12 849 cpds/3% decrease in size but with considerable qualitative changes, Δ<MW> (2010–2017) = −26; Δ<logP> (2010–2017) = −0.36); (iii) growing: Chem-Bridge (+328 157 cpds/44%); (iv) extremely growing: ChemDiv (+643 496 cpds/82%), Enamine (+809 017 cpds/66%), UORSY (+963 219 cpds/120%), and Asis Chem (+2 076 986 cpds/634%); and (v) companies that proposed building blocks in mg quantities: Oakwood and FluoroChem. The latest category appears to be growing, with seven vendors currently included: Otava Chemicals, Pharmeks, TimTec, Specs, Princeton Biomolecular Research, Vitas-M, and Zelinsky Institute. Despite the increased number of compounds, these collections include a few unique structures (Fig. S2 in the Supplementary information online). We carried out further analysis of cross-overlapping of these collections (Table 2) that revealed that the libraries of five vendors (Otava Chemicals, Tim-Tec, Princeton, Vitas-M, and Zelinsky Institute) substantially overlap, which is an indirect proof of common source of these compounds and questions the production ability of these compounds.

At a cursory glance, the space was sufficiently diverse and covered significant PhysChem parameters for most screening campaigns; thus, it could deliver an appropriate HTS set. To verify this statement, several case studies were performed.

Case study: an 'ideal' million

Among the variety of screening paradigms that exist to identify hits [53], we chose an example comprising building a compound set to screen against a novel target with an unknown structure, with few known active chemotypes, or without existing small-molecule modulators. In this case, HTS is the method of choice for its potential to identify quality leads because it does not require

---

§§ https://bbj.hu/business/
albany-molecular-closes-comgenex-acquisition_9580.

¶¶ www.thepharmaletter.com/article/tripos-to-sell-drug-discovery-business.

Space

(a)

Scaffold diversity
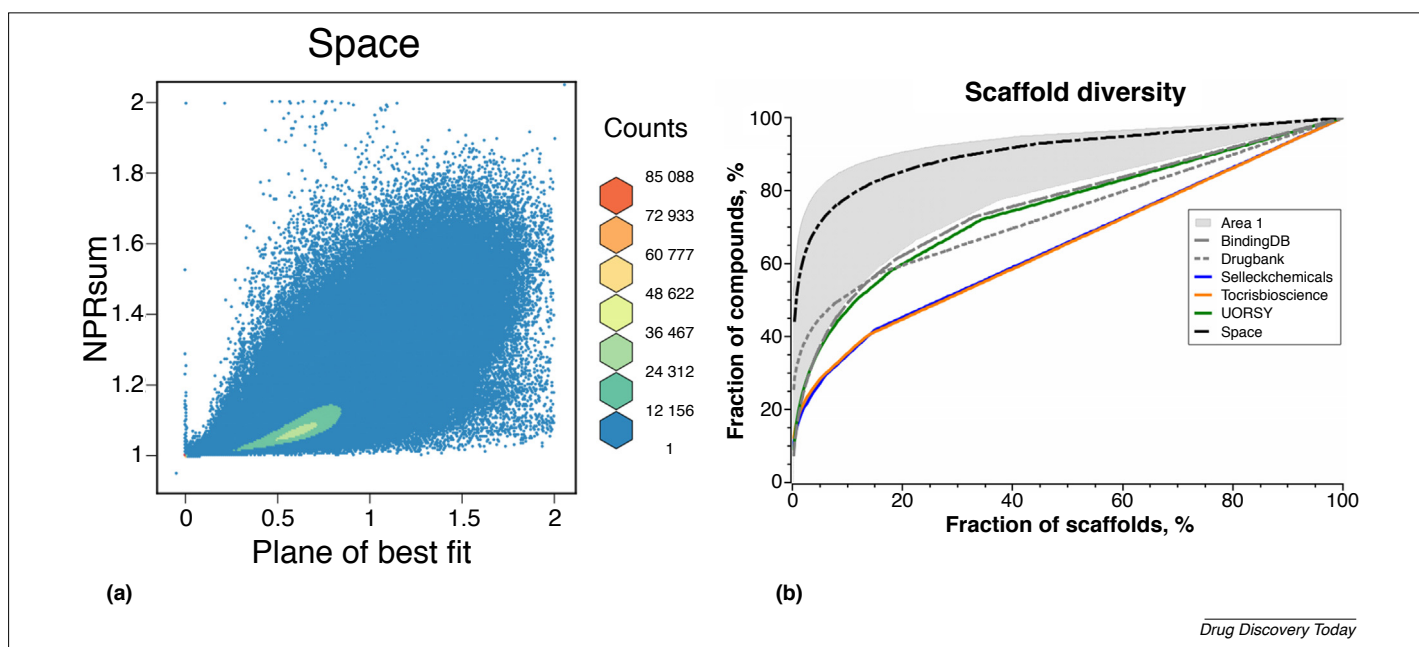
(b)

Drug Discovery Today

FIGURE 3

3D shape and scaffold diversity of the purchasable chemical space. **(a)** Density plot of Plane of Best Fit (PBF) score versus the sum of normalized principal moments of inertia (NPR). **(b)** Cumulative Scaffold Frequency Plots of the scaffold with 'vendor areas' and outliers compared with Binding DB and DrugBank.

information about the target. However, determining the optimal size of such a screening deck is problematic. Several studies have addressed this question but the optimal size of a screening collection [54,55] has remained undefined and varied.

The technical possibilities of modern HTS are almost unlimited. Nowadays, 384-well microtitre plates are the 'golden standard,' whereas 1536-well plates are increasing in popularity, and even 3456-well microtitre plates are used in some projects. Throughputs of $\geq$100 000 compounds screened per day are routine in leading HTS practitioner laboratories using *in vitro* biochemical, functional cell-based, reporter gene, and phenotypic assays [56]. According to reports on screening campaigns, the number of compounds used in an 'all-or-nothing' screening mode ranges from 50 000 to 1 500 000 [57]: a maximum mean value of 800 000 compounds per screen was reported in 2003, whereas this number had decreased to 500 000 in 2009 [58]. Despite a low true positive hit rate (<1% in 2010 [59]), in 2018, AZ concluded that increasing success could be achieved by gaining access to as many compounds as possible [13]. Moreover, choosing the 'relevant region' of the chemical space [28] would decrease further attrition and increase the true positive hit rate [60]. Support for the trend to use several million screening compound campaigns is the multiplexing of more than one compound per well during primary HTS to increase the capacity without compromising screening quality [61]. Thus, we assembled a screening deck of 1 million lead-like compounds, based on 50 000 scaffolds with 20 representatives each, belonging to clusters that were as diverse as possible for the first case study. We limited the number of the compounds to eliminate the molecular redundancy [62], but left a sufficient number of compounds per cluster to efficiently identify latent hit series and rapid preliminary structure–activity relationships (SARs), and to avoid any singletons [63]. Currently, there is controversy over the optimal size of compounds per cluster per scaffold. The first papers

discussing the issue were published in early 2000, although their conclusions varied from 10 [64] to 50–100 [65] compounds per scaffold. By contrast, the 'Open Scaffolds' collection from Compounds Australia was build with $\leq$30 SAR-meaningful compounds per scaffold (avarage value 28) [66]. Nevertheless, a series of 5–20 compounds was most frequently used by Pfizer [67] during plate-based diversity subset generation 2 (PBDS2). Therefore, we selected a model value of 20 compounds per scaffold, also in agreement with the opinion of Bostwick.*** For comparison, we also ran the study using 50 compounds per scaffold.

To build an 'ideal million' set, we initially subjected the purchasable chemical space of 16 902 208 compounds to structural filtering against PAINS (despite recent criticism [68], the filters are routinely used) and toxicology/reactive Eli Lilly Rules [27,28], which selected 15 968 338 compounds. Further application of the lead-likeness [69] and Ro3/75 [23] criteria resulted in two spaces with 6 544 044 and 3 705 803 compounds, respectively. Bemis–Murcko loose framework analysis of the sets gave only 39 101 and 22 162 scaffolds bearing more than 20 compounds per scaffold and 13 156 and 8006 scaffolds bearing more than 50 compounds per scaffold (Table 3). Given that the first model ideal million set (20 compounds per scaffold) would require 50 000 scaffolds and fewer than this were available from drug-like space, we targeted a 0.5 million set represented by 25 000 scaffolds with 20 compounds per scaffold and used the 6 544 044 set. From this set of 39 101 scaffolds, we extracted 25 000 of the most diverse using the MaxMin algorithm [70]. If the scaffolds had more than 20 compounds in the lead-like space, we selected the 20 most diverse structures using the above-mentioned MaxMin algorithm for compounds from overpopulated scaffolds [70]. In this 'ideal half million', the unique structures from all 33
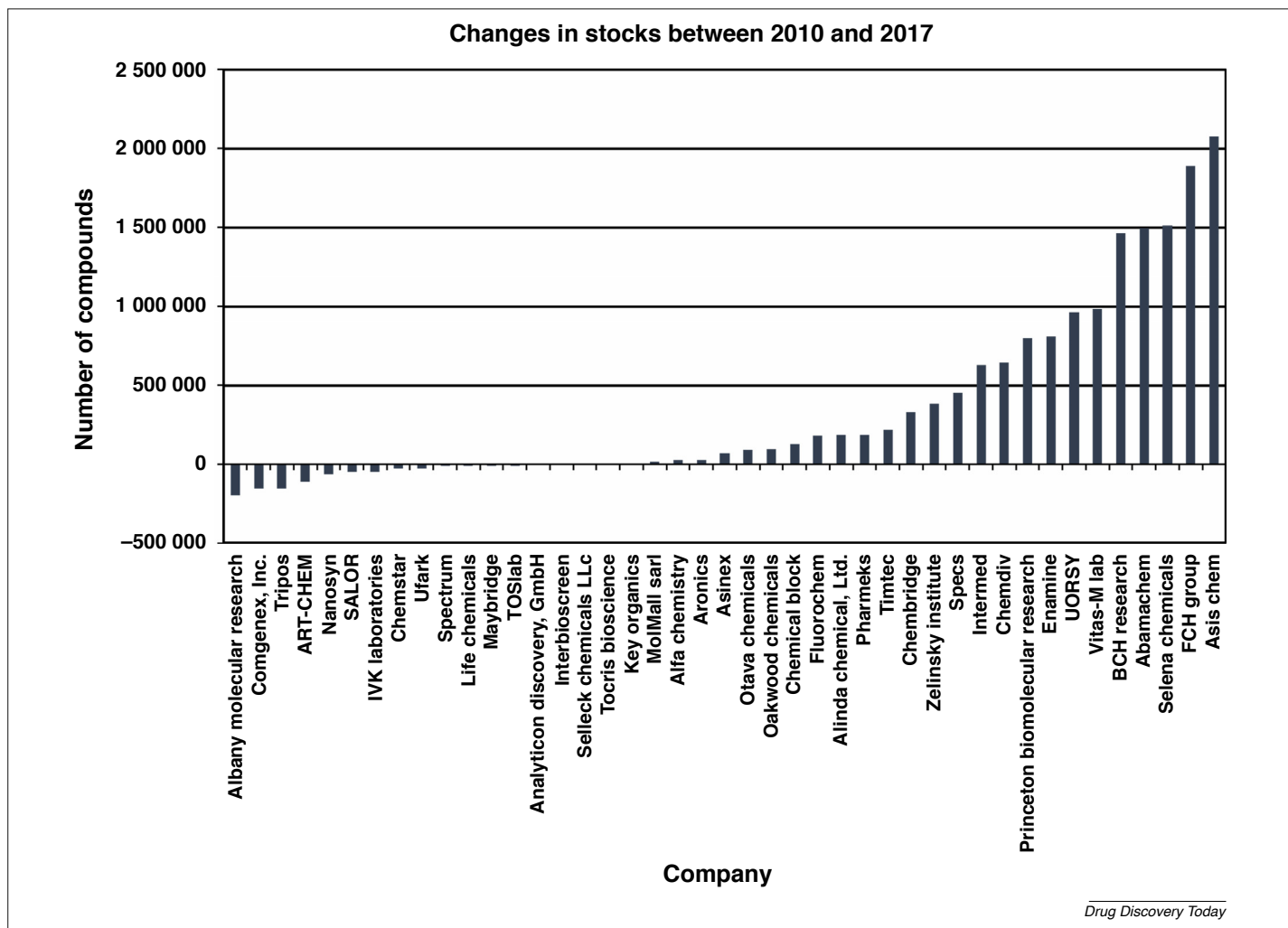

*** www.uab.edu/medicine/adda/images/BostwickHTS.pdf.

FIGURE 4

Changes in suppliers' compound libraries (SCL) size from 2010 to 2017.

TABLE 2

Cross-overlapping of the 'seemingly growing' vendors[a,b]

	Otava	Pharmek	Princeton	Specs	Timtec	Vitas-M	Zelinsky
Otava Chemicals		9	14	4	11	12	2
Pharmek	12		15	4	4	16	3
Princeton Biomolecular Research	62	52		37	51	64	86
Specs	10	7	21		24	21	33
Timtec	36	10	38	32		32	80
Vitas-M	66	69	77	44	51		84
Zelinsky Institute	3	3	28	19	34	23	

[a] The fraction (%) of vendor 1 compounds <in column> that are present in the vendor 2 database <in string>.
[b] XXXXX.

suppliers were presented, although the contribution of each supplier varied significantly (Fig. 5). To simplify compound management (as mentioned in the Introduction), we studied the dependence of the quality of the selected set on the number of suppliers. Based on the obtained data (Fig. 5), we selected 12, six, and three suppliers that contributed the most. The above-mentioned procedure for the 'ideal half million' selection was applied for the chemical space covered by these 12, six, and three suppliers, respectively. For the 12 and six suppliers, the generated space

contained 0.5 million compounds, whereas for three suppliers, the size of the space decreased to 384 520 compounds based on 19 226 scaffolds. We then compared these three spaces with the initial space from 33 suppliers at the compound and scaffold levels. Diversity at the compound level as well as QED were similar for all the three spaces (Figs. S7.01 and S7.02 in the Supplementary information online). However, a similar analysis at the scaffold level showed a significant decrease in diversity from the 33 to the three supplier sets (Fig. 7a).

TABLE 3

Bemis–Murcko loose framework scaffolding of the prefiltered chemical space covering 15 968 338 compounds

Number of structures per scaffold	Number of scaffolds			Resulting number of structures		
	Lead-like	3/75 rule	Drug-like	Lead-like	3/75 rule	Drug-like
≥50	13 156	8006	28 815	657 800	400 300	1 440 750
≥20	39 101	22 162	78 756	782 020	443 240	1 575 120
≥10	88 155	47 375	169 072	881 550	473 750	1 690 720
≥5	198 649	102 369	365 419	993 245	511 845	1 827 095
Total number of structures	6 544 044	3 705 803	14 191 016			
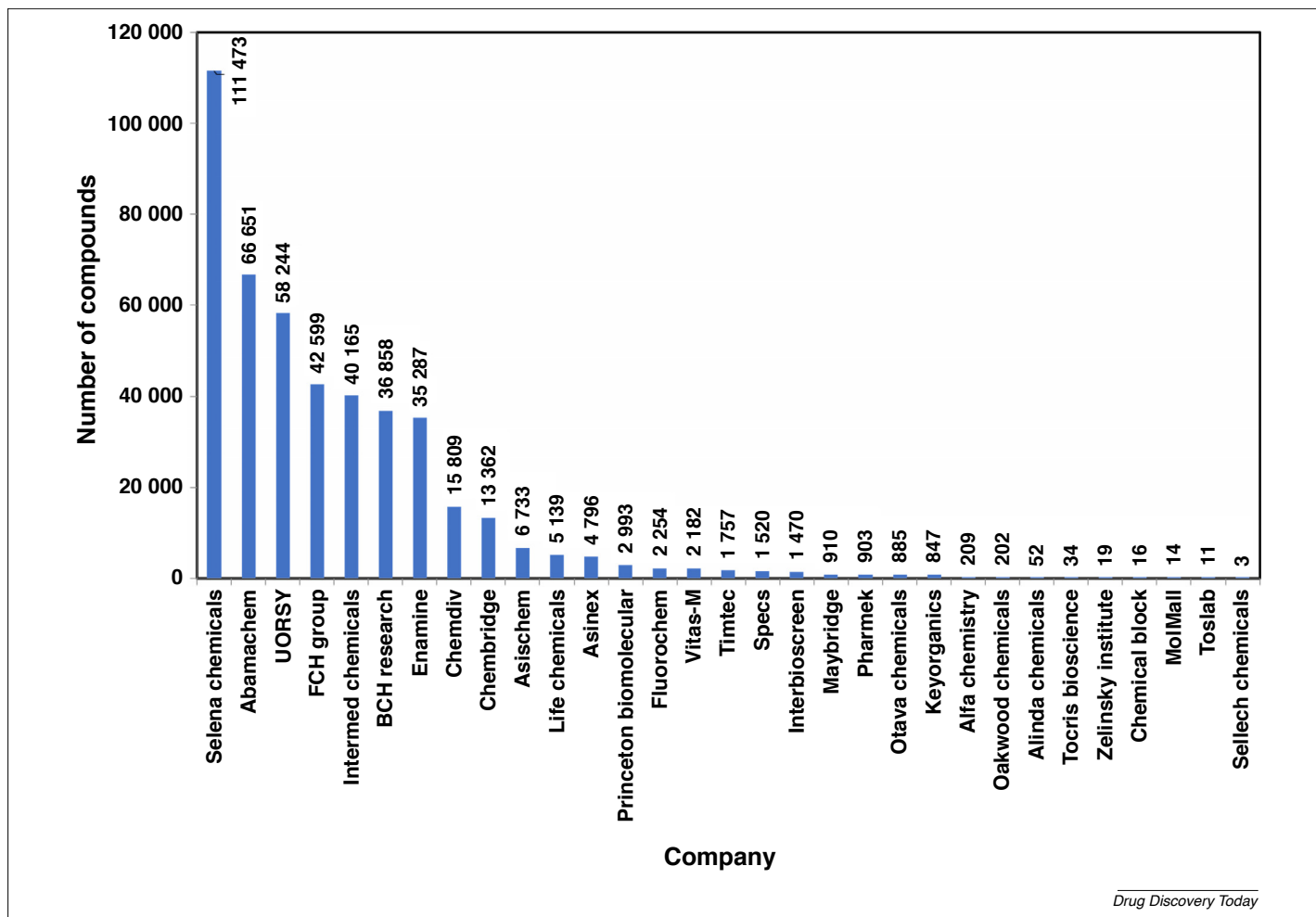
FIGURE 5

The contribution of vendors to the 'ideal half million' set.

The second model 'ideal million' set (50 compounds per scaffold) was collected using the above-mentioned algorithm. Similarly, for 50 compounds per scaffold set, only an 'ideal half million' could be generated. However, in contrast to the previous analysis, this resulted in a different level of contribution from each supplier (Fig. 6). We also analyzed the contribution from the top 12, six, and three suppliers. For 12 suppliers, applying the algorithm resulted in a 0.5 million compound set, whereas for six and three suppliers, the size of the r sets was 494 450 and 306 200 compounds based on 9889 and 6124 scaffolds, respectively. Compared with the 20 compounds per scaffold set analysis, decreasing the number of suppliers did not significantly influence the Tanimoto diversity at the compound level or the QED (Figs. S7.03 and S7.04

in the Supplementary information online), but did significantly decreased diversity at the scaffold level (Fig. 7b). In general, the comparison of the two sets (20 and 50 compounds per scaffolds) showed that the 50 compounds per scaffold set was significantly less diverse at the scaffold level. Therefore, the 20 compounds per scaffold set with the number of suppliers reduced to six or three subsets would be a pragmatic way to build a useful set of compounds for HTS screening campaigns based on compounds purchased from commercial sources.

The last step of our investigation was to compare the results from 33, 12, six, and three suppliers (for the libraries bearing 20 compounds per scaffold). For this purpose, we utilized the recently developed Generative Topographic Mapping (GTM) [71,72] be-
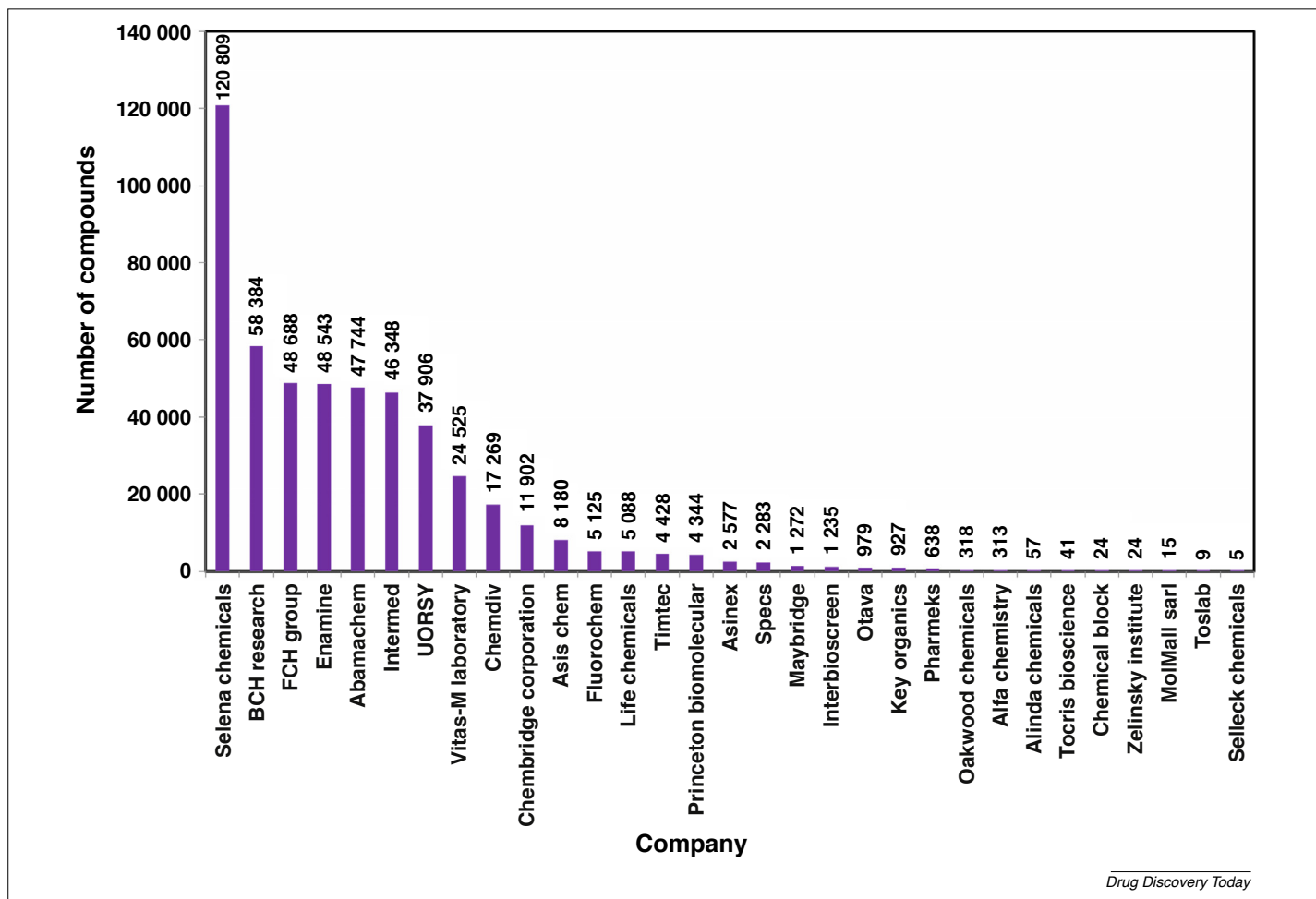
FIGURE 6

The contribution of vendors to the 'ideal half million 50 compounds per scaffold' set.
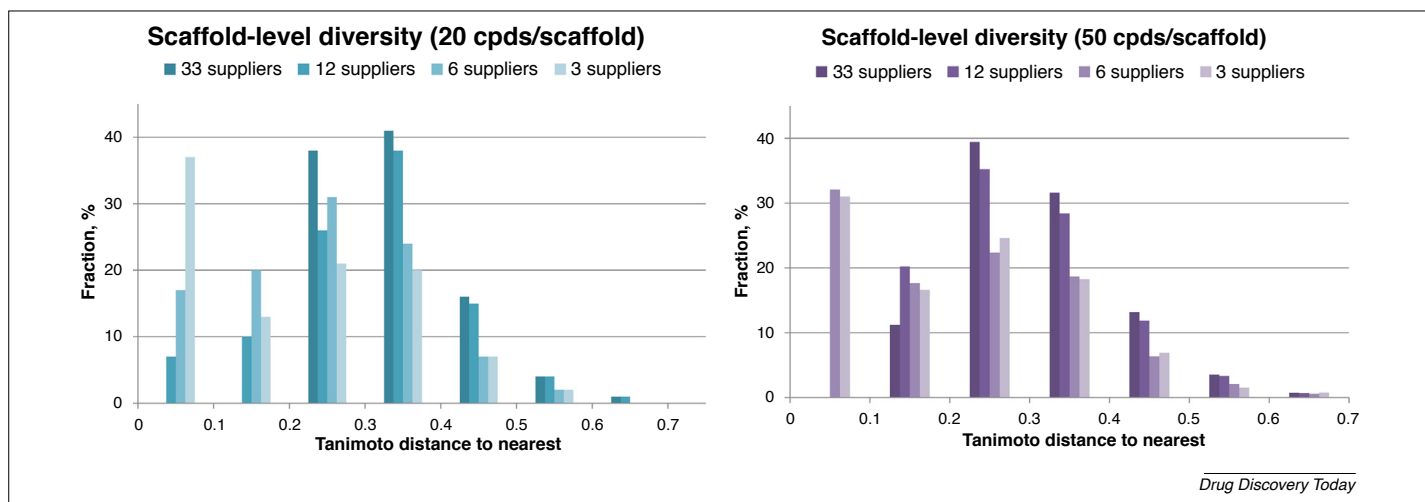

FIGURE 7

Comparison of the scaffold diversity of the libraries collected from 33, 12, six, and three suppliers. (a) For 20 compounds per scaffold set; (b) For 50 compounds per scaffold set.

cause it is considered the most efficient tool among the published methods for multiple descriptor chemical space comparison. The 1.5-million ChEMBL compound data set was used as a reference database. The four compound sets corresponded to three, six, 12, and 33 suppliers. These were mapped against the background of ChEMBL compounds, with blue zones corresponding to chemical space areas dominated by supplier compounds, versus dark–red zones containing (almost) exclusively ChEMBL compounds, after applying Bayesian normalization to compensate for the initial imbalance of set size (300 000–500 000 for supplier sets, versus 1.5-million ChEMBL compounds). Intermediate colors, from light red through yellow and green, corresponded to chemical space zones in which supplier and ChEMBL compounds mingled (increasing relative density of supplier compounds corresponding to a 'blue shift'). Three maps were built on the basis of the aforementioned principles, shown in Fig. 8.

Map #1 was based on ISIDA [73] force-field-type colored atom sequence counts acting as molecular descriptors. The force field types assigned to atoms (the CVFF forcefield typing rules were applied) were specific to their chemical environment and, therefore, this class of ISIDA fragment descriptors provides a fine-grained analysis of chemical space. The three-supplier set dominated the 'north-eastern' chemical space zone, clearly separated by a ChEMBL-dominated central part from some secondary 'islands' in both the north-western and south-eastern regions. Increasing the number of suppliers resulted in a gradually growth of overlap with the ChEMBL set, by embracing more compounds in the central area, which remained dominated by ChEMBL compounds while also starting to be populated by supplier molecules. The extent of library overlaps, calculated as the Tanimoto score of the mean vectors responsible from the supplier and ChEMBL libraries, respectively, increased from 0.28 (three suppliers) to 0.33 (six suppliers) to 0.42 (12 suppliers) and remained constant when all suppliers were considered.

Map#2 relied on ISIDA pharmacophore-type colored atom sequence count descriptors (i.e., it monitors pharmacophore pattern diversity). Therefore, it ignored the precise chemical nature of the atoms, rendered as hydrophobes, aromatics, HBA and HBD, cations, and anions, respectively. The three-supplier set provided significant coverage of the chemical space, with the only ChEMBL-dominated area close to the 'south pole' of the map. The addition of compounds from further suppliers gradually filled this initial diversity hole. The degree of library overlap was generally higher than in the more fine-grained map #1, and gradually increased from 0.51 (three suppliers) to 0.54 (six suppliers), 0.63 (12 suppliers), and 0.65 (all suppliers).

Map#3 was based on plain ISIDA atom sequence counts. Similar to map#1, it also focused on chemical constitution and connectivity patterns, but was less fine-grained than the latter; thus, the libraries are strongly overlap. On this map, the three-supplier library appears as a core collection that gradually expands (in particular, into the north-west and south-west regions) as
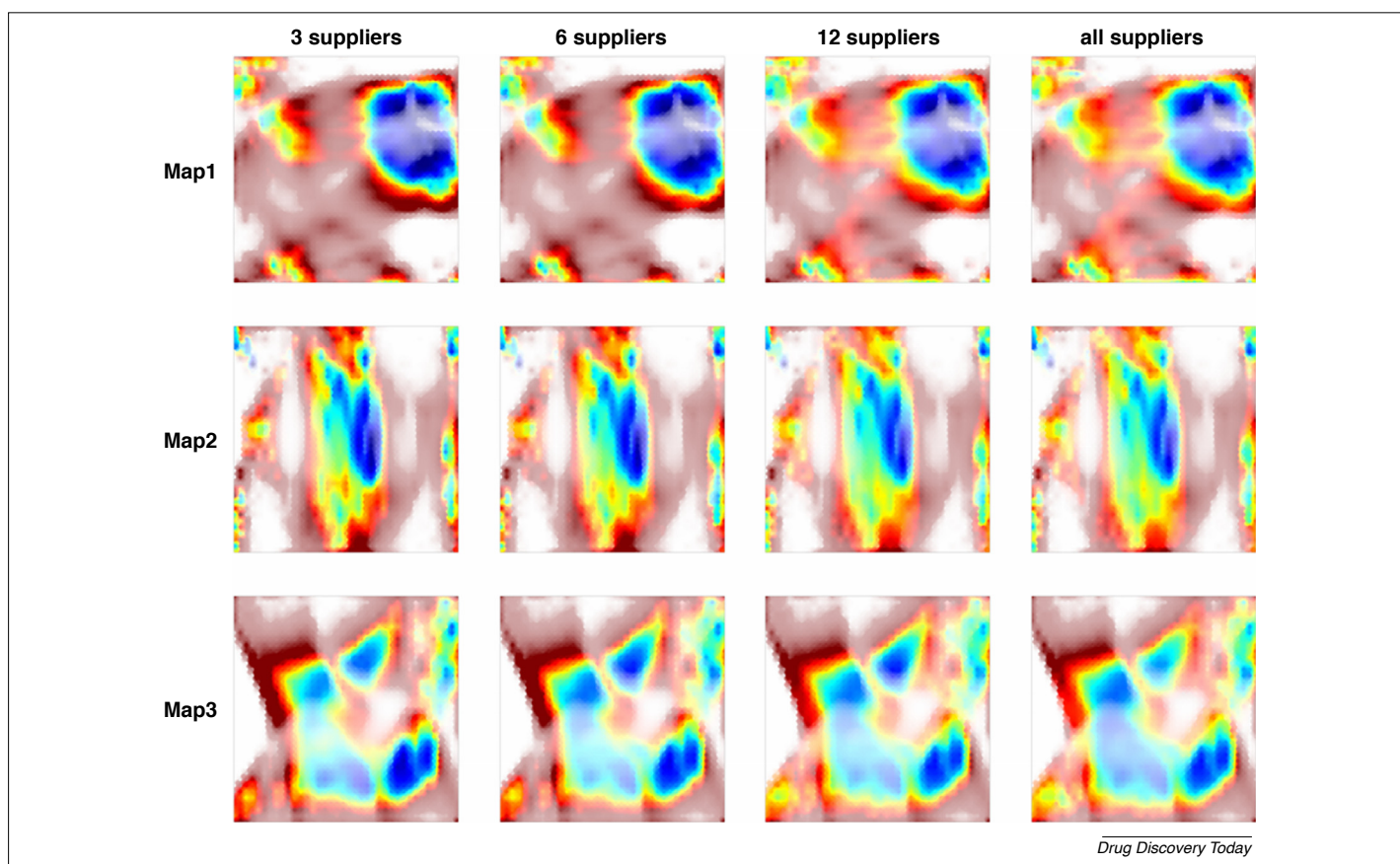
FIGURE 8

Generative Topographic Mapping (GTM) maps of four compound sets corresponding to three, six, 12, and 33 suppliers on the ChEMBL compounds background. See main text for key to colors.

compounds from further suppliers were added. Overlap degrees varied from 0.34 (three suppliers) to 0.40 (six suppliers), 0.47 (12 suppliers), and 0.49 (all suppliers).

Concluding remarks

As HTS has matured, our understanding of what features constitute a quality hit and lead has evolved. It is generally regarded that low lipophilic, and higher Fsp^3 properties are preferred. From our analysis, it appears that, over the past 10 years, the market has evolved to meet these demands, with new compounds from many suppliers having modern physiochemical properties. Currently, it is not possible to purchase an 'ideal' 1-million compound set (50 000 scaffolds, minimum of 20 compounds per scaffold). However, it appears that an 'ideal' 500 000 set can be purchased. If sample logistics is an issue, then we have shown that it is possble to purchase the 500 000 set from only six suppliers, with a 350 000 set available from just three suppliers. Many large companies have been through similar exercises and have built their screening decks accordingly. If you are considering building a screening deck *ab initio*, then it is possible to achieve this from purchasable space. In the interest of open innovation, we have made our data available online (www.awridian.co.uk/Resources). We are confident that, as new challenges in sample supply emerge, the market place will respond positively.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.10.016.

References

1 Winquist, R.J. *et al.* (2014) The fall and rise of pharmacology – (re-)defining the discipline? *Biochem. Pharmacol.* 87, 4–24

2 Erlanson, D.A. *et al.* (2016) Twenty years on: the impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* 15, 605–619

3 Goodnow, R.A., Jr *et al.* (2017) DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* 16, 131–147

4 Moffat, J.G. *et al.* (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.* 16, 531–543

5 Scannell, J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11, 191–200

6 Prasad, V. and Mailankody, S. (2017) Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA Intern. Med.* 177, 1569–1575

7 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–716

8 Macarron, R. *et al.* (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195

9 Peakman, M.-C. *et al.* (2015) Experimental Screening Strategies to Reduce Attrition Risk. In *Attrition in the Pharmaceutical Industry: Reasons, Implications, and Pathways Forward* (Alex, A., ed.), pp. 180–214, John Wiley & Sons

10 Feher, M. and Schmidt, J.M. (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 43, 218–227

11 Bakken, G.A. *et al.* (2012) Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* 52, 2937–2949

12 Koge, T. *et al.* (2013) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. *Drug Discov. Today* 18, 1014–1024

13 Morgan, P. *et al.* (2018) Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* 17, 167–181

14 Njoroge, M. *et al.* (2014) Recent approaches to chemical discovery and development against malaria and the neglected tropical diseases human African trypanosomiasis and schistosomiasis. *Chem. Rev.* 114, 11138–11163

15 Cooper, C.B. (2013) Development of *Mycobacterium tuberculosis* whole cell screening hits as potential antituberculosis agents. *J. Med. Chem.* 56, 7755–7760

16 Peña, I. *et al.* (2015) New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource. *Sci. Rep* 5, 8771

17 Scott, A. (2015) Sanofi off-loads R&D activities in France to Evotec. *C@EN* 93, 6

18 Cabrera, A.C. *et al.* (2016) Aggregated compound biological signatures facilitate phenotypic drug discovery and target elucidation. *ACS Chem. Biol.* 11, 3024–3034

19 Anon (2012) AstraZeneca and Bayer share their entire compound libraries. *Nat. Rev. Drug Discov.* 11, 739

20 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26

21 Todeschini, R. and Consonni, V., eds (2009) *Molecular Descriptors for Chemoinformatics*, Wiley-VCH Verlag GmbH & Co

22 Teague, S.J. *et al.* (1999) design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* 38, 3743–3748

23 Hughes, J.D. *et al.* (2008) Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* 18, 4872–4875

24 Congreve, M. *et al.* (2003) A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877

25 Jadhav, A. *et al.* (2010) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* 53, 37–51

26 Walters, W.P. and Namchuk, M. (2003) A guide to drug discovery: designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* 2, 259–266

27 Baell, J.B. and Holloway, G.A. (2010) Compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740

28 Bruns, R.F. and Watson, I.A. (2012) Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* 55, 9763–9772

29 Gorse, A.-D. (2006) Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* 6, 3–18

30 Gillet, V.J. (2008) New directions in library design and analysis. *Curr. Opin. Chem. Biol.* 12, 372–378

31 Nadin, A. *et al.* (2012) Lead-oriented synthesis: a new opportunity for synthetic chemistry. *Angew. Chem. Int. Ed. Engl.* 51, 1114–1122

32 Senger, M.R. (2016) Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discov. Today* 21, 868–872

33 Kitchen, D.B. and Decornez, H.Y. (2015) Computational Techniques to Support Hit Triage. In *Small Molecule Medicinal Chemistry: Strategies and Technologies* (Czechtizky, W. and Hamley, P., eds), pp. 191–210, John Wiley & Sons

34 Janzen, W.P. (2014) Screening technologies for small molecule discovery: the state of the art. *Chem. Biol.* 21, 1162–1170

35 Mullard, A. (2013) European lead factory opens for business. *Nat. Rev. Drug Discov.* 12, 173–175

36 Schuhmacher, A. *et al.* (2016) Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* 14, 105

37 Green, C. and Taylor, D. (2016) Consolidating a distributed compound management capability into a single installation: the application of overall equipment effectiveness to determine capacity utilization. *J. Lab. Automat.* 21, 811–816

38 Baell, J.B. (2013) Broad coverage of commercially available lead-like screening space with fewer than 350,000 compounds. *J. Chem. Inf. Model.* 53, 39–55

39 Baurin, N. *et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* 44, 643–651

40 Siroisa, S. *et al.* (2005) Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* 29, 55–67

41 Verheij, H.J. (2006) Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Divers.* 10, 377–388

42 Lucas, X. *et al.* (2015) The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* 55, 915–924

43 Chuprina, A. *et al.* (2010) Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* 50, 470–479

44 Petrova, T. *et al.* (2012) Structural enrichment of HTS compounds from available commercial libraries. *Med. Chem. Commun.* 3, 571–579

45 Wigglesworth, M.J. *et al.* (2015) Increasing the delivery of next generation therapeutics from high throughput screening libraries. *Curr. Opin. Chem. Biol.* 26, 104–110

46 Karawajczyk, A. *et al.* (2015) Expansion of chemical space for collaborative lead generation and drug discovery: the European Lead Factory Perspective. *Drug Discov. Today* 20, 1310–1316

47 Besnard, J. *et al.* (2015) The Joint European Compound Library: boosting precompetitive research. *Drug Discov. Today* 20, 181–186

48 Bickerton, G.R. *et al.* (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98

49 Firth, N.C. *et al.* (2012) A novel method to characterize the three-dimensionality of molecules. *J. Chem. Inf. Model.* 52, 2516–2525

50 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893

51 Langdon, S.R. *et al.* (2011) Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* 51, 2174–2185

52 Shang, J. *et al.* (2017) Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J. Cheminform.* 9, 25

53 Hughes, J.P. *et al.* (2011) Principles of early drug discovery. *Br. J. Pharmacol.* 162, 1239–1249

54 Lipkin, M.J. *et al.* (2008) How large does a compound screening collection need to be? *Comb. Chem. High Throughput Screen.* 11, 482–493

55 Renner, S. *et al.* (2011) Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem.* 3, 751–766

56 An, W.F. and Tolliday, N. (2010) Cell-based assays for high-throughput screening. *Mol. Biotechnol.* 45, 180–186

57 Mayr, L.M. and Fuerst, P. (2008) The future of high-throughput screening. *J. Biomol. Screen.* 13, 443–448

58 Downey, W. *et al.* (2010) Compound profiling: size impact on primary screening libraries. *Drug Discov. World Spring* 81–86

59 Glaser, V. (2010) High throughput screening retools for the future. *Bio-IT World Mag.* 8, 20–24

60 Hansson, M. *et al.* (2014) On the relationship between molecular hit rates in high-throughput screening and molecular descriptors. *J. Biomol. Screen.* 19, 727–737

61 Elkin, L.L. *et al.* (2015) Just-in-time compound pooling increases primary screening capacity without compromising screening quality. *J. Biomol. Screen.* 20, 577–587

62 Bakken, G.A. *et al.* (2012) Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* 52, 2937–2949

63 Kitchen, D.B. and Decornez, H.Y. (2015) Computational techniques to support hit triage. In *Small Molecule Medicinal Chemistry: Strategies and Technologies* (Czechtizky, W. and Hamley, P., eds), pp. 211–214, John Wiley & Sons

64 Harper, G. *et al.* (2004) Design of a compound screening collection for use in high throughput screening. *Comb. Chem. High Throughput Screen.* 7, 63–70

65 Nilakantan, R. *et al.* (2002) A novel approach to combinatorial library design. *Comb. Chem. High Throughput Screen.* 5, 105–110

66 Preston, S. *et al.* (2017) Screening of the 'Open Scaffolds' collection from Compounds Australia identifies a new chemical entity with anthelmintic activities against different developmental stages of the barber's pole worm and other parasitic nematodes. *Int. J. Parasitol. Drugs Drug. Resist.* 7, 286–294

67 Bell, A.S. *et al.* (2016) Plate-based diversity subset screening generation 2: an improved paradigm for high-throughput screening of large compound files. *Mol. Divers.* 20, 789–803

68 Chakravorty, S.J. *et al.* (2018) Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discov.* 23, 532–545

69 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263

70 Ashton, M. *et al.* (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Struct. Act. Relat.* 21, 598–604

71 Horvath, D. *et al.* (2017) Generative topographic mapping approach to chemical space analysis. In *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences* (Roy, K., ed.), pp. 167–199, Springer

72 Gaspar, H.A. *et al.* (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J. Chem. Inf. Model.* 55, 84–94

73 Ruggiu, F. *et al.* (2010) ISIDA property-labelled fragment descriptors. *Mol. Inf.* 29, 855–868

Summary

In this work, catalogs of 33 leading chemical suppliers have been analyzed separately and as a whole unity forming the chemical space of commercially available compounds. It was shown that over the past decade, commercially available chemical space has evolved to meet the main criteria for the quality drug candidates according to the current beliefs - like low lipophilic and higher Fsp3 properties etc. The feasibility of compiling an "ideal" diverse 1-million compound set (50 000 scaffolds, with a minimum of 20 compounds per scaffold) was also evaluated. However, it appeared that currently, it is impossible to purchase it even by combining catalogs of 33 vendors. In contrast, the "ideal" 500 000 compound set can be gathered from only six suppliers, with a 350 000 set available from just three vendors. Many large companies have built their screening decks in a similar way.

Several «ideal» screening datasets were created using compounds proposed by the different number of suppliers. Four differently collected «ideal» datasets have been mapped against ChEMBL collections using three uGTMs of the biologically relevant chemical space described previously. In all of them, it was clearly seen that there is a large area of biologically active chemical space (represented by ChEMBL compounds) that is not covered by any of the «ideal» datasets. Partially, it was caused by the filtration procedure applied while compiling ideal datasets. In any case, the presence of ChEMBL-specific areas on GTMs has raised a question of general correspondence between the biologically relevant and commercially available chemical space. This question became the main focus of the next project, described in the following chapter.

4.2 Searching for hidden treasures in commercially available and biologically relevant chemical spaces

Introduction

Nowadays, commercially available compounds are one of the primary sources of potential drugs. However, the currently known chemical space is far from being fully studied and apprehended by medicinal chemists. The existing studies of the purchasable collections are usually limited to the statistical analysis of chemical collections in terms of four groups of characteristics: physicochemical properties (e.g., molecular weight, log P, polar surface area, etc.), molecular complexity, diversity, and novelty (usually based on a simple scaffold analysis). Moreover, the scope of the mentioned works does not cover the entire chemical market but only up to 2% of the currently available compounds.

Trying to fill this gap, we compared almost a billion commercially available molecules from ZINC library with 1.6 million biologically tested molecules from ChEMBL. Depending on the selected hit identification strategy, ZINC and ChEMBL compounds were split into four groups: *fragment-like*[6, 7], *lead-like*[8, 9], *drug-like*[10, 11], *and PPI-like*[12] subfamilies. The purchasability of ZINC molecules was also assessed: commercial compounds were further split into "*ZINC-Real*" and "*ZINC-Tangible*" subsets. The latter concerns compounds that were not yet synthesized but can be

Main terminology

Fragment-based drug discovery – method of lead identification, based on the serach for small chemical fragments, which may bind to the biological target, and then combining them to produce a lead with a higher affinity.

Druglikeness concept - used in drug design to estimate compound oral bioavailbility by considering physicochemical properties influencing comppund's ADME profile.

Leadlikeness concept – implies usage of cut-off values in thephysico-chemical profile of chemical libraries used in drug design for lead identification. It is based on the observation that effective leads have lower molecular weight and complexity, smaller number of rings and rotatable bonds, are more polar comparing to drugs.

Tangible libraries - contain compounds that were designed as a result of the stock enhancement programs and have not been synthesized yet. Thus, 8−10 weeks are needed for their delivery and associated acquisition success rate ≈ 70%.

prepared upon request with an 70% success purchasability rate.[61]

The first uGTM, described in Chapter 3, was chosen as the main general-scale map that provides a bird's eye view of the biologically relevant chemical space. The density landscapes were used to analyze the chemotype distribution over each chemical subspace. Comparison of the density landscapes of ZINC-Real and ZINC-Tangible chemical spaces allows evaluation of the success of the enhancement strategies that first of all affect tangible collections.

The comparative landscapes featuring commercially available chemical space as opposed to the reference library containing biologically tested compounds allow to evaluate the extent of the biological relevance of purchasable libraries. In order to improve the resolution and level of detalization of such analysis, hierarchical GTM (hGTM) was used to reach down to the smallest clusters in the chemical space. Structural comparison of ChEMBL and ZINC compounds on the last level of such hierarchy allows detection of the previously hidden features of each library, identify what has been missed by chemical suppliers in the race to improve their catalogs and by medicinal chemists during the experimental biological exploration of the available chemical space.

Data preparation

Commercially available chemical space was represented by 1 369 004 023 compounds with a standard reactivity from the ZINC15[61] website retrieved in January 2019. Four purchasability categories were included:

- In stock - delivery in under two weeks, 95% typical acquisition success rate;
- Procurement agent - in stock, delivery in 2 weeks, 95% typical acquisition success rate;
- Make-on-demand - delivery typically within 8 to 10 weeks, 70% typical acquisition success rate;
- Boutique, where the cost may be high but still likely cheaper than making it yourself, 70% typical acquisition success rate.

The first two groups were combined, resulting in the Real subset of 13 196 748 compounds. All the rest forms the Tangible subset.

1 879 206 compounds were collected from the ChEMBL database version 25[59] in March 2019.

All datasets were standardized accordingly to the procedure implemented on the VS server of the Laboratory of Chemoinformatics at the University of Strasbourg (infochimie.u-strasbg.fr/webserv/VSEngine.html) using the ChemAxon Standardizer[117]. That included dearomatization and final aromatization (heterocycles like pyridone were not aromatized); conversion to canonical SMILES; salts and mixtures removal; neutralization of all species, except nitrogen (IV); the major tautomer generation and stereochemical information removal.

101 M compounds from the PubChem database were collected after analysis of ChEMBL- and ZINC-specific maximum common substructures were finished (December 2019) as an external independent dataset of biologically tested compounds that were not included in ChEMBL (mostly results of HTS). Those compounds were also standardized, and after removal of the stereoisomers, 80M molecules were left. 3.1 M of those compounds have been tested in at least one biological assay, while only 1.1M compounds were labeled as "active".

After standardization and stereoisomers deletion, 800 million ZINC compounds and 1.6 million compounds from CHEMBL database have been submitted to the removal of unwanted chemical functionalities due to potential toxicity reasons or unfavorable pharmacokinetics[118, 119]. These included potentially mutagenic groups such as nitro groups, groups likely to have unfavorable pharmacokinetic properties such as sulfates and phosphates; and reactive groups such as 2-halopyridines or thiols. Furthermore, compounds that are likely to interfere with typical HTS assays were also excluded[120]. It was realized by applying BRENK[118] and PAINS[120] substructure filter sets from RDKit[121] and standalone Lilly Med Chem filters[119]. Apart from substructure filters, each of the three resulting subsets (ChEMBL, ZINC-Real, and ZINC-Tangible) have been separated into four segments based on the medicinal chemistry concepts (**Table 4**).

Table 4. Target specification of the profiled compounds.

Parameters	Drug-like[19]	Lead-like[18]	Fragment-like[17]	PPI-like[20]
MW	≤500	≤400	≤300	[400; 700]
LogP	≤5	[-3.5;4]	≤3	[1,5; 6,5]
HBD	≤5	≤5	≤3	-
HBA	≤10	≤8	≤3	[4; 9]
RNG	≤10	≤4	-	[3; 6]
RTB	-	≤10	≤3	-
TPSA	-	-	≤60	-

Article

Chemography: Searching for Hidden Treasures

Yuliana Zabolotna, Arkadii Lin, Dragos Horvath, Gilles Marcou, Dmitriy M. Volochnyuk, and Alexandre Varnek*

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The days when medicinal chemistry was limited to a few series of compounds of therapeutic interest are long gone. Nowadays, no human may succeed to acquire a complete overview of more than a billion existing or feasible compounds within which the potential "blockbuster drugs" are well hidden and yet only a few mouse clicks away. To reach these "hidden treasures", we adapted the generative topographic mapping method to enable efficient navigation through the chemical space, from a global overview to a structural pattern detection, covering, for the first time, the complete ZINC library of purchasable compounds, relative to 1.6 million biologically relevant ChEMBL molecules. About 40 000 hierarchical maps of the chemical space were constructed. Structural motifs inherent to only one library were identified. Roughly 20 000 off-market ChEMBL compound families represent incentives to enrich commercial catalogs. Alternatively, 125 000 ZINC-specific compound classes, absent in structure–activity bases, are novel paths to explore in medicinal chemistry. The complete list of these chemotypes can be downloaded using the link https://forms.gle/B6bUJj82t9EfmttV6.
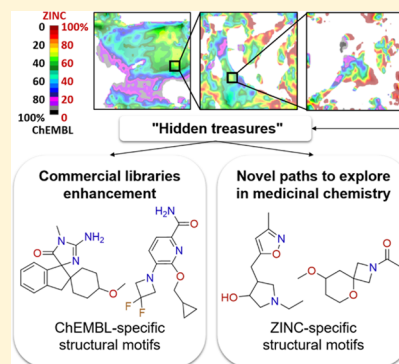
INTRODUCTION

Nowadays, the number of molecules available to medicinal chemists is huge. The ZINC database merges commercial catalogs proposed by numerous chemical suppliers and contains more than 1.4 billion compounds.[1] It includes both already synthesized or in-stock compounds and tangible molecules. Despite being just a tiny fraction of the estimated number of possible drug-like molecules (around 10^{33} structures),[2] the currently known chemical space is far from being fully studied and apprehended by medicinal chemists. For example, ChEMBL,[3] containing biologically studied compounds extracted from the scientific literature, is a thousand times smaller than ZINC. Thus, while chemical suppliers compete to enumerate the higher number of new virtual molecules,[4] already existing compounds are largely unexplored from a drug discovery perspective.

Within the 2 last decades, the usefulness of purchasable screening libraries playing the role of a source of potential drugs has been evaluated in numerous reports.[5−12] These studies typically rely on a statistical analysis of chemical collections in terms of four groups of characteristics: physicochemical properties (e.g., molecular weight, log P, polar surface area, etc.), molecular complexity, diversity, and novelty (usually based on a simple scaffold analysis[13]). All of these reports provide an important insight into the evolution of medicinal chemistry-relevant properties of commercially available compounds and their distribution across screening libraries of different chemical suppliers. Yet, the scope of the mentioned works does not cover the entire chemical market but only up to 2% of the purchasable compounds (16M out of

800M unique ZINC molecules). Moreover, there is a lack of chemical analysis of commercially available libraries. Indeed, direct references to molecular structures were limited to the typical scaffold population analysis—a convenient and yet biased way to comprehend structural diversity.[14] The same scaffold may be adorned with radically different pharmacophore patterns and, hence, have completely different biological effects. On the other hand, the same pharmacophore may be "incarnated" by radically different scaffolds and yet exhibit similar activity.[15]

All of those works aim to analyze only the current state of the chemical market without trying to identify and, if possible, fill the gaps in the purchasable chemical space. One way to evaluate such possible incompleteness is a comparison of commercial catalogs with a reference subset of molecules possessing desired properties. Such an approach was previously adopted by Shelat and Guy in their study of the biological relevance of screening libraries.[16] They compared some purchasable chemical collections (≈2M unique structures) with a set of known drugs (≈8k compounds). The results have shown that there is only a 14% scaffold overlap between analyzed subsets, which brings us to the conclusion that commercial chemical space at that time was not sufficiently
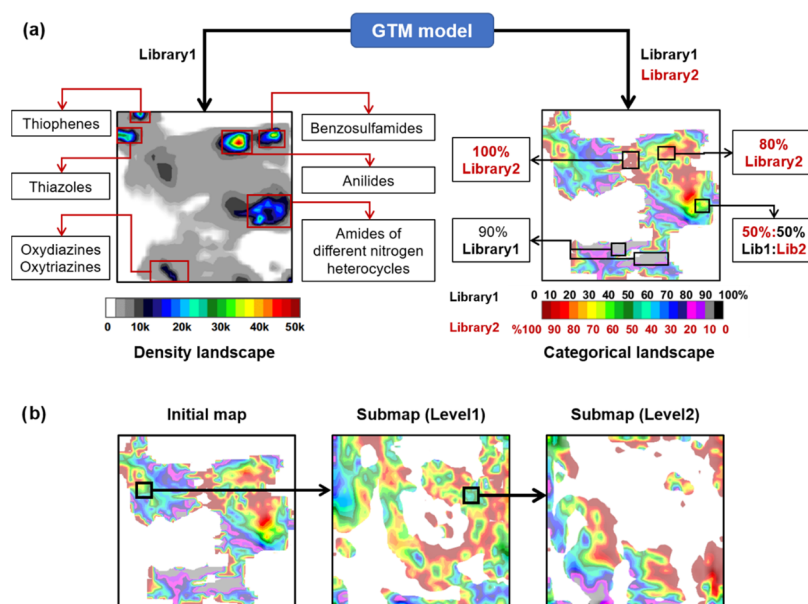
Figure 1. Generic scheme of library analysis and comparison with GTM. (a) Left: density landscape used to analyze the distribution of different compound classes across the chemical space (color spectrum matches the cumulated responsibility, corresponding to the number of resident compounds); right: a categorical landscape rendering chemical space regions occupied by two libraries (the color code matching the proportion of residents from each library). (b) Schematic overview of the Hierarchical GTM (HGTM) navigation through the highly populated areas of the chemical space — compounds, extracted from the zone of interest, are used for constructing a new map, now focused only on this region of chemical space.

covering biologically relevant compounds. The challenging goal of increasing that coverage can hardly be achieved by unguided compound enumeration. It requires a deep understanding of the main features of both purchasable and biologically relevant chemical space.

In this context, our study focuses on two goals: (i) commercial chemical space enhancement and (ii) its exploration. The first means identification of biologically relevant compounds that are absent from the current chemical market. Such molecules, being synthesized in academic laboratories, small startups, big pharmaceutical companies, or coming from natural product-based programs,[17] are also entering biological assays and results of these tests eventually become publicly available. These biologically relevant compounds and especially their untested analogs, if added to the commercial catalogs, could be highly useful in further screening campaigns and SAR studies and, thus, become good starting points for the development of new "best sellers" of the chemical market. Reciprocally, not all commercially available compounds have been tested in biological studies. The compound classes that have been overlooked by medicinal chemists can be used for expanding the scope of the biological exploration of the commercially available chemical space.

To find such "hidden treasures", we performed a thorough chemical analysis of the drug discovery-oriented commercial chemical space, featuring (after standardization and duplicate removal) 800M ZINC compounds, versus 1.6M molecules that have already attracted the attention of medicinal chemists and were therefore captured in the ChEMBL database together with their observed biological activities. Both ZINC and ChEMBL compounds were split into four groups depending on the type of biological tests and selected drug design strategy, resulting in fragment-like,[18] lead-like,[19] drug-like,[20] and protein−protein interaction (PPI)-like[21] subfamilies. The purchasability of ZINC molecules was also assessed: they were

further split into ZINC-Real, in-stock compounds directly available for purchase, and ZINC-Tangible, compounds that can be synthesized upon request.

Thousands of chemotypes, specific only to ChEMBL or ZINC libraries, were detected for each of the mentioned subspaces. It was done using one of the most efficient chemography methods of dimensionality reduction, generative topographic mapping (GTM),[22] that has already proven to be a successful approach for visualization and versatile analysis of large chemical libraries.[23] It produces easily readable two-dimensional (2D) maps of chemical space—a very convenient way for navigating through billions of compounds.

It was found that commercially available libraries are missing numerous compound families known to include biologically active members—highly potent inhibitors of important biological targets. Some examples of ChEMBL- and ZINC-specific chemotypes are discussed in the text, while the full list of these structures—a potential source of inspiration for synthetic and medicinal chemists—can be downloaded using the link https://forms.gle/B6bUJj82t9EfmttV6. It is noteworthy that the ZINC-specific maximum common substructures (MCSs) identified in this work, which were absent in both ChEMBL and PubChem[24] (revealed by the secondary substructure check), were then in silico profiled against 749 ChEMBL targets. It was done with the help of the GTM Profiler tool[25] used to evaluate their potential usefulness in drug design (http://infochim.u-strasbg.fr/webserv/VSEngine. html).

Chemography as a Versatile Tool for Chemical Space Analysis. Both chemography, as an "art of navigating" in chemical space,[26] and activity/property prediction should be used for chemical space analysis. The first is needed to navigate through the complex structure of the chemical data, and the second might serve to set the landmarks (identify compounds potentially possessing desired properties, by predicting those

properties, in the absence of experimental data). Also, the chosen approach must be "Big Data"-compatible. Generative topographic mapping, or GTM, conveniently fulfills all of these requirements. Briefly speaking, it translates compounds from the initial multidimensional descriptor space to a 2D latent space, called a 2D map. In contrast to self-organizing maps,[27] GTM distributes molecule projection over the map with node-specific probabilities (responsibilities) instead of unambiguously assigning each compound to only one point on the map. This smoothness enables the creation of GTM landscapes—cumulated compound responsibility patterns, colored by average values of different properties, e.g., density, biological activity, assigned class, etc. (see examples in Figure 1a). The details of the method are provided in the Supporting Information.

Walking over this map and performing an in-depth chemotype analysis of the residents of the local map zones is a rational and intuitive way to systematically "browse" the chemical space and get acquainted with the structural patterns it hosts. In this work, those patterns were characterized by maximum common substructures (MCSs)—the largest structural fragments that aim to generalize common features of the group of molecules they represent.[28] These MCSs were defined as substructural fragments that contain at least 30% of each molecule they represent. An MCS was preferred over the widely used scaffold concept because it is open-ended and adaptive: it may coincide with the scaffold or be more specific by including key substituents (side chains) if appropriate. The algorithm that combines both GTM and MCS detection was presented by Lin et al.[29] and is briefly discussed in the Supporting Information.

Yet, 2D maps cannot accommodate a huge number of compounds while capturing fine differences between close neighbors: a hierarchical zooming approach will be required to let the user capture the details of the chemical population at any point of the global map and reach down to hidden treasures buried beneath millions of compounds. Hierarchical GTM (HGTM),[29,30] a.k.a "Zooming", is a technique that trains a new map on a set of compounds extracted from a given zone on the parent map to ensure a locally optimal mapping (Figure 1b). The zoomed map is free to fit the local compound distribution, with no constraints to simultaneously match all of the other compounds—which is the key benefit, beyond the obvious gain in resolution (the latter could have been easily achieved by imposing a finer grid mesh on the global map).

Last but not least, with a robust structure—activity set used to create an activity landscape (a landscape colored by activity values), the map can be turned into a potent quantitative structure—activity relationship/quantitative structure—property relationship (QSAR/QSPR) model.[25,31−33] Predictivity of those models can be quantitatively determined and serve as a guide in the search for "the best map" parameters configuration. In this way, our group built seven optimized "Universal" maps of the drug design-relevant chemical space, selected for their ability to host as many predictive activity landscapes, for different drug targets with enough structure—activity data reported in ChEMBL.[25] Those maps are the basis of the GTM Profiler—a virtual screening tool that allows to predict the compound activity against 749 biological targets. It is extremely time-effective for already mapped molecules. The previously reported "top" Universal map serves here as the principal tool for the biologically biased analysis of the commercial compound space.

RESULTS AND DISCUSSION

Chemical Analysis of the Commercially Available Chemical Space. Initially, 1.3 billion (out of total 1.5 billion) compounds from ZINC15, passing built-in "standard reactivity" filter, and 1.8 million molecules from ChEMBL (version 25) were collected for this project. After structure standardization and stereoisomer "fusion" into a common, stereochemistry-depleted representation, 800 million ZINC and 1.6 million ChEMBL unique structures remained. Compounds with unwanted functionalities were filtered out (Table S1), and four subsets associated with different stages and strategies of drug discovery were defined (Table 1). Commercially available

Table 1. Size of the Medicinal Chemistry-Relevant Subsets after Standardization and Appropriate Filtration

	ChEMBL	ZINC-Real	ZINC-Tangible
fragment-like	15 398	103 530	2 772 851
lead-like	361 051	3 253 343	329 893 210
drug-like	668 222	5 158 676	516 492 788
PPI-like	229 570	1 248 875	63 632 835

compounds were split according to their purchasability into ZINC-Real and ZINC-Tangible. The first group contains all compounds that have been already synthesized in a sufficient quantity and thus can be delivered within 2 weeks to the buyer with a 95% acquisition success rate. The second one, in contrast, contains compounds that were designed by suppliers as a result of the stock enhancement programs and have not been synthesized yet. Thus, 8−10 weeks are needed for their delivery and acquisition success rate is about 70%.[1] Tangible libraries are considered as the source for the chemical enhancement of the Real ones. They can be readily made from existing building blocks according to the well-defined procedures,[34] approved by synthetic chemists. Therefore, ZINC-Tangible compounds were used in this study rather than de novo generated molecules[35,36] of uncertain chemical feasibility. Further details about data preparation and filtering rules can be found in the Supporting Information.

The present analysis employs Universal map #1 as the best one out of the previously built general-purpose chemical space maps.[25] It was constructed in a way to be able to predict 618 biological activities present in ChEMBL database. Being multitarget-oriented, this map can be considered as a generalized framework for biologically biased chemical space visualization. It is based on one of the ISIDA fragment descriptors—atom sequences with a length from 2 to 3 atoms labeled by CVFF Force Field types and Formal Charges labels.[37] See more details about the construction of Universal map #1 in the Supporting Information.

First, each of the above-mentioned ZINC subsets was projected onto the universal map. Density landscapes of the subsets were built to obtain a general overview of the structural features of the purchasable chemical space (Figure 2). Interestingly, the commercial compounds are distributed in a highly imbalanced manner: the major part of the map area is rather sparsely populated (gray zones), by contrast to a few outstanding density peaks (multicolored regions). In Figure 3, the structural analysis of the densest regions of the lead-like ZINC-Real part of the chemical space is presented: characteristic MCSs of some zones are shown. The density imbalance goes in correspondence with the previously reported unequal compound distribution across different compound classes.[11,12]
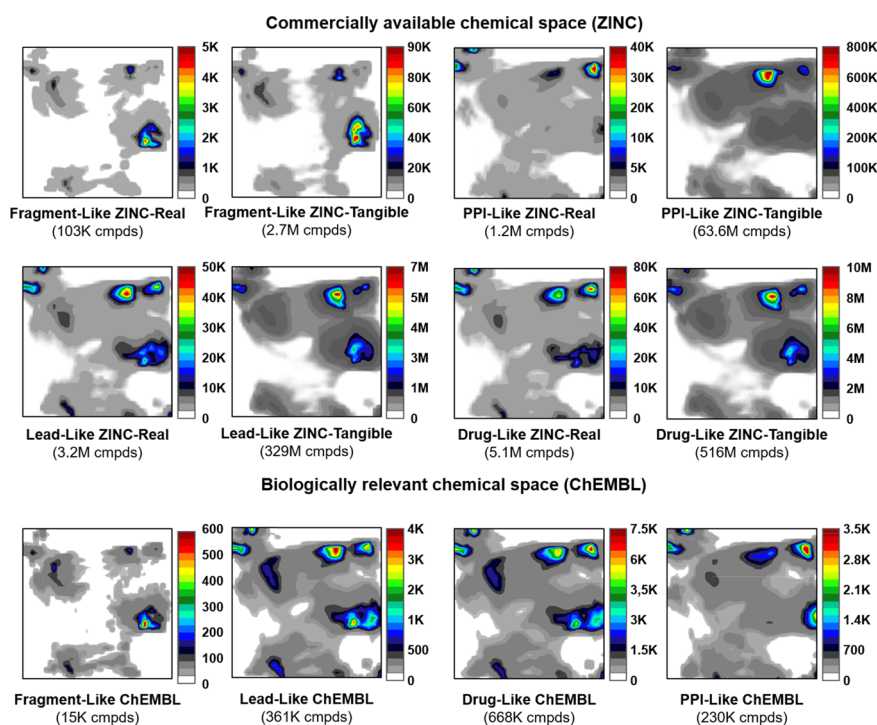
Figure 2. Density landscapes of commercially available (ZINC) and biologically relevant (ChEMBL) subsets. The color scale renders the corresponding number of compounds residing in each colored node of the map.
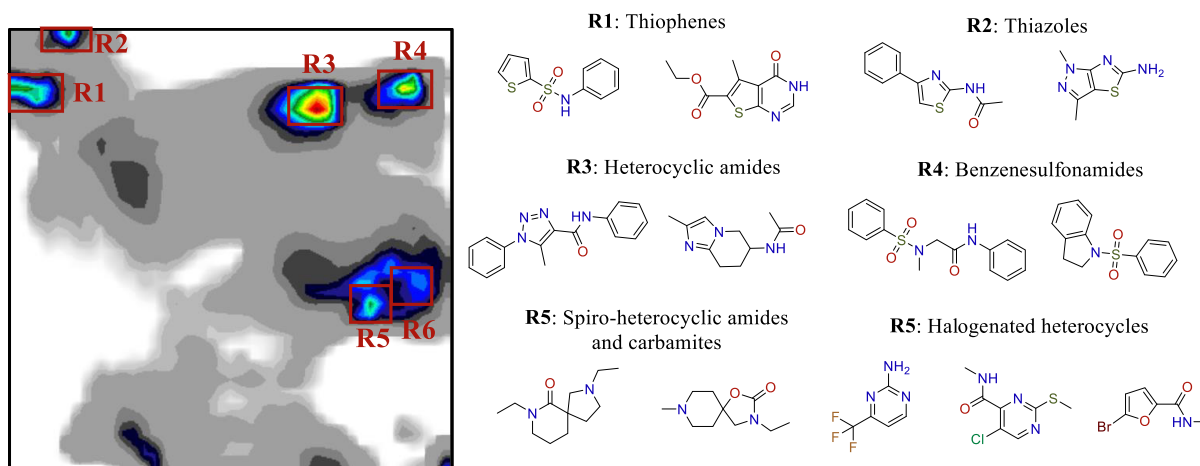

Figure 3. Examples of the most frequent structural motifs from the densest regions of the lead-Like ZINC-Real map.

An overrepresentation of synthetically accessible benzenesulfonamides, anilides, and other amides is noticed (Figure 3: regions R3, R4, and R5). These chemical subfamilies echo, first, the extreme popularity of combinatorial chemistry methods in the 20th century. Based on the limited sets of building blocks and simple reactions, they allowed synthesis of large numbers of compounds at the cost of limited chemical diversity. At the same time, the complexity of the synthetic path for some compounds prevented the mass production of their analogs.

The second reason is medicinal chemistry demand, which has also reshaped purchasable libraries significantly. For example, sulfonamides, the main inhabitants of the R4 region, are known for their antibacterial properties for almost 100 years. Back in time, together with antibiotics, they revolutionized the medicinal approach for treatment of various infections, moving it from immuno- to chemotherapy.[38] Other examples are thiophene-containing compounds (region R1) that possess diverse therapeutic properties such as antimicrobial, anticancer, anti-inflammatory activity, etc.[39] In addition, the thiophene cycle is highly popular in medicinal chemistry due to its bioisosteric correspondence with phenyl.

The previous century's synthetic methods and medicinal chemistry demands are still influencing the current chemical market.[40] This historical bias can be a dangerous limitation for discovering new valuable patterns in medicinal chemistry, novel chemotypes with a specific activity. Since tangible ZINC libraries have been designed rather recently, in theory, their compound distribution should be more balanced than those of in-stock collections. In practice, all of the analyzed subsets of ZINC-Real and ZINC-Tangible are very similar. Shapes of occupied areas and positioning of high-density regions are
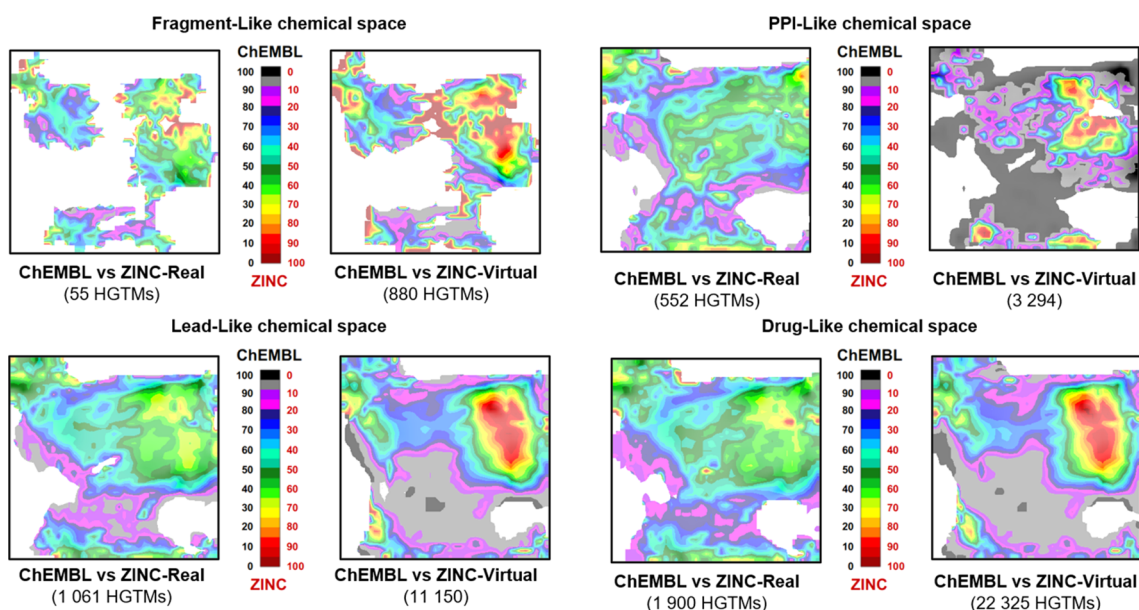
Figure 4. Categorical landscapes of the medicinal chemistry-relevant subsets of commercially available chemical space. Each map visualizes compounds both from ChEMBL (zones colored in black) and ZINC (colored in red). White regions correspond to the empty areas. All colors in between correspond to the various normalized proportion of compounds from different subsets, projected into a particular node of the map (see the Supporting Information). Numbers in parentheses show how many subsidiaries or "zoomed" GTMs were built.
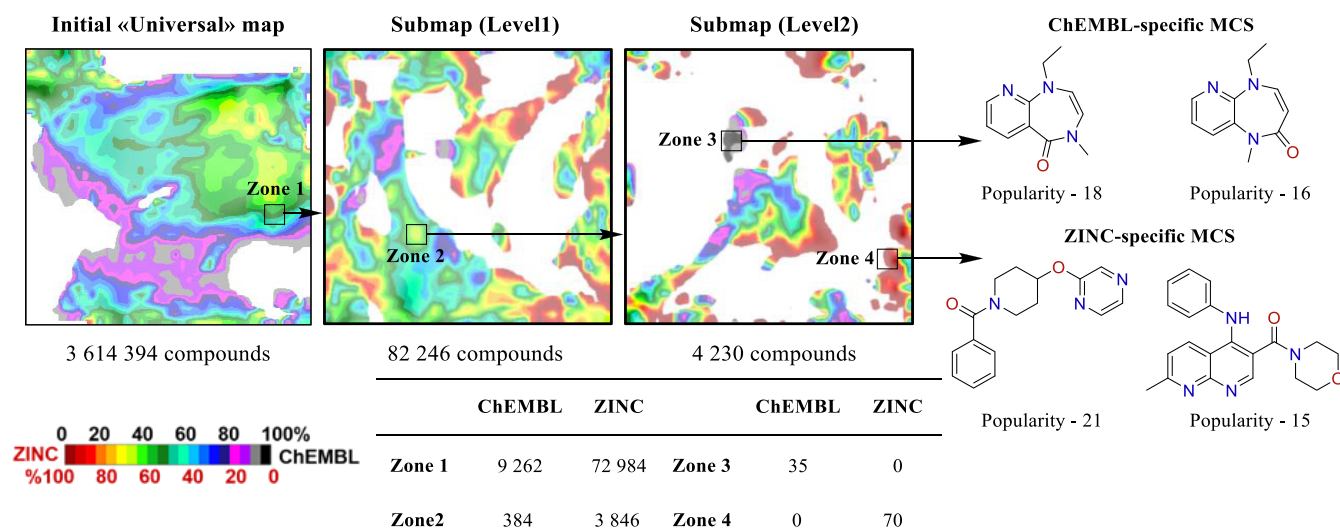
Figure 5. HGTM navigation of the highly populated areas of the chemical space: Lead-Like ChEMBL vs ZINC-Real example. The table provides the composition of each highlighted zone. Starting from the dense mixed zone 1, through the two levels of zoom, small purely ChEMBL (zone 3) and ZINC (zone 4) subareas are detected. Corresponding MCSs and their popularity (number of compounds that contain each structural fragment) are also reported.

similar (Figure 2). Although tangible libraries increase the total number of compounds on the market, they still tend to sample the same areas of the chemical space that are already overpopulated by in-stock libraries. This means that the current strategies of the commercial library enhancement do not provide a uniform chemical space sampling, and thus there is an urgent need for their improvement.

In Search of Hidden Treasures. Commercial chemical space is huge and thus expected to include novel chemotypes that were never subjected to biological testing so far. Moving them from the chemical store onto a shelf of the medicinal chemistry lab might open new opportunities in drug discovery. At the same time, suppliers might miss some important types of compounds, highly potent drug design candidates, that were

developed and tested in small companies or academic laboratories. These compounds are of high interest for medicinal chemists, and their presence in the commercial catalogs will certainly enrich the latter.

In search of these hidden treasures, a detailed comparison of ZINC and ChEMBL libraries was performed. From a "bird's-eye" perspective, the ChEMBL and ZINC chemical spaces coincide fairly well: in Figure 4, for each of the landscapes, there are only a few small zones in which the extremes of the color spectrum (local population exclusively stemming from one of the libraries) can be observed. However, this resolution level is certainly not sufficient, as one single node of the map may contain up to several millions of compounds (Figure 2), forcing dissimilar compounds to share common zones. The
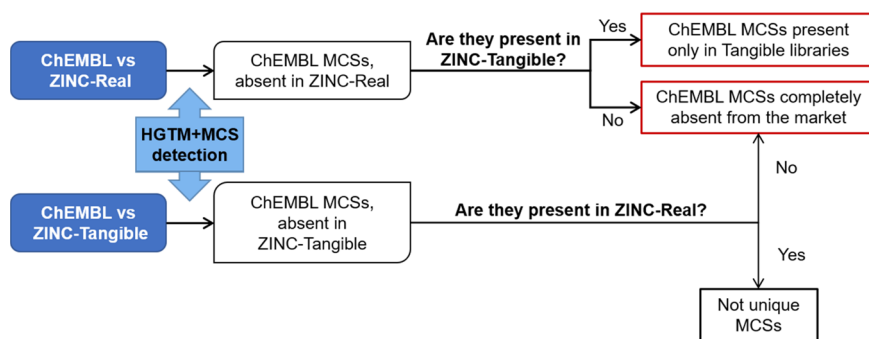
E

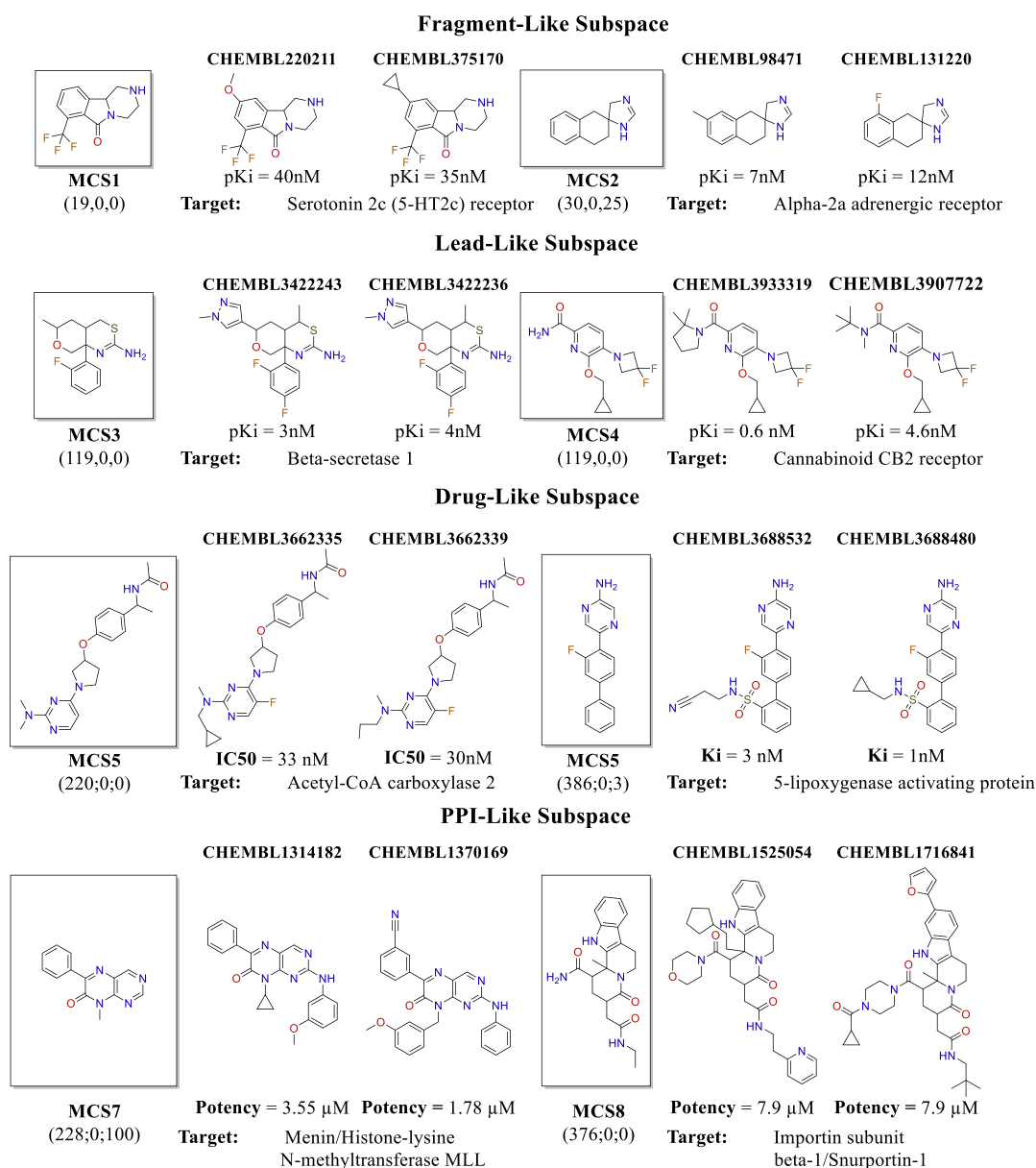**Figure 6.** Schematic workflow: searching for ChEMBL-specific MCSs with no commercial coverage.

Figure 7. Examples of the highly potent inhibitors, incarnating one of the reported unique ChEMBL substructures, recommended for the chemical space enhancement. Numbers in parentheses under each MCS identify the number of corresponding compounds containing this MCS in ChEMBL, ZINC-Real, and ZINC-Tangible libraries, respectively. All reported targets are *Homo sapiens* proteins with high therapeutic importance.

HGTM approach has been used to further navigate through highly populated areas. Up to five zooming levels were used to build about 40 000 "child" maps (Figure 4). All zones

containing in total more than 1 000 compounds were zoomed, while others were subjected directly to the MCS detection protocol.[29] For example, in the landscape hosting 3.6M lead-

like [ChEMBL + ZINC-Real] compounds (Figure 5), zone 1 is equally frequented by both libraries and contains more than 82 000 compounds. Two zooming iterations of this zone reveal a detailed landscape where areas with unique substructures (and, hence, chemotypes) can be found for each library (zone 3 and zone 4).

First, we focused on MCSs present in ChEMBL but not in ZINC. The workflow of their search is depicted in Figure 6. ChEMBL subsets (fragment-like, lead-like, drug-like, and PPI-like) were compared pairwise to ZINC-Real and ZINC-Tangible. The ChEMBL-specific MCSs, locally discovered as a result of such comparison, were used as queries in a substructure search against the corresponding ZINC-Real and ZINC-Tangible subsets. The absence of substructure hits means that these MCSs are not only zone-specific but unique to the respective subspace of biologically tested compounds and absent from the supplier libraries. Several examples of the potent nanomolar inhibitors containing some of the specific substructures for each of the analyzed subsets are shown in Figure 7. For more examples of ChEMBL-specific MCSs, see Table S2.

Most of the new ChEMBL substructures are much more complex than simple Bemis−Murcko scaffolds. For some substructures, it is the side chains that make them unique—the corresponding scaffolds with different decorations can be present on the market. This is the key advantage of our MCS-based search for characteristic substructures over a rigid scaffold-based approach. Figure 7 includes compounds active against therapeutically important targets. Those compounds and especially their analogs can be useful not only in the context of their known activities but also (and more so) in other drug design campaigns featuring other biological targets.

The absence from the commercially available chemical space of so many potentially very important compound families, known to include biologically (very) active members, is somehow intriguing—after all, those molecules were produced and tested, but somehow left no trace of precursors or analogs in commercial space. Several plausible explanations may exist—the "unique" MCS may emerge during the reaction, thus not be present in commercial building blocks, the compound was produced from proprietary building blocks, etc. Some of the ChEMBL-specific chemotypes can be missing from the vendors' libraries because they are part of the intellectual property space, which covers compounds protected by the patents. Unfortunately, the analysis of the intellectual property chemical space is not straightforward. A majority of patented structures are represented in a form of Markush structures, making these libraries impossible to cartograph (as prerequisite individual enumeration and molecular descriptor calculation for the combinatorially enumerated structures covered by a Markush formula may be too costly or outright unfeasible). Furthermore, not all of the mechanically enumerable Markush substituent combinations stand for chemically stable compounds—and even less represent confirmed actives. Specific tools for Markush-targeted substructure querying and even (connectivity-driven) similarity search tools exist but more sophisticated approaches involving information-rich descriptors, such as topological pharmacophore patterns, cannot be applied. Users will be free to submit any species of interest highlighted by our tool to a state-of-the-art check against patent libraries, but in our opinion no closer integration can be envisaged—the rigorist, connectivity-centric

legal status of a compound is not easy to reconcile with its fuzzy-logics-based responsibility patterns.

It should also be noted that the presence of a particular chemotype in the patents libraries yet does not mean that respective compounds cannot be synthesized or used in drug design campaigns. The point is that some patents protect only compound usage against a specific biological target or family of targets, leaving the freedom to operate outside of the specified research area. Such compounds can still be used in primary screening campaigns against novel biological targets.

The entire list of concerned MCS is freely available and, in our opinion, is an interesting source of enrichment of the purchasable in-stock libraries enhancement.

Biological Exploration of the Currently Available Chemical Space. The complementary application of this work is the detection of biologically unexplored regions of chemical space, e.g., ZINC-specific MCS. The same approach highlighted two sets of ZINC-Real and ZINC-Tangible-specific substructures derived from compounds not found in ChEMBL. Table S3 shows a diverse set of examples.

One might argue that some of those compounds could have been not "overlooked" by medicinal chemists but rather intentionally discarded from the screening campaigns. However, the herein employed standardization and filtering procedure should have eliminated most of the obviously reactive compounds or potential pan-assay interference compounds (PAINS) from the 800M filtered pool of ZINC compounds (albeit there is no absolute consensus of what precisely "unwanted" structures are). Thus, to dispel remaining doubts, additional analysis of the key substructures as a potential source of the highly potent hits was performed.

The ultimate pertinence of herein highlighted ZINC-specific MCSs for biological exploration of the chemical space will only be completely validated by actual experimental screening of those compounds, by MedChem groups pursuing specific drug discovery projects. This path is beyond the present work, which limits itself to present some indirect hints of the usefulness of these compounds, notably by (i) investigating whether those types of compounds have been tested already, without being reported yet in ChEMBL database, or (ii) predicting biological properties of the compounds of interest using the same universal map-based property landscapes—a fast, robust, and intuitive approach directly emerging from the chemographic concept.

Not being present in ChEMBL is not yet synonymous with being "off the beaten path". ChEMBL focuses mostly on the higher-level (dose−response) biological data, but some of the ZINC-specific MCSs might have served in high-throughput screening (HTS) campaigns reported elsewhere. PubChem, the largest collection of structure−activity data including high-throughput screening (HTS) reports, has been chosen in this study as an alternative external subset. A total of 101M compounds, 80M of which are unique structures (stereo-isomers were considered duplicates), were collected after the analysis of ChEMBL- and ZINC-specific maximum common substructures (December 2019); 3.1M of those compounds have been tested in at least one biological assay, while only 1.1M compounds were labeled as "active" by PubChem.

In a search for the potential drug candidates out of ZINC-specific subspace, around 24k of lead-like ZINC-Real unique MCSs (absent not only in lead-like subset but also in the unfiltered version of ChEMBL) were used as queries against 3.1M biologically tested PubChem compounds, but only

molecules marked as active were reported as hits. The lead-like real subset was selected as the most relevant with respect to the HTS demands and instant purchasability of corresponding compounds.

As a result, 9 575 ZINC MCSs were found in PubChem. For 1 628 of those MCSs, there were 4 520 PubChem compounds labeled as actives in 1 772 different biological assays. Among them, one of the recent studies of natriuretic polypeptide receptor (hNpr1) antagonism[41] was published in July 2019 and therefore could have not been included in ChEMBL version 25 used here, which was released in March 2019. Using HTS, the authors identified potent hNPR1 inhibitors. One of these compounds (JS-11) was further tested in vivo in mouse, causing a decrease of the behavioral response. Interestingly, this molecule contains one of the ZINC-specific substructures identified earlier, MCS12. Figure 8 shows examples of MCSs
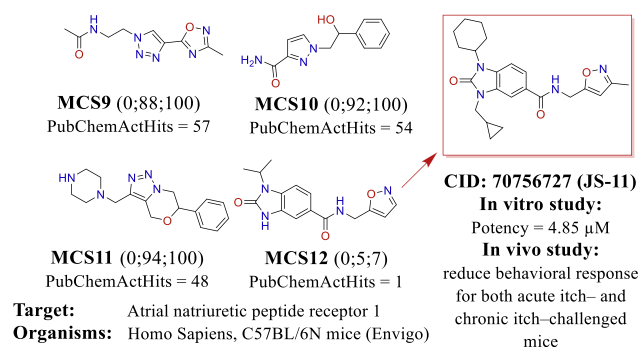
Figure 8. Examples of the ZINC-specific MCSs, generalizing compound classes, tested in hNpr1 antagonism studies. Compound on the right (JS-11) has been ranked as the best inhibitor and was tested in an in vivo model, showing a decline in the behavioral response for itch-challenged mice.

that were found in the active PubChem subset, including MCS12 and the corresponding compound JS-11. These examples prove that previously unexplored regions of chemical space may contain hidden treasures—potential drug candidates or at least starting points for their design.

Remaining 13 891 ZINC-specific MCSs absent from PubChem were considered as overlooked by medicinal chemists and, thus, suggested as a guide for the more efficient exploration of the purchasable chemical space. To assess their potential biological activity, 149k lead-like ZINC-Real compounds incarnating those MCSs were profiled against 749 ChEMBL biological targets, using the in-house GTM-based Profiler.[25] These results are not intended to represent any specific "virtual screening campaign" pending experimental validation but are shown as an illustration of the power of this multifaceted tool—both a chemical space map and an activity profile predictor, at the same time. Their accuracy is, of course, essential, but that issue was already addressed in many other publications, both benchmarking studies[33] and prospective virtual screens.[42,43] The conclusion is that they are slightly less accurate than machine-learned models but acceptable because unlike the former "black box" models they are visual and intuitive.

As a result, 41k compounds (around 30% of the virtually screened molecules) were marked as potentially active against 525 ChEMBL biological targets. Half of the hits (Table 2) were predicted to be active only against a single target, another 21% against two targets, and remaining compounds are

Table 2. Target Specification of the Profiled Compounds

type of target	number of targets	number of predicted actives
receptors	181	25 395
enzymes	148	30 300
kinases	108	5 860
other targets	88	14 453

predicted to be highly promiscuous (cumulating up to 18 activities). The MCSs with the highest number of compounds predicted as actives are shown in Figure 9.
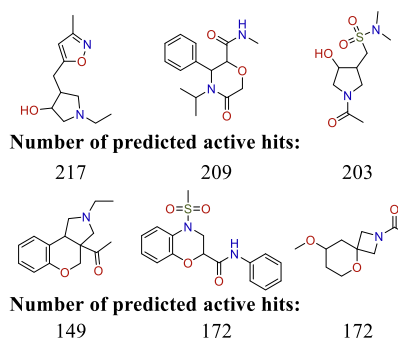

Figure 9. ZINC-Specific PubChem absent MCSs that had the higher number of corresponding compounds, predicted as actives using GTM-based Profiler.

■ CONCLUSIONS

This HGTM analysis of the chemical space has provided a better understanding of the structural features of the purchasable chemical space. For the first time, all commercially available compounds have been taken into consideration, focusing on the detection of specific "open-ended" chemotypes (by contrast to scaffolds, maximum common substructures can be more specific by containing side-chain substituents). It was shown that the chemical market is highly unbalanced, with a bias toward sulfonamides, amides, etc. Comparison of the main features of the in-stock and tangible compounds distribution demonstrated that tangible libraries still sample the same areas of the chemical space that were already overrepresented by in-stock molecules. Thus, there is a need for novel strategies of commercial library enhancement, which can provide a uniform chemical space sampling, avoiding the synthesis of a large number of close analogs. It goes without doubt that chemoinformatics and machine learning methods will be of paramount importance for the development of such strategies in the future.

At the same time, the biological relevance of the purchasable chemical space was assessed in this work. On the one hand, it was discovered that a lot of compound families, known to include biologically active members, are absent from the in-stock catalogs of chemical suppliers. Some of them can be conveniently found in the tangible libraries, the most straightforward source of compounds for the in-stock enhancement campaign, while others are completely unavailable. In both cases, those substructures represent a potential source of inspiration for synthetic chemistry in search of enriching the commercial compound portfolio. On the other hand, the high number of ZINC-specific substructures demonstrates the limited extent of the biological exploration of purchasable libraries. Tens of thousands of such chemotypes encountered

in neither ChEMBL nor PubChem can be used as a "novelty" guide for the further screening campaigns. More than 40 000 HGTMs generated in this work can be used in future investigations of chemical space of any other library.

Finding library-specific substructures by comparing a 1.6M- to an 800M-compound library is rendered possible only by means of the combination of the fast, zone-based clustering of compounds on GTMs and hierarchical zooming, allowing to focus on detailed chemical space zones within which the maximum common substructure detection algorithm can be technically applied. A smooth and comprehensive link is herewith established between the universal map, providing a bird's-eye view of the "Big Data" library, and the specific substructures found in the particular chemical space zones.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00936.

> Computational methods details; examples of the unique biologically relevant MCS for the commercially available libraries enhancement; examples of the unique ZINC MCS for the biological exploration of the commercially available chemical space; link to the complete list of unique MCS for Fragment-Like, Lead-Like, Drug-Like and PPI-Like subsets - https://forms.gle/B6bUJj82t9EfmttV6 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Alexandre Varnek − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* ⓞ orcid.org/0000-0003-1886-925X; Email: varnek@unistra.fr

Authors

Yuliana Zabolotna − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Arkadii Lin − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France*

Dragos Horvath − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* ⓞ orcid.org/0000-0003-0173-5714

Gilles Marcou − *University of Strasbourg, Laboratoire de Chemoinformatique, Strasbourg 67081, France;* ⓞ orcid.org/0000-0003-1676-6708

Dmitriy M. Volochnyuk − *Institute of Organic Chemistry National Academy of Sciences of Ukraine, Kyiv 02660, Ukraine; Enamine Ltd., Kyiv 02094, Ukraine*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.0c00936

Notes

The authors declare no competing financial interest.
The ISIDA/GTM software used in this work is available from the authors upon the request.

■ ABBREVIATIONS USED

HTS, high-throughput screening; MCS, maximum common substructure; GTM, generative topographic mapping; HGTM, hierarchical generative topographic mapping; QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; PPI, protein−protein interaction; ACC2, acetyl-CoA carboxylase 2; PAINS, pan-assay interference compounds

■ REFERENCES

(1) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(2) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(3) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(4) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116−1124.

(5) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.

(6) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and Lead-likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470−479.

(7) Lucas, X.; Grüning, B. A.; Bleher, S.; Günther, S. The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, *55*, 915−924.

(8) Petrova, T.; Chuprina, A.; Parkesh, R.; Pushechnikov, A. Structural enrichment of HTS compounds from available commercial libraries. *MedChemComm* **2012**, *3*, 571−579.

(9) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K.-C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29*, 55−67.

(10) Verheij, H. J. Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Divers.* **2006**, *10*, 377−388.

(11) Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of commercially available compounds for HTS. *Drug Discovery Today* **2019**, *24*, 390−402.

(12) Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T. Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *Chem. Soc. Rev.* **2017**, *9*, 25.

(13) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(14) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59*, 4062−4076.

(15) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162−1171.

(16) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442−446.

(17) Beutler, J. A. Natural Products as a Foundation for Drug Discovery. *Curr. Protoc. Pharmacol.* **2019**, *86*, No. e67.

(18) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(19) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817−834.

(20) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(21) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and structural lessons from recent successes in protein−protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011**, *15*, 475−481.

(22) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.

(23) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13*, 540−554.

(24) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.

(25) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J. Chem. Inf. Model.* **2019**, *59*, 564−572.

(26) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(27) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59−69.

(28) Cao, Y.; Jiang, T.; Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **2008**, *24*, i366−i374.

(29) Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. Diversifying chemical libraries with generative topographic mapping. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 805−815.

(30) Tino, P.; Nabney, I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE Trans.-Pattern Anal. Mach. Intell.* **2002**, *24*, 639−656.

(31) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34*, 348−356.

(32) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301−312.

(33) Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-task generative topographic mapping in virtual screening. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 331−343.

(34) Hann, M. M.; Leach, A. R.; Green, D. V. S. Computational Chemistry, Molecular Complexity and Screening Set Design. In *Chemoinformatics in Drug Discovery*; Wiley-VCH Verlag GmbH & Co. KGaA, 2005; pp 43−57.

(35) Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **2020**, *11*, No. 10.

(36) Reymond, J.-L.; Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.

(37) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

(38) Chast, F. Chapter 1 - A History of Drug Discovery: From first steps of chemistry to achievements in molecular pharmacology. In *The Practice of Medicinal Chemistry*, 3rd ed.;Wermuth, C. G., Ed.; Academic Press: New York, 2008; pp 1−62.

(39) Jha, K. K.; Kumar, S.; Tomer, I.; Mishra, R. Thiophene: the molecule of diverse medicinal importance. *J. Pharm. Res.* **2012**, 560−566.

(40) Grygorenko, O. O.; Volochnyuk, D. M.; Ryabukhin, S. V.; Judd, D. B. The Symbiotic Relationship Between Drug Discovery and Organic Chemistry. *Chem. - Eur. J.* **2020**, *26*, 1196−1237.

(41) Solinski, H. J.; Dranchak, P.; Oliphant, E.; Gu, X.; Earnest, T. W.; Braisted, J.; Inglese, J.; Hoon, M. A. Inhibition of natriuretic peptide receptor 1 reduces itch in mice. *Sci. Transl. Med.* **2019**, *11*, No. eaav5464.

(42) Orlov, A. A.; Khvatov, E. V.; Koruchekov, A. A.; Nikitina, A. A.; Zolotareva, A. D.; Eletskaya, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Horvath, D.; Osolodkin, D. I.; Varnek, A. Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol. Inf.* **2019**, *38*, No. 1800166.

(43) Casciuc, I.; Horvath, D.; Gryniukova, A.; Tolmachova, K. A.; Vasylchenko, O. V.; Borysko, P.; Moroz, Y. S.; Bajorath, J.; Varnek, A. Pros and cons of virtual screening based on public "Big Data": In silico mining for new bromodomain inhibitors. *Eur. J. Med. Chem.* **2019**, *165*, 258−272.

Summary

For the first time, structural analysis of all purchasable compounds, represented by 800M unique structures from the ZINC15 database, was performed, followed by their comparison with the 1,6M biologically relevant molecules from the ChEMBL(v.25) database. It was the first study featuring detailed structural analysis and comparison of the ultra-large compound libraries. The usage of hGTM enabled a 40-fold increase in the size of analyzed libraries compared to previously published reports (800M against 20M analyzed in the work of Lin et al.[13]). Detailed analysis of the chemical space at such a scale provided a better understanding of the structural features of the purchasable chemical space and its biological relevance.

It was shown that the chemical market is highly unbalanced with a shift towards sulfonamides, amides, etc. Since tangible libraries have been designed rather recently as an attempt to enrich existing catalogs with high-quality, diverse screening compounds, in theory, tangible compound distribution should be more balanced. However, a comparison of the density landscapes of the in-stock and tangible compounds revealed that the latter continue to sample the same areas of the chemical space that were already overpopulated by the former. This observation forces one to question current strategies of commercial library enhancement. Indeed, they may need some improvements in order to ensure a uniform chemical space sampling, avoiding the synthesis of a large number of close analogs.

Performed in this work assessment of biological relevance of the purchasable chemical space was never performed before in such scale. On the one hand, it was found that in-stock commercially available libraries are missing around 20K compound families known to include biologically active members - highly potent inhibitors of important biological targets (**Table 5**). Some of them are already represented in the tangible libraries, the most straightforward source of compounds for the in-stock enhancement campaign, while others are completely unavailable. On the other hand, more than 100K ZINC-specific compound families are awaiting to have their potential assessment in screening research programs (**Table 6**). Such a high number of ZINC-specific substructures demonstrates the limited extent of the biological exploration of purchasable libraries.

Table 5. Examples of the unique biologically relevant maximum common substructures (MCS) for the commercially available libraries enhancement. A total number of detected MCS for the particular subset is provided. Numbers in parenthesis under each MCS identify the number of corresponding compounds containing this MCS in ChEMBL, ZINC-Real, and ZINC-Tangible libraries, respectively.
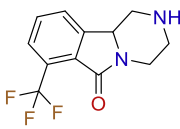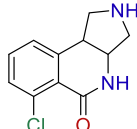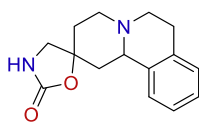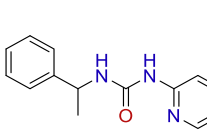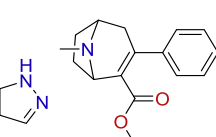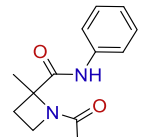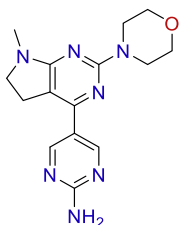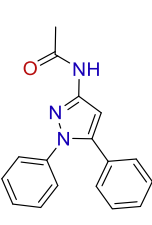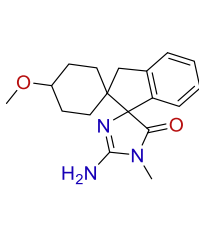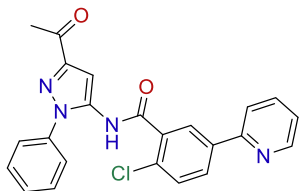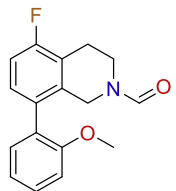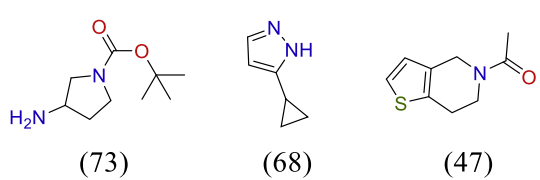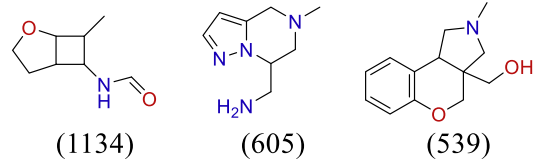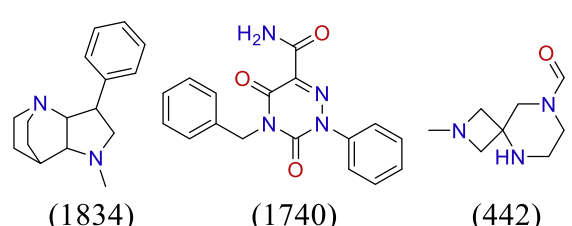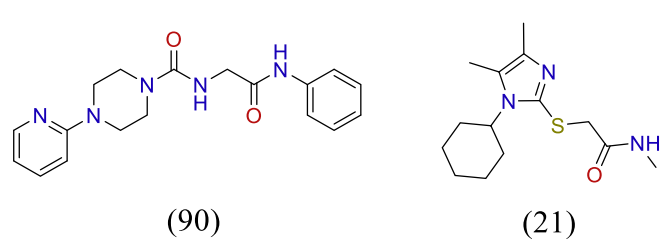
	ChEMBL-specific MCS completely absent on the chemical market
Fragment-Like	Total number of MCS − 38 (19; 0; 0) (8; 0; 0) (5; 0; 0)
Lead-Like	Total number of MCS − 1 966 (415; 0; 0) (73; 0; 0) (50; 0; 0)
Drug-Like	Total number of MCS − 8 239 (295; 0; 0) (271; 0; 0) (230; 0; 0)
PPI-Like	Total number of MCS − 5 969 (625; 0; 0) (326; 0; 0)

Table 6. Examples of the unique ZINC maximum common substructures (MCS) for the biological exploration of the commercially available chemical space. A total number of detected MCS for the particular subset is provided. Number in parenthesis under each MCS identify the number of corresponding compounds containing this MCS in the ZINC-Real library.

	ZINC-Real-specific MCS for chemical space exploration
Fragment Like	Total number of MCS – 1049 (73) (68) (47)
Lead-Like	Total number of MCS – 31 294 (1134) (605) (539)
Drug-Like	Total number of MCS- 67 967 (1834) (1740) (442)
PPI-Like	Total number of MCS- 25 952 (90) (21)

Such an informative analysis of the ultra-large chemical space was only rendered possible by means of the combination of the fast, zone-based clustering of compounds on GTMs and hierarchical zooming, allowing to focus on detailed chemical space zones within which the maximum common substructure detection algorithm can be technically applied. As a result, a smooth and comprehensive link was established between the bird's eye universal map and the specific chemical space zones populated by structurally very similar

compounds. More than 40 000 hGTMs generated in this work can be used in future investigations of chemical space of any other library. Thanks to that, this extensive hierarchy of maps was used as a basis of ChemSpace Atlas in its chapters concerning conventional screening libraries.

4.3 DNA-encoded libraries

Introduction

Apart from classical well-studied techniques of hit identification, like HTS or fragment-based lead discovery[7], several new methodologies have become available recently. One of the most promising among them is affinity selection with DNA-Encoded Libraries (DEL)[15]. Although it was proposed by Brenner and Lerner in 1992, DEL became actively developed only in the 2000s when a squall of papers from researchers all around the world discussing new methods of creating, screening, and evaluating DELs appeared. The advancements in DEL synthesis and screening allowed it to emerge as an efficient option for hit compounds identification.

DEL technology of hit identification comprises three main stages:

- water-based combinatorial synthesis of ultra-large libraries containing up to billions of molecules labeled with single or double-stranded DNA (ususally using split-and-pool method[122]);
- their screening against soluble target proteins using binding affinity selection;
- identification of strongest binders by their DNA tags (using amplification and sequencing techniques[123]).

In such libraries, DNA plays a role of a "barcode" that encodes information about the BBs composing each compound. This DNA barcode allows

Main terminology

DNA-encoded libraries (DEL) technology consists in i) the synthesis of ultra-large libraries of DNA-encoded compounds using water-based combinatorial chemistry; ii) their screening against soluble target proteins using binding affinity selection with iii) further identification of the hits by sequencing the DNA barcode.

Split-and-pool synthesis – a step-wise method in combinatorial chemistry realized in repetitive cycles: i) "splitting" the mixture into several parts, ii) coupling different BB to each portion; iii) pooling and mixing the portions.

DNA sequencing is the process of determining the nucleic acid sequence – the order of nucleotides in DNA.

DNA amplification – a process of producing multiple copies of a specific DNA sequence.

Pool of DELs – complex mixture of multiple DELs synthesized separately but screened together all at once.

to easily identify successful ligands competitively bound to the protein during affinity selection.

This young technology offers many advantages to drug discovery compared to the conventional HTS approach. First of all, chemically versatile library of enormous size can be screened all at once in a single vessel in contrast to individual compound screening in HTS. Moreover, a simple experimental setup of affinity selection accessible both to academic laboratories and small startups allows cheap and fast hits identification. Many success stories of employing this technology in drug discovery have been published, involving the ones when the DEL-derived hits progressed to the clinic.

Even though it gains more and more popularity each day, there are almost no reports of chemoinformatics analysis of DEL chemical space. Therefore, our efforts were directed towards the detailed analysis of the compounds that can be produced via DEL technology. For that, thousands of possible DELs were designed using commercially available BBs and recently reported by Martin et al.[14] freely available tool for multimillion DELs generation, called eDesigner. For each DEL, 1 million representative set was generated. The resulted multibillion DEL chemical space was subjected to GTM-based comparison with the reference library (ChEMBL v.28), representing the chemical space of biologically relevant compounds. The main goal of such comparison is to identify a so-called "golden" DEL or a set(s) of DELs that would cover the chemical space of biologically tested compounds to the highest extent. Such libraries would be particularly useful for the primary screening against novel biological targets.

Exploration of the chemical space of DNA-encoded libraries

Yuliana Zabolotna[1], Regina Pikalyova[1], Dmitriy M.Volochnyuk[3,4], Dragos Horvath[1], Gilles Marcou[1], Alexandre Varnek[1,2]*

Abstract: DNA-Encoded Library (DEL) technology has emerged as an alternative method for bioactive molecule discovery in medicinal chemistry. It enables simple synthesis and screening of compound libraries of enormous size. Even though it gains more and more popularity each day, there are almost no reports of chemoinformatics analysis of DEL chemical space. Therefore, in this project we aimed to generate and analyze the ultra-large chemical space of DEL. Around 2500 DELs were designed using commercially available BBs resulting in 2,5B DEL compounds that were compared to biologically relevant compounds from ChEMBL using Generative Topographic Mapping. This allowed to choose several optimal DELs covering the chemical space of ChEMBL to the highest extent and thus containing the maximum possible percentage of biologically relevant chemotypes. Different combinations of DELs were also analyzed to identify a set of mutually complementary libraries allowing to attain even higher coverage of ChEMBL than it is possible with one single DEL.

Keywords: DNA-encoded libraries, libraries design and comparison, GTM**,** drug design**,** hit identification

INTRODUCTION

Identifying compounds that bind to a biomacromolecule and show a desired therapeutic effect is a fundamental step in any drug discovery process. The most common method to find such molecules is high throughput screening (HTS)[1, 2]. Since its emergence in the 1990s, HTS has delivered numerous lead molecules for drug development[3]. Nevertheless, this technology has several limitations, such as expensive robotic equipment and compound libraries, that are available mostly to large pharmaceutical companies[4]. The number of compounds that can be screened in one HTS campaign is usually limited to a million[5], while the chemical space of synthetically accessible molecules is far larger[6].

DNA-encoded library (DEL) technology has partially solved these problems[7]. It consists of the creation of ultra-large libraries of DNA-encoded compounds using water-based combinatorial chemistry and their screening against soluble target proteins using binding affinity selection[8]. DNA-encoded compounds are molecules labeled with single or double-stranded DNA. The latter plays a role of a "barcode" that encodes information about the building blocks (BBs) from which the compounds were synthesized. This DNA barcode allows to quickly identify successful ligands bound to the protein after affinity selection. The creation and screening of DELs offer many advantages compared to the conventional HTS approach. First of all, they are usually synthesized using a combinatorial split-

1. *University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France) *e-mail: varnek@unistra.fr*

2. *Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan*

3. *Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine*

4. *Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine*

and-pool approach[9] and thus allow to produce chemically versatile libraries of enormous size[10, 11]. DEL compounds are screened all at once in a single vessel in contrast to individual compound screening in HTS[8]. Simple experimental setup of affinity selection accessible both in industry and university laboratories allows cheap and fast hits identification.[12] Many successful stories of employing this technology were published, including DEL-derived hits that progressed to clinic[9].

However, up to this point, most efforts were focused on the analysis of the libraries of BBs or identified active compounds[4]. Authors were less keen to explore the entire chemical space covered by DELs because it is extremely vast. To our best knowledge, only one paper reported the analysis of DEL space using Reduced Complexity Molecular Frameworks (RCMF) methodology[13]. However, in that work, the analysis was limited to only four DELs (>5 × 10^8 compounds). Since DEL technology is actively being developed and new methodologies for DEL synthesis were being elaborated, the aforementioned pioneering work no longer reflects the status quo.

This work is focused on the generation of possible DELs from commercially available BBs using a tool for DELs generation called eDesigner[14]. Since screening thousands of DELs containing billions of compounds is unfeasible, we suggest choosing the so-called "golden" DEL(s) that covers the chemical space of biologically tested compounds to the highest extent. Such a library would have high structural diversity and contain the majority of biologically relevant chemotypes, which is critical for the success of the primary screening against novel biological targets. It was identified by comparing the generated DEL space to the chemical space of biologically relevant ChEMBL[15] compounds using Generative Topographic Mapping (GTM) – a very efficient dimensionality reduction method[16]. GTM has proved to be a powerful tool for "Big Data" analysis and visualization (up to 1B compounds)[17]. Notably, the prior development of quantitatively validated, polypharmacologically competent Universal Maps (uMaps) allowed us to propose a chemically meaningful representation of the to-date explored drug-like chemical space.[18] Only

one of the several uMaps (uMap1, see corresponding article) has been used in this study for simplicity, but the study could be extended to consensus mapping on several uMaps.

METHODS

General workflow

The workflow consists of seven parts, as shown in **Figure 1.** First, DEL-compatible chemical building blocks (BBs) were selected from the eMolecules and Enamine in-stock BB libraries described in the Data section. It was done on the basis of the Goldberg rule of two (Ro2)[19] and eDesigner built-in filters for selecting DNA-compatible BBs. Using these BBs, thousands of DELs were designed and generated with the help of eDESIGNER. The size of each DEL varied from 1M to 1B, but for easier and quicker analysis, only a representative subset of 1M compounds per DEL was enumerated using the random sampling approach. In the third step, generated compounds were standardized according to the protocol explained in the Data section. ISIDA descriptors[20] were used to represent molecular structures in a machine-readable form of numerical N-dimensional vectors. They were then projected onto uMap1. Comparative landscapes were created and visualized to compare DEL compounds to biologically relevant molecules from the ChEMBL database. Then a so-called "golden" DEL that provides the highest coverage of ChEMBL chemical space was identified using responsibility patterns (RPs)[21]. To achieve even better coverage, complementary DELs were added to the "golden" one to give a "platinum" pool of DELs.

BBs selection

Before DEL design and generation, input BBs were filtered according to Ro2 with the help of SynthI[22]. Ro2 is a guideline to choose high-quality BBs that can give access to drug-like molecules[19]. According to it, BBs should contribute to the final molecule only structural fragments that satisfy the following rules: MW<200 Da, clogP<2, number of H-bond donors <=2, and number of H-bond acceptors <=4. This filtration allows to limit the size of DEL compounds shifting corresponding

libraries towards drug-like subspace of the chemical space. In addition to physicochemical properties, eDesigner built-in DNA-compatibility filters were also applied. The selection of building

blocks by eDesigner is made by excluding compounds with unwanted functionalities that can lead to the reaction with water such as imines, benzyl halides, etc.
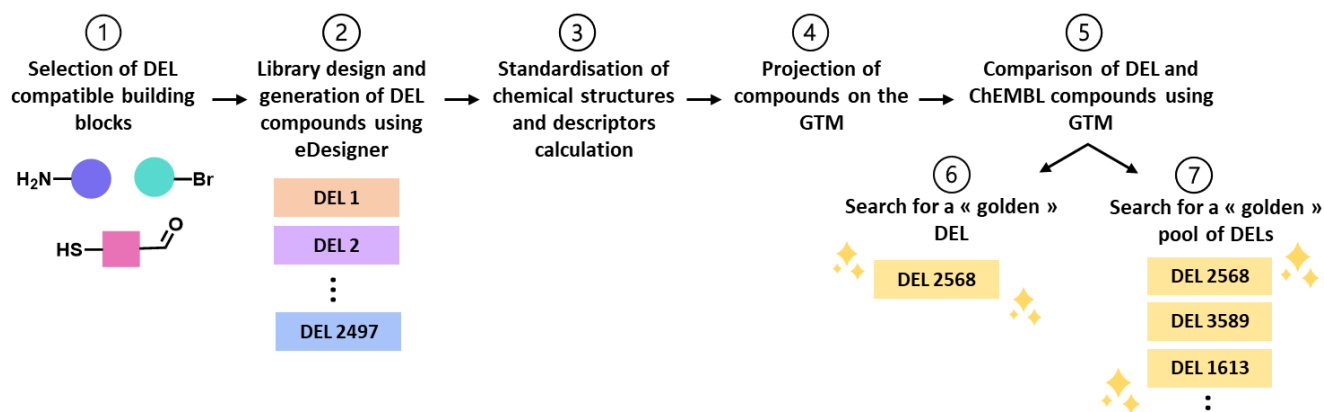

Figure 1. Workflow of the project. The rectangles represent separate DNA encoded libraries (DELs).

DEL generation with eDesigner

For the generation of chemical space of DELs, the eDESIGNER[14] tool was used. At first, based on the list of the most efficient DNA-compatible reactions encoded in the tool (see Supporting Information of respective article[14]) and a user-provided list of BBs, it generates a special set of instructions for DEL compound enumeration called libDESIGNs. Each libDESIGN contains information about the starting headpiece (the whole DNA part for computational convenience is formally represented as a ^{13}C atom), the reaction types, and BBs which will be used in them, as well as deprotection reactions for the final stage of DEL generation. There are also several restrictions that can be applied to control some of the properties of the resulting DEL. They include, for example, the maximum and the median value of heavy atom count in the generated molecules, minimum library size, etc. Once the libDESIGNs are created, the representative DELs subsets of the selected size can be enumerated by the LillyMol tool.[23] An example of such enumeration is shown in **Figure 2**. The isotopic mark on the carbon atom specifies the place of attachment of the DNA tag. For clarity reasons, before physicochemical properties calculation and GTM analysis, the ^{13}C atom is removed, therewith obtaining the compound that would have been resynthesized off-DNA for validation in case of being selected during a real screening campaign.

Generative Topographic Mapping (GTM)

In the chemical space molecules are represented as data points, with their position being defined by a vector of numerical values called descriptors. The main idea of GTM[16] consists in inserting a flexible hypersurface called manifold into the high-dimensional descriptor space with a subsequent projection of these data points into a 2D latent space grid.

The manifold is defined by a grid of Radial Basis Functions (RBFs, represented by Gaussian functions). It generates a probability distribution and is fitted to maximize the likelihood of the training set. The probability distribution generated by the GTM is evaluated over another grid of predefined locations, termed nodes. The number of RBFs is the key user-defined operational parameters; the number of nodes controls the map's resolution: it impacts the rendering but not the model itself. The GTM algorithm "bends" the manifold to pass through the densest areas of the data cloud formed by the points representing molecules of the input dataset. Then, the molecules are projected from the high-dimensional space onto the 2D map by associating each molecule to the several closest grid nodes. The degrees of association of each molecule to each node of the grid are called "responsibilities". The responsibility of a node for a compound is the contribution of this node to the likelihood of this compound. Therefore responsibilities are real

numbers vectors summing up to 1 over all nodes. Finally, the manifold is flattened out to obtain a 2D

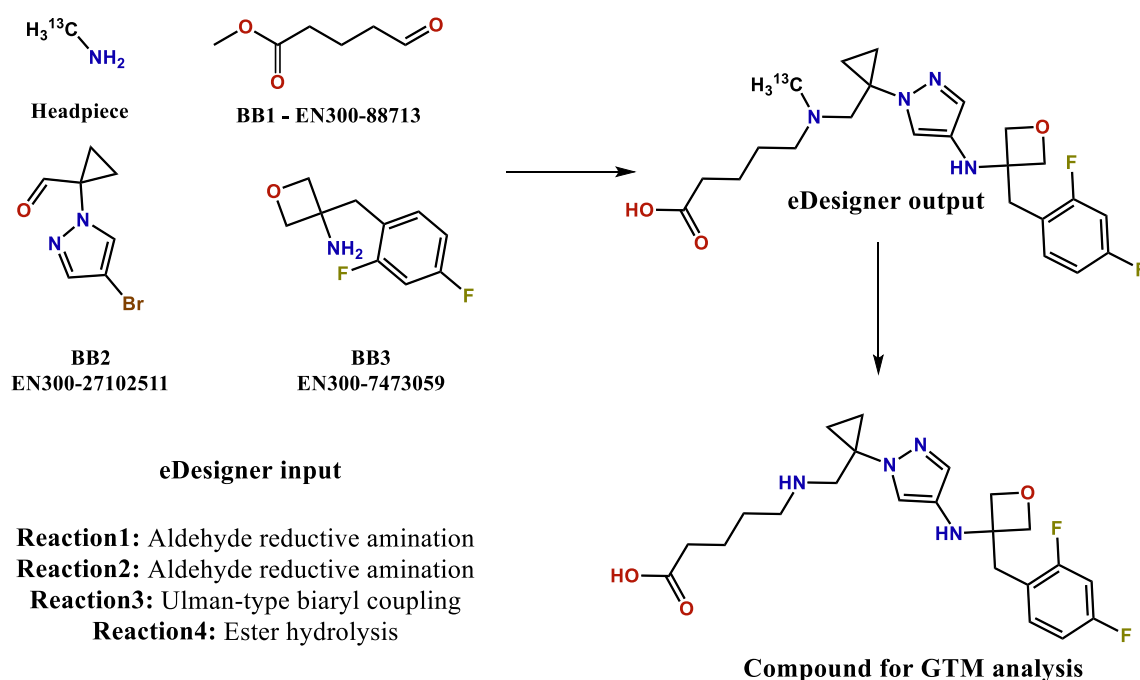representation of the map with compounds projected onto it.

Figure 2. Example of DEL compound generation by eDesigner. The user should provide headpiece and the list of BBs; an appropriate list of reactions will be selected automatically by eDesigner, and respective compounds are generated. The isotopic mark is placed by eDesigner in order to know the position of DNA attachment and is removed prior to GTM analysis and physicochemical properties.

Based on the responsibility vectors, different types of landscapes can be created, where each node is colored using the weighted average of the properties of the compounds projected there. Properties assigned to each node are calculated as a weighted average of the properties of all residents, where weights are compound responsibilities to reside in this node. Depending on the information used for its coloration, there are two types of landscapes: class and property. The class landscape is used to analyze the distribution of the molecules of two classes in the chemical space. In this work, the class landscapes are used to visualize and analyze the distribution of the molecules of two classes – DEL (library1) and ChEMBL (library2) compounds. Property landscapes represent the distribution of molecular property or activity values. Using these landscapes, GTM can be applied for chemical space analysis, library comparison, or even virtual screening[24].

Universal GTM

The concept of Universal GTM (UGTM) was introduced by Sidorov et al.[25] and further developed by Casciuc et al.[18] as a general-purpose map that can accommodate ligands of diverse biological targets on the same GTM manifold. A genetic algorithm was used to choose the best descriptors set and GTM operational parameters (number of nodes and RBFs, manifold flexibility controls, etc.) so as to maximize the mean predictive performance over hundreds of biological activities from ChEMBL. The resulting best uMap1 allowed to separate molecules by their activity class (active/inactive) against 618 (later extended to 749) biological targets, which makes it "polypharmacologically competent". This map was built based on ISIDA atom sequence counts with a length of $2-3$ atoms labeled by CVFF force field types and formal charge status[20]. The size of the map was chosen to be 41x41 nodes and the number of RBFs - 18x18.

Since the ChEMBL database is the most reliable source of the compounds with experimentally measured biological activity[15], the universal maps trained on the ChEMBL data series are highly oriented towards biologically relevant compounds. Apart from predicting biological activity, these maps can also be used as frameworks for analyzing large chemical libraries in medicinal chemistry and drug design context. The uMap1 has been used in this project to compare biologically relevant compounds from ChEMBL with the DNA-encoded compounds. This choice was motivated by previous results in identifying biologically relevant molecules missing from the chemical market, as well as untested commercially available compounds when comparing ChEMBL and ZINC[17].

Responsibility patterns

As mentioned previously, compounds are mapped on the GTM with certain responsibilities - probabilities of these compounds to populate a specific node of the map. Since these values are real numbers, finding two molecules with identical responsibility vectors is highly improbable. This makes it challenging to identify structurally similar compounds by their responsibility vectors – they may be slightly different even for very similar

compounds. To solve this problem, it was suggested by Klimenko et al.[26] to discretize the vector, with all responsibility values less than 0,01 being reassigned to zero and all others - to a number from 1 to 10. This discretized vector is referred to as Responsibility Pattern (RP) and is calculated for each compound according to the formula in **Figure 3.**

Molecules whose R vectors round up to the same RP are considered to be grouped in the same cell of the chemical space and thus to form a cluster of similar structures[24]. For example, in **Figure 3,** a GTM density landscape, featuring compound sets associated with two different RPs is shown. Colors encode the cumulative sum of responsibilities of all compounds residing in the particular node (grey regions are moderately populated, while colored ones contain a higher number of compounds). RP1 corresponds to the 221 indoles that contain additional amino and/or guanidino functional groups. These compounds occupy a small compact area of the chemical space distanced from the island of RP vector 2, populated by 173 naphthols, polyphenols, and their methyl ethers. In this work, RPs were used to compare each separate DEL with ChEMBL, i.e. to evaluate the proportion of ChEMBL RPs ("structural motifs") also covered by a given DEL.

Left box:

$$RP = [10 \times R + 0,9]$$

Where [] – truncation

R – Responsibility value

RP – Responsibility Pattern

If $0,01 \leq R \leq 0,1 \Rightarrow RP = 1$

If $0,11 \leq R \leq 0,2 \Rightarrow RP = 2$, etc.

Indols with additional amino or guanidino groups

221 compounds
RP vector 1

Naphthols, polyphenols and their methyl ethers
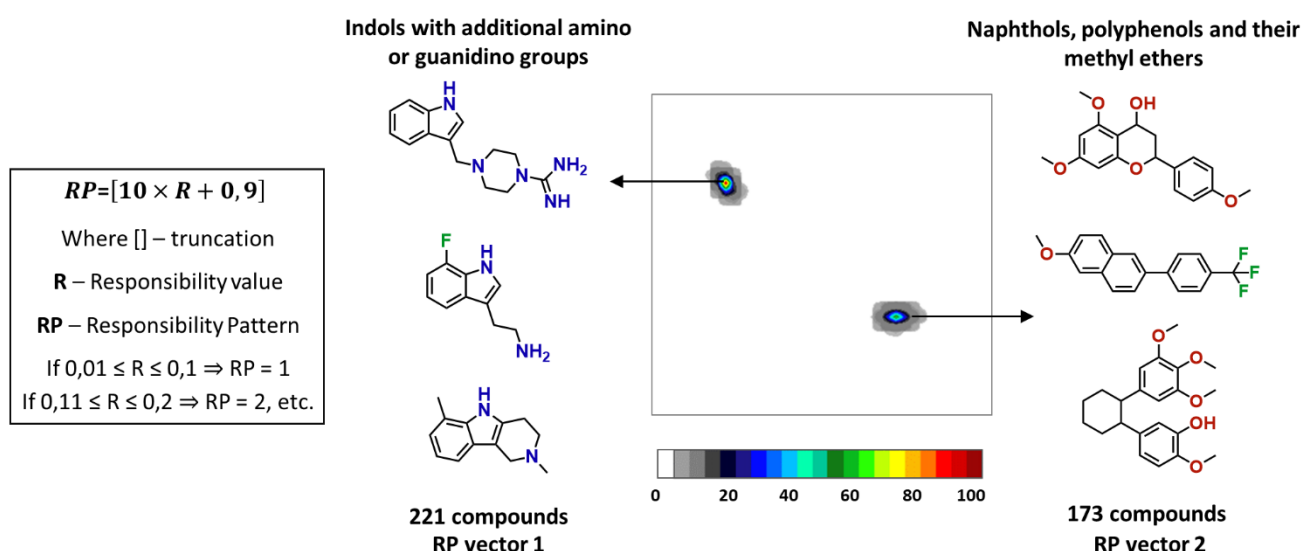
173 compounds
RP vector 2

Figure 3. Left: formula for responsibility pattern (RP) calculation. Right: example of compounds sharing the same RPs and their position on the density landscape - a map colored by local density of compounds. Highly populated zones are colored in red, underpopulated ones - in grey.

ChEMBL coverage estimation

First, RPs for all compounds are calculated as described above. Then the pairwise overlap

between each DEL and ChEMBL is determined by dividing the number of common RPs for both libraries by the total number of ChEMBL RPs:

$$ChEMBL \; RPs \; coverage \; \% = \frac{Number \; of \; ChEMBL \; RPs \; present \; in \; DEL}{Total \; number \; of \; ChEMBL \; RPs}$$

However, the analysis of the percentage of covered ChEMBL RPs does not consider the number of compounds corresponding to each RP, although different RPs can be populated differently – from 1 to ≈12 000 compounds. As a result, increasing RP

coverage does not necessarily mean significantly increasing the compound coverage. Thus the ChEMBL RPs coverage (%), weighted by RP population (the number of ChEMBL compounds per RP), is also used:

$$Weighted \; ChEMBL \; RPs \; coverage \; \% = \frac{\sum Population \; of \; ChEMBL \; RPs \; present \; in \; DEL}{\sum Population \; of \; all \; ChEMBL \; RPs}$$

DATA

Commercially available BBs

A set of 450K commercially available BBs was provided by eMolecules Inc[27]. They were complemented by an "orthogonal" (i.e., containing completely different BBs) dataset of 10K Enamine[28] in-stock BBs. Among them, only 79,141 BBs that satisfy Ro2 and eDesigner build-in DNA-compatibility filters were selected.

ChEMBL (biologically tested compounds)

ChEMBL is a database containing >2M diverse and biologically relevant compounds against >14K biological targets[15]. The major goal of this project was to find structurally diverse DELs suitable for primary screening. Since similar structures tend to have similar properties, finding a DEL containing compounds structurally similar to ChEMBL means finding a DEL that contains biologically relevant molecules. Such DEL will have a high potential to contain hit compounds. Hence, ChEMBL (version 28) was used as a reference library that guides our choice of the best DEL for primary screening. First, 2 086 898 molecules were downloaded from ChEMBL. After standardization, 1,853,565 unique compounds with known biological activities remained. The standardization of chemical structures was done using ChemAxon

Standardizer[29] according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics in the University of Strasbourg.[30]It included dearomatization and final aromatization (heterocycles like pyridone are not aromatized), dealkalization, conversion to canonical SMILES, removal of salts and mixtures, neutralization of all species, except nitrogen(IV), generation of the major tautomer according to ChemAxon. After the standardization, the ISIDA fragment descriptors used to construct the first universal map (described in Experimental section 4) were calculated for all molecules. The same procedure was also applied to generated in this work DEL compounds.

RESULTS AND DISCUSSION

DNA-compatible BBs and reactions for DEL generation

The scope of synthetic procedures used in DEL chemistry is limited to high-yielding DEL compatible reactions. Synthetic efforts to adapt reactions for use in DEL technology have been underway for several years, but the number of optimized for DEL chemistries is still rather restricted[31]. For example, only a few heterocyclisations optimized for DEL synthesis were described, such as benzimidazole,

imidazolidinone, thiazole synthesis, and some others[32]. Nevertheless, even a few reactions can give rise to structurally diverse DELs if abundant

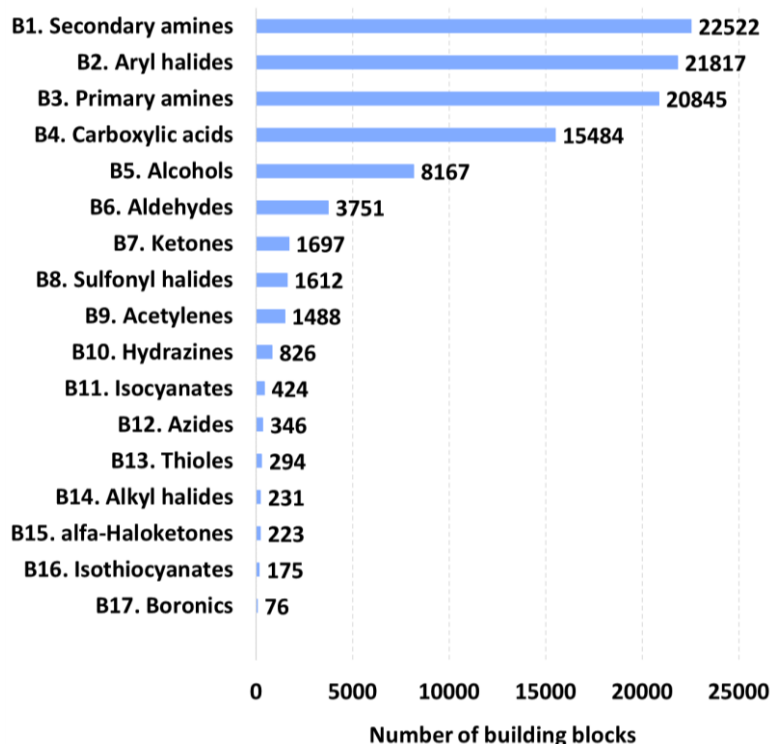building blocks (BBs) sets are employed for their generation.

Figure 4. Monofunctional DNA-compatible commercially available BBs.
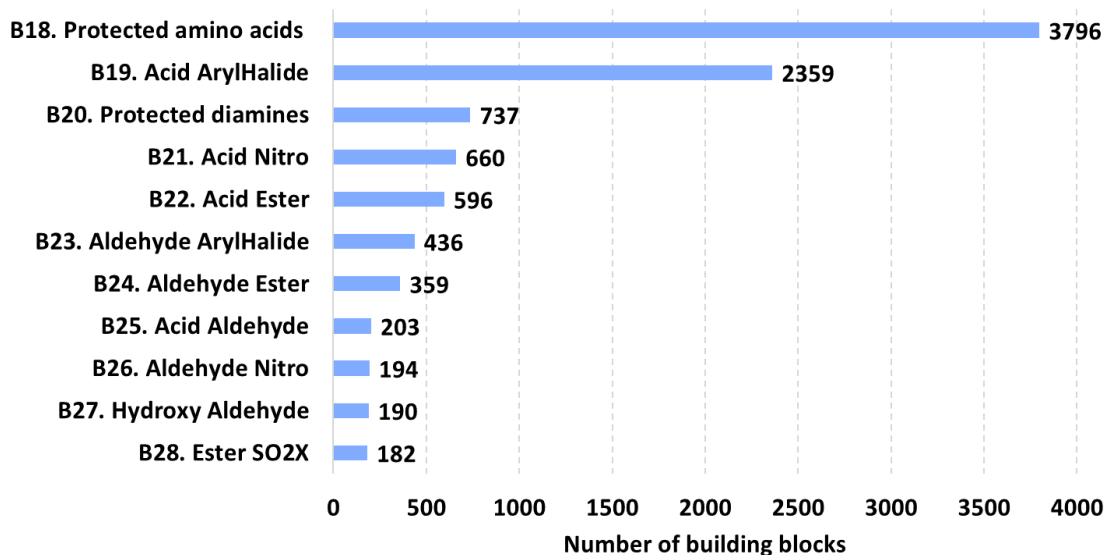

Figure 5. Bifunctional DNA-compatible commercially available BBs.

In this work, 79,141 mono-, bi-, and trifunctional BBs were used for DEL generation. They were obtained by applying the Goldberg rule of two and built-in eDesigner DEL-compatibility filters to the combined in-stock library provided by eMolecules
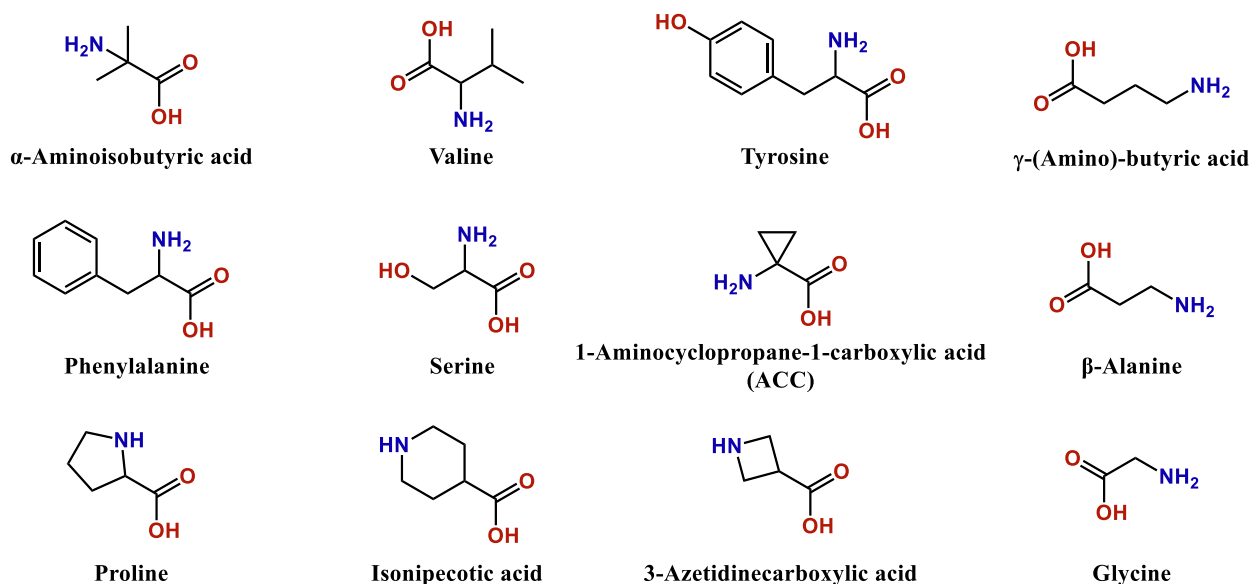
and Enamine. Prevalent monofunctional BB classes in the resulting dataset are secondary and primary amines, aryl halides, and carboxylic acids (**Figure 4**). Due to their participation in common DNA-compatible combinatorial reactions (such as

condensation of carboxylic acids with amines, aldehyde reductive amination, bromo-Sonogashira coupling, etc.), there is an active development of such BBs, making these four classes more structurally rich and widely available commercially. Note that in this work, all structures were stereochemistry-depleted (a unique skeleton graph being used to represent all stereoisomers). Therefore, the number of different BBs is higher.

In the case of bifunctional BBs (**Figure 5**), protected amino acids (AA) (such as amino esters, N-Boc-AA, N-Fmoc-AA, etc.) represent the most abundant class (3,796). The reason for such abundance is the popularity of peptide bond formation for DEL compounds' synthesis that requires this type of reagents. However, the number of actual AA fragments available from BBs with multiple protective groups is slightly smaller (2,885). It appears that the majority of AA fragments (2,173) occur in only one protected form, and only 712 AA were found in the library more than once with different protecting groups. **Figure 6 (I)** shows an example of AAs that occur in the maximum number of protected combinations in the BB library.

Amino acids with the highest number of protecting group variations in the commercially available libraries

α-Aminoisobutyric acid   Valine   Tyrosine   γ-(Amino)-butyric acid

Phenylalanine   Serine   1-Aminocyclopropane-1-carboxylic acid (ACC)   β-Alanine

Proline   Isonipecotic acid   3-Azetidinecarboxylic acid   Glycine

Diamines with the highest number of protecting group variations in the commercially available libraries

Piperidin-4-amine   Azetidin-3-amine   1-(Piperidin-3-yl)methanamine   4-Piperidinemethanamine
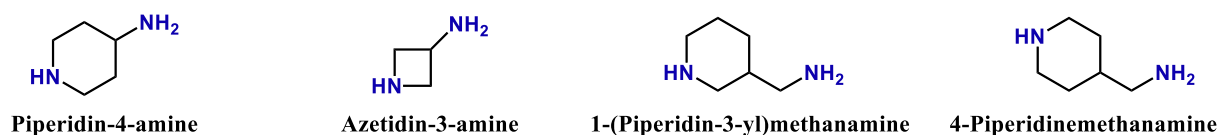
Figure 6. AA (I) and diamines (II), represented in the commercially available libraries of DNA-compatible BBs with the highest number of protected variations (N-Boc, N-Fmoc, various esters etc.)
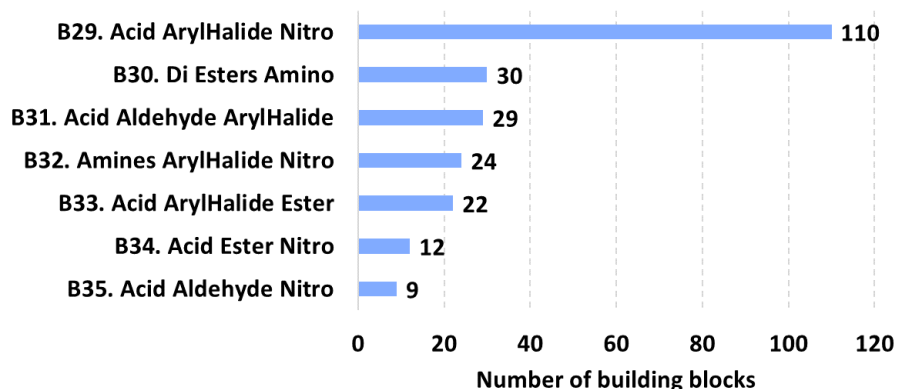
Figure 7 bar chart:

- B29. Acid ArylHalide Nitro — 110
- B30. Di Esters Amino — 30
- B31. Acid Aldehyde ArylHalide — 29
- B32. Amines ArylHalide Nitro — 24
- B33. Acid ArylHalide Ester — 22
- B34. Acid Ester Nitro — 12
- B35. Acid Aldehyde Nitro — 9

X-axis: Number of building blocks

Figure 7. Trifunctional DNA-compatible commercially available BBs.

Figure 8 bar chart:

- R1. Condensation of carboxylic acids with amines — 1611
- R2. Aldehyde reductive amination — 1204
- R3. 1,2,3-Triazole synthesis — 700
- R4. Guanidinylation of amines — 615
- R5. Migita thioether synthesis — 557
- R6. Bromo Sonogashira coupling with TMS-acetylene — 512
- R7. Ullmann-type N-aryl coupling — 451
- R8. SnAr aniline synthesis — 360
- R9. Iodo Sonogashira coupling with TMS-acetylene — 328
- R10. Imidazole synthesis — 280
- R11. SnAr ether synthesis — 246
- R12. Aminothiazole synthesis — 232
- R13. Buchwald Hartwig amination with bromoaryles — 210
- R14. 1,2,3-Oxadiazole synthesis — 189
- R15. Schotten Baumann sulfonamide synthesis — 154
- R16. Buchwald-Hartwig amination with iodoaryles — 148
- R17. Larock indole synthesis — 138
- R18. Isocyanate-amine coupling — 100
- R19. Isothiocyanate-amine coupling — 84
- R20. Alkyne coupling — 56
- R21. Pyrazole synthesis — 56
- R22. Thioether synthesis — 38
- R23. N-arylation with fluoronitrobenzenes — 13
- R24. Benzimidazole synthesis — 10
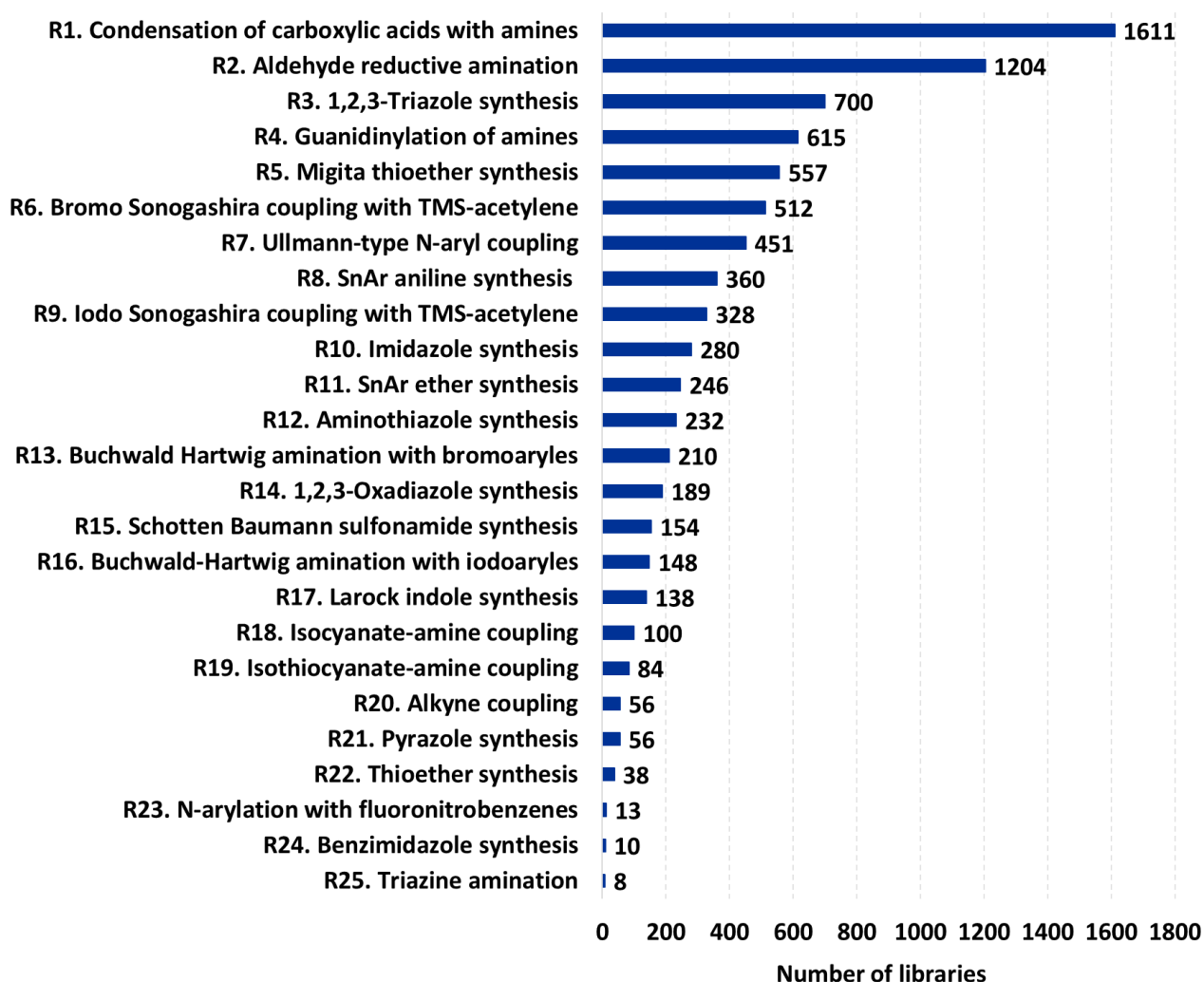- R25. Triazine amination — 8

X-axis: Number of libraries

Figure 8. Frequency of the use of a particular reaction in DELs generation.

A similar tendency is also observed for protected diamines that occupy third place in the bar chart in **Figure 5** after BBs containing both aryl halide and carboxylic functionality (2 359). A total of 737 protected diamines are equivalent to only 632 unique diamine fragments. Among them, 510 are represented by only one protected variant, while the other 122 occur in several differently protected copies. Four diamines, each occurring in the highest observed protected variations, are shown in **Figure 6 (II)**. The number of trifunctional BBs is significantly lower than other reagents due

to higher structural complexity (**Figure 7**). The most highly populated class of trifunctional BBs is haloaryl nitrocarboxylic acids containing 110 members. In DEL technology nitro group usually pose as a latent amino group that can be obtained upon reduction.

Using these BBs and user-defined library limitations in eDesigner, 2,495 DELs were designed. The maximal number of heavy atoms in DEL compounds was set to be 45, and at least half of all compounds in the library needed to have less than 35 non-hydrogen atoms. The frequency of the use of a particular reaction to generate all DELs is shown in **Figure 8**. The most frequently used reactions, each being exploited in more than 500 libraries, were: condensation of carboxylic acids with amines (R1), aldehyde reductive amination (R2), 1,2,3-triazole synthesis (R3), guanidinylation of amines (R4), Migita thioether synthesis (R5), and bromo-Sonogashira coupling with TMS-acetylene (R6). The high frequency of reaction usage is mainly caused by the prevalence of the

respective BB classes in the input library (B1, B2, B3, B4 in **Figure 4**). Indeed, the amines are coupling partners in three reactions mentioned above (R1, R2, and R4), aryl halides - in two (R5 and R6), and carboxylic acids in R1.

Not all compounds were enumerated for every DEL, but random sets of 1M representative compounds were produced by eDesigner. In order to verify that such a library core is indeed representative, the whole library of 88M has been enumerated for one of the DELs, and density landscapes have been built for the whole library and 1M dataset on the same density scale. As one can see in **Figure 9**, each region of the map, occupied by the members of the whole library, also has representatives in the 1M randomly generated dataset – colored regions coincide on both maps, and only the density of residents differs. Therefore, 1M randomly enumerated compounds will be considered in this work as a sufficient representation of large DELs for GTM-based analysis.
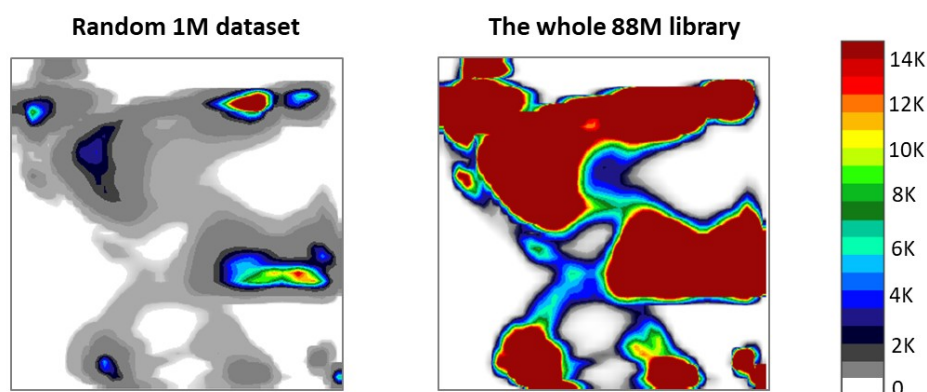
Random 1M dataset   **The whole 88M library**

Figure 9. Comparison of the density distribution for the 1M randomly generated compounds and the whole DEL(88M). The color scale encodes the corresponding number of compounds residing in each colored node of the map.

Physicochemical properties of generated libraries

Out of total 2,495 generated DELs, 77 are produced by a single coupling reaction of 2 BBs (hence the label "2BB libraries"). The remaining 2,418 DELs are "3BB libraries". The physicochemical properties were calculated using RDKit[33]. Drug-like[34] (MW ≤ 500; LogP ≤ 5; the

number of H-bond donors ≤ 5; the number of H-bond acceptors ≤ 10; ring counts ≤ 10) and lead-like[35] (MW ≤ 400; -3.5≤LogP ≤ 4; the number of H-bond donors ≤ 5; the number of H-bond acceptors ≤ 8; ring counts ≤ 4; rotatable bonds≤10) filters were applied. **Figure 10** depicts how many of 2BB and 3BB libraries (in percentage) contain a specified portion of drug-like **(Figure 10 (I))** and lead-like **(Figure 10 (II))** compounds.
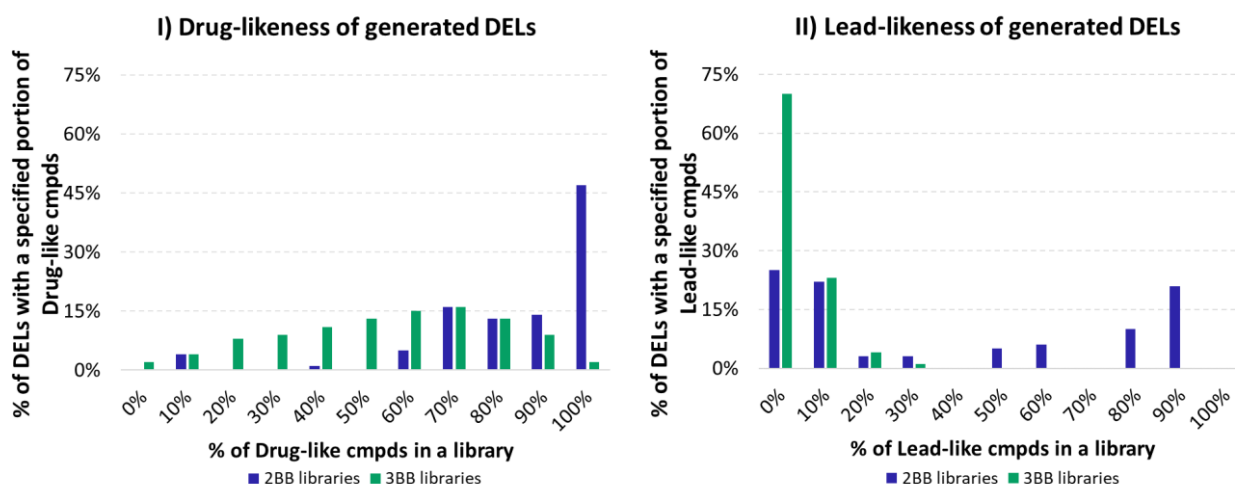
Figure 10. Comparison of (I) drug- and (II) lead-likeness of 2BB and 3BB libraries: percentage of 2BB and 3BB libraries having a particular portion of compounds satisfying respective filters is given.

As expected, 2BB libraries contain smaller compounds, and thus the portion of drug- and lead-like compounds for them is higher than for 3BB DELs. For almost a half of 2BB libraries, all generated compounds fall into the category of drug-like, while in the case of 3BB DELs, only 2% of libraries are fully drug-like. However, the content of such compounds in 3BB libraries is still relatively high – the majority of DELs (68%) contain at least 50% of drug-like compounds. At the same time, the number of lead-like compounds is significantly lower for both categories of DELs. Almost a quarter of all 2BB libraries do not contain them, and another quarter is less than 50% lead-like. In the case of 3BB libraries, the lead-like compounds are almost entirely absent – 70% of DELs do not contain such molecules at all, and the remaining 30% of libraries have only up to 30% of lead-like molecules.

Search for the "golden" DEL

The "golden" DEL can be defined as a library that is diverse enough to cover the highest possible proportion of biologically relevant compounds from ChEMBL. This coverage was calculated in terms of common responsibility patterns (RPs) explained in Methods section. In **Figure 11(a)** one can see the number of libraries with particular coverage of ChEMBL RPs. The majority of libraries cover 10-20% of ChEMBL chemical space in terms of unweighted RPs coverage score.

64 DELs showed the highest coverage of ChEMBL RPs – 30-33%. **Figure 11 (b)** depicts the coverage of the ChEMBL RPs weighted by the number of compounds that correspond to each RP. This time, 90 DELs showed high coverage of ChEMBL chemical space, ranging from 50 to 60%.

Figure 12 displays three comparative landscapes: DEL1857 with 13%, DEL167 with 27%, and DEL2568 with 60% coverage of ChEMBL (here, weighted coverage is considered). Dark grey zones are populated exclusively by ChEMBL molecules, while all other colors indicate areas also containing DEL compounds in a different ratio. Below each landscape, the IDs of reactions used for the corresponding library generation are given (see **Figure 8** for reaction IDs). From the landscape of DEL1857, it is apparent that this library does not cover many areas of ChEMBL chemical space – there are few multicolored spots on the landscape. It is an indicator that DEL1857 is not chemically diverse enough, and there are plenty of biologically relevant chemotypes absent from this library. DEL167, in its turn, allows achieving higher coverage of ChEMBL. However, DEL2568 is the leader among all 2,5K DELs - multicolored areas are not focused in one place of the map, but rather distributed on different islands that correspond to different chemotypes, and dark grey areas are less present.
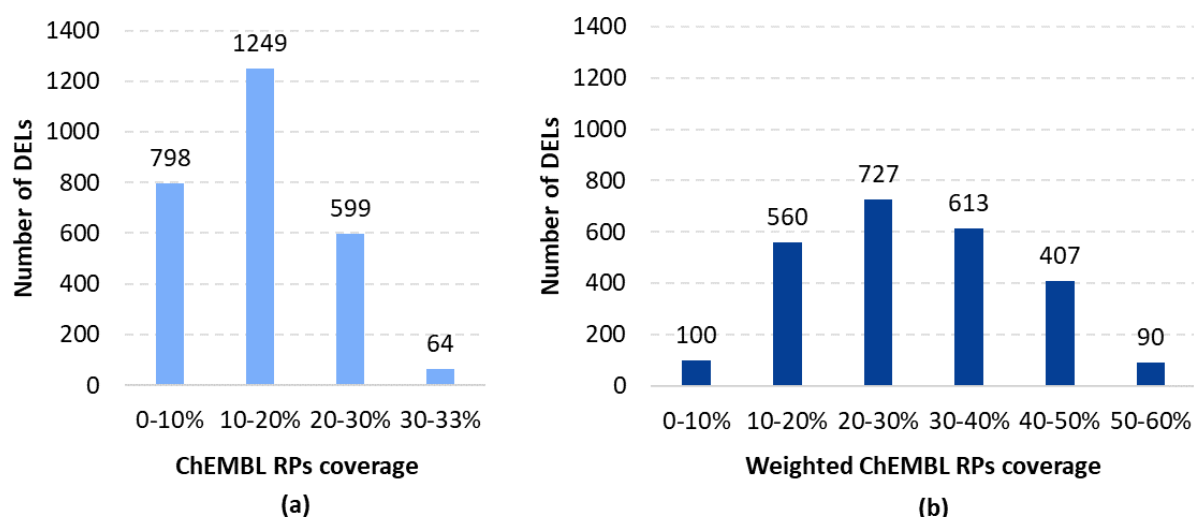
Figure 11. (a) Number of DELs with different coverage of ChEMBL responsibility patterns (RPs) (b) Number of DELs with different percentages of ChEMBL RPs coverage weighted by the RPs population (number of ChEMBL compounds per RP).
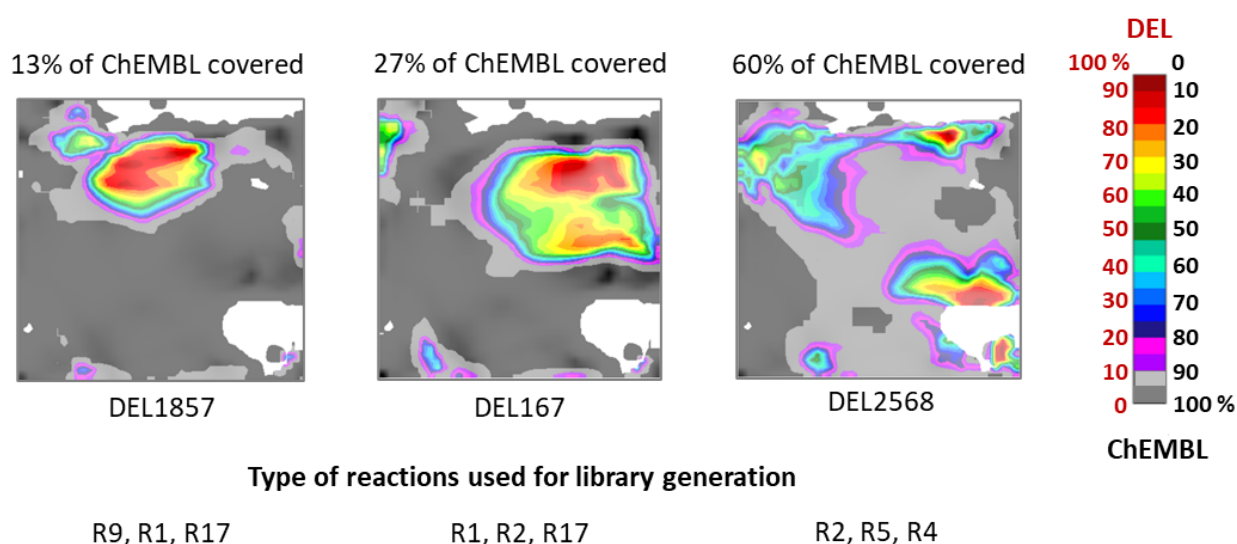

Figure 12. Class landscapes comparing a particular DEL with ChEMBL. From left to right: comparison of ChEMBL to DEL1857, DEL167, and DEL2568. Dark grey zones are populated exclusively by ChEMBL compounds, while all other colors indicate areas also containing DEL compounds in a different ratio. White regions correspond to the empty areas of the chemical space. Below each landscape, , a library ID and IDs for corresponding reaction types are given.

There are around 60 libraries with similar chemical space coverage and diversity, but here, we will limit the discussion to the DEL2568 as an example of a "golden" DEL. 88 Million compounds from this DEL can be obtained by sequentially employing three reactions: aldehyde reductive amination, Migita thioether synthesis, and guanidinylation of amines (see **Figure 14**, DEL2568). BBs used for this DEL design are three aromatic mercaptoaldehydes, 8,914 aryl bromides, and 3,311 amines. As was discussed earlier, the last two are the classes with the highest number of diverse BBs (**Figure 4**). Therefore, a random selection of BBs for DEL generation from such various and numerous collections results in higher coverage of ChEMBL chemical space. DEL2568 was chosen here as an example of a "golden" library because it outruns all

other libraries by 3% of weighted ChEMBL coverage, corresponding to approximately 45K of biologically relevant compounds. However, if the presence of thioether or guanidine groups is not desirable, there is still a diverse choice of DELs that do not contain such moieties.

Search for the "platinum" set of DELs

As shown on the class landscape for DEL2568 in **Figure 11**, there are still some dark-grey zones left that are not covered even by this "golden" DEL, which means there is space for improvement. To fill uncovered parts of the chemical space, the approach of library pools[36, 37] was considered. According to it, several distinct DELs may be further combined to create another more complex mixture, called "library pool", which can then be simultaneously screened. In order to obtain the highest coverage of ChEMBL, composing DELs for constructing such library pools should be complementary to each other, and each new DEL should cover previously unrepresented areas of the biologically relevant space.

To achieve that, first of all, 64 DELs that have the highest coverage of ChEMBL RPs were chosen. Each of these DELs was then iteratively completed with up to 14 other libraries. Every complementary DEL was chosen in a way to cover the maximal portion of the ChEMBL chemical space that was not covered in the previous steps. Each time a complementary DEL was added to the pool, the weighted ChEMBL coverage was calculated. The chart in **Figure 13** was used to identify a pool of DELs that can enhance ChEMBL coverage to the highest possible extent. It shows how the weighted ChEMBL coverage increases over the addition of complementary libraries. According to this chart, after the fifth DEL, each complementary library provides less than 1% of additional weighted ChEMBL coverage. Considering that the size of each DEL can vary from 1M to 1B compounds, adding a library of such large size to the pool only to increase ChEMBL coverage by 1% is not worth it. Therefore, it is irrational to use a pool of DELs composed of more than five libraries.
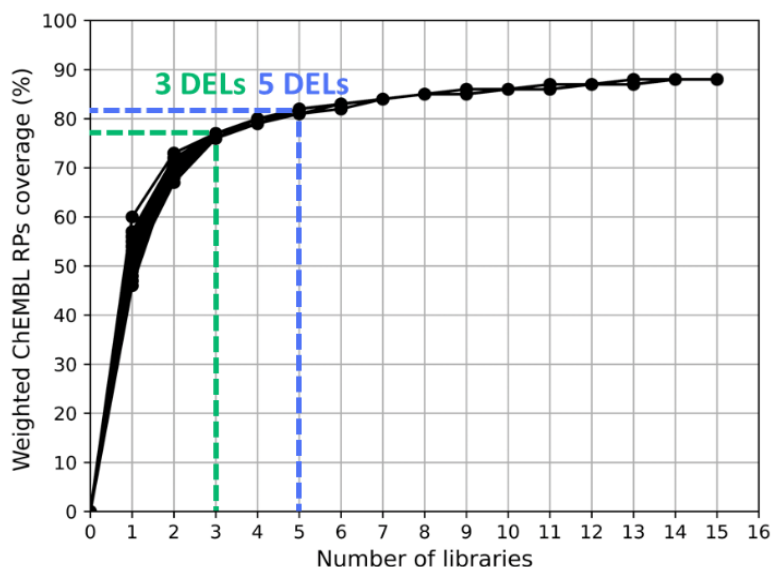

Figure 13. The percentage of the ChEMBL coverage, weighted by the number of compounds sharing common RPs, as a function of the number of libraries in the set. Green and blue dashed lines highlight the points for three and five DELs.
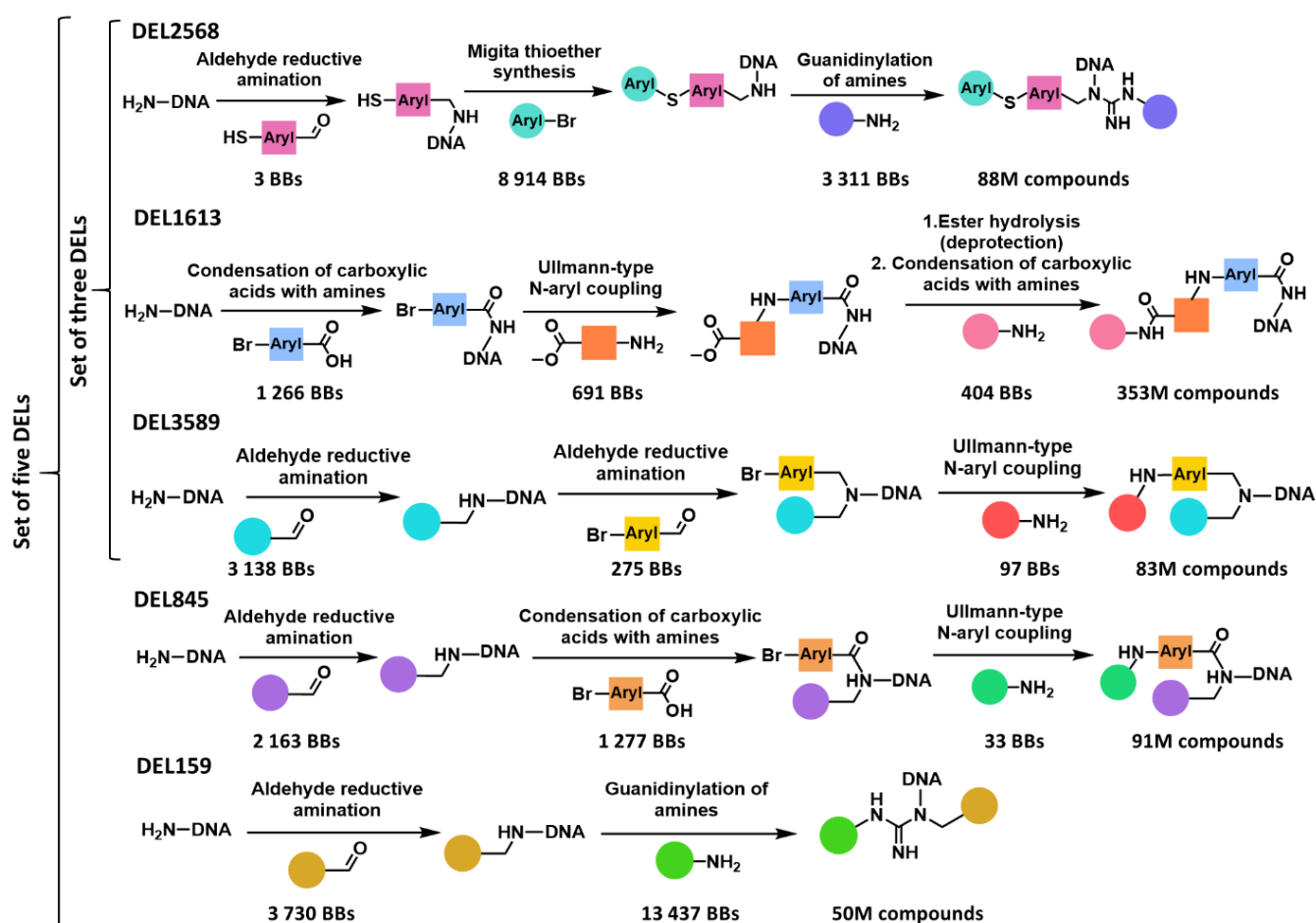
Figure 14. Reactions and BBs required for synthesis of the "golden" DEL and libraries composing "platinum" pools of libraries.

If described above DEL2568 is used as a starting DEL, the "platinum" pool of five DELs will be composed of such libraries: DEL2568, DEL1613, DEL159, DEL845, and DEL3589. Overall, they contain 665M compounds. Reactions used for the generation of these five DELs are shown in **Figure 14:** aldehyde reductive amination (R2), Migita thioether synthesis (R5), Ullmann-type N-aryl coupling (R7), condensation of carboxylic acids with amines (R1), and guanidinylation of amines (R4). All of them are among the most frequently used reactions for DEL generation (**Figure 8**) that employ BBs from highly represented classes **(Figure 4).** On the other hand, a pool of three DELs (DEL2568, DEL1613, DEL3589) can be even more convenient since it contains fewer compounds (524M) and yet still allows to cover a large portion of ChEMBL (78%).

The physicochemical properties of the selected libraries have been calculated and analyzed (**Table 1**). It appears that half of DEL2568 compounds are drug-like, while the portion of lead-like molecules is almost negligible. Complementary DELs forming a "platinum" pools of three and five DELs possess higher drug- and lead-likeness, which influenced the number of corresponding compounds. Indeed, the percentage of drug-like compounds is increasing for the pool of 3 DELs (60.8%) and even more so in the case of 5 DELs (70.4%). Likewise, the portion of lead-like compounds peaks at 21% for the pool of 5 DELs.

To better illustrate how ChEMBL coverage increases when a pool of DELs is used instead of a single DEL, four comparative landscapes – featuring the "golden" DEL, the "platinum" pools of three and five DELs, and ≈2,5K DELs against ChEMBL were created (**Figure 15**). Structural

analysis of underrepresented in DELs zones was carried out (**Figure 16**). The obtained landscapes show that as we go from one (**Figure 15 (I)**) to three DELs (**Figure 15 (II)**), the ChEMBL coverage increases drastically. On the landscape of the "platinum" pool of three DELs, the ChEMBL

areas from A1 to A7 became a lot more populated. However, the addition of the following two libraries does not have the same impact. There are almost no new previously uncovered areas, only the increase in the population of previously occupied areas is observed (**Figure 15 (III)**).

Table 1. The portion of drug-like and lead-like compounds in the selected "golden" DEL and "platinum" pools of three and five DELs.

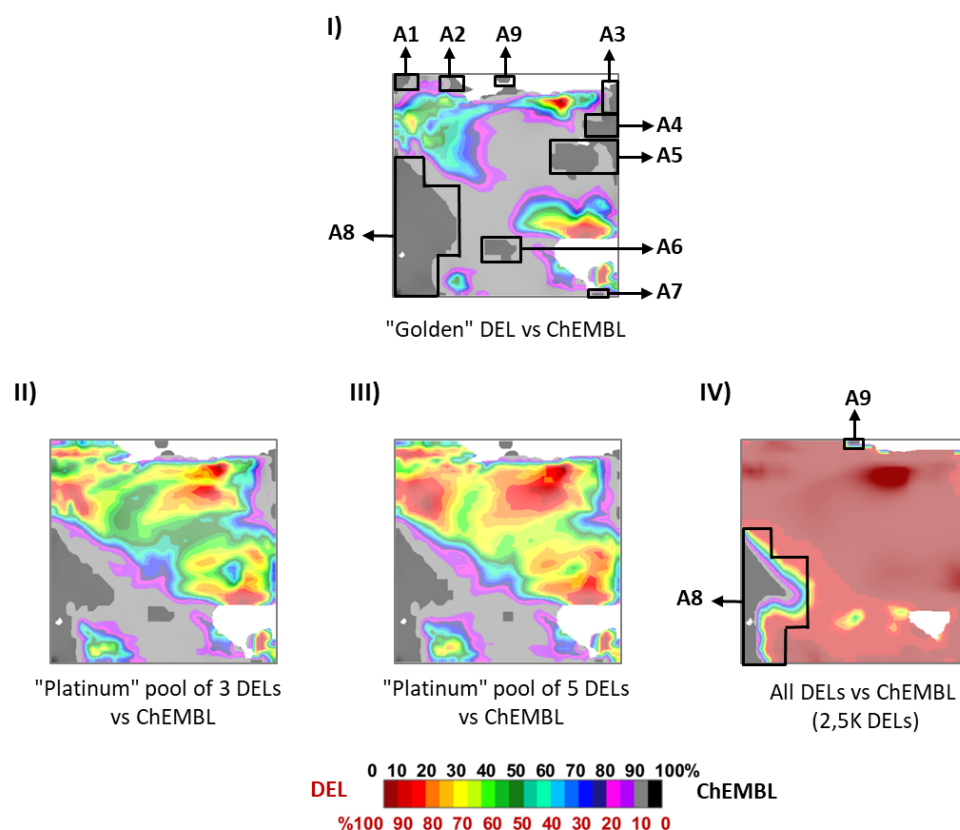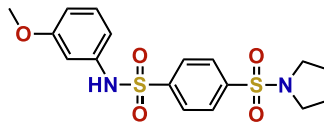|	Portion of drug-like compounds	Portion of lead-like compounds
"Golden" DEL2568	50%	1.5%
"Platinum" pool of 3 DELs	60.8%	6.2%
"Platinum" pool of 5 DELs	70.4%	21.7%

Figure 15. Comparison of ChEMBL and I) "golden" DEL, II) a pool of three DELs, III) a pool of five DELs, and IV) all 2,5K DELs. Multicolored zones are populated by both ChEMBL and DEL compounds, dark grey zones – only by ChEMBL compounds. White regions correspond to the empty areas of the chemical space. Examples of compounds populating highlighted areas A1-A9 are provided in **Figure 16**
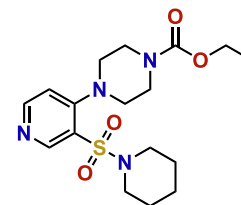
A1: Thiophene-containing compounds
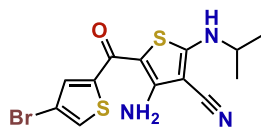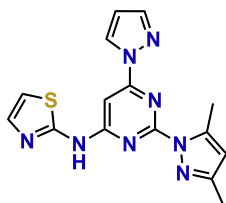
CHEMBL4454199
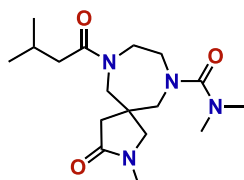
A2: Thiazoles and thiadiazoles

CHEMBL3100167

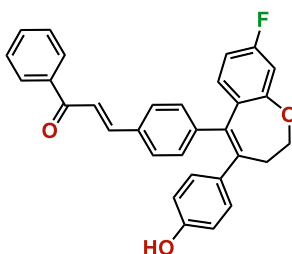**A3: Benzosulfonamides (with two or more PhSO2N groups)**

CHEMBL1729230

A4: Sulfonamides

CHEMBL1346964

A5: Polyamides, ureas, and carbamates
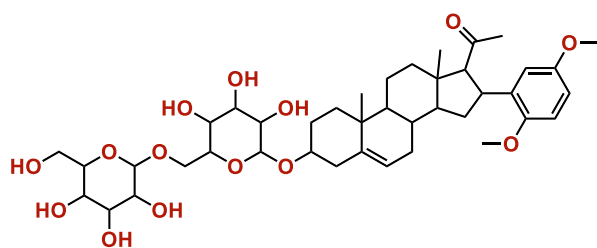
CHEMBL3444791

A6: Aromatic compounds with long conjugated systems
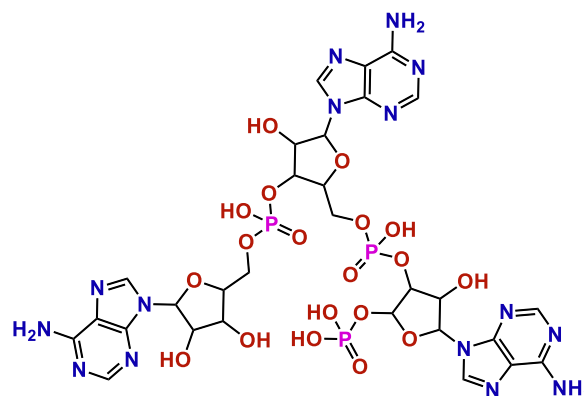
CHEMBL4225431

A7: Dihydropyrazoles and hydrazones with sulfonamide group

CHEMBL1950243

A8: Natural products and NP-like compounds

CHEMBL2096828

A9: Nucleotides

CHEMBL605454

Figure 16. Examples of CHEMBL compounds populating areas from A1 to A9 highlighted in landscapes in **Figure 15**.

However, neither three nor five libraries succeeded in covering areas A8 and A9 completely. To see whether it is even possible to do so, a comparative landscape for all DELs versus ChEMBL was created (**Figure 15 (IV)**). It appears that neither of the DELs can cover these regions of the chemical space – areas A8 and A9 remained dark-grey. This result is not surprising because they contain natural products (NP) and NP-like compounds such as cardiac glycosides, steroids, and steroid-like compounds, saccharides, nucleotides, oligopeptides, coumarins, macrolides, chalcones, etc., which are indeed inaccessible by DEL technology as employed in this analysis.

CONCLUSIONS

In this work, for the first time, the ultra-large chemical space of DNA-encoded libraries (DELs) containing 2,5B compounds in total (2.5K libraries 1M each) was designed and generated using eDesigner and analyzed with the help of GTM. Owing to the probabilistic nature of GTM and efficiency of the libraries analysis and comparison based on the responsibility patterns, it was possible to develop a GTM-based approach for quick selection of DELs occupying the same areas of the chemical space as a reference library. In this work, the goal was to detect the "golden" DEL or "platinum" pool of DELs for primary screening - the libraries containing the highest portion of biologically relevant chemotypes. Therefore, ChEMBL, as the largest database of dose-response activity tests and thus an optimal representation of biologically relevant space, was used as a reference. However, the approach described herein could be applied to any reference library, e.g., actives of a particular biological target.

This approach allowed to identify the so-called "platinum" pools of five and three DELs providing the highest coverage of ChEMBL chemical space – 82% and 78%, respectively. Our results suggest that an optimal set for primary screening is the one encompassing three DELs,

which, even though containing fewer compounds than in five DELs, still succeeds in covering a large portion of ChEMBL chemical space. Analysis of physicochemical properties of the "golden" DEL revealed that half of the compounds are drug-like, and in the case of the pool of 3 DELs, this percentage rises to 60%. The portion of lead-like molecules, however, is negligible.

In this project, only a brief structural analysis of DEL chemical space was performed. Without a doubt, a more detailed GTM-based analysis of chemical structures composing DELs and their comparison to ChEMBL and commercially available HTS libraries will improve our understanding of the chemical space accessible via this technology. Further GTM analysis and comparison of generated DELs can be helpful for the enhancement of available BBs libraries and prioritizing some promising synthetic procedures in order to improve the biological relevance of DEL chemical space.

ACKNOWLEDGEMENTS

REFERENCES

1.   Attene-Ramos, M. S.; Austin, C. P.; Xia, M. High Throughput Screening. In *Encyclopedia of Toxicology*, Wexler, P., Ed.; Academic Press: Oxford, 2014, pp 916-917.
2.   Inglese, J.; Auld, D. S., High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. *Wiley Encyclopedia of Chemical Biology* **2008**, 1-15.
3.   Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S., Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011**, 10, 188-95.

4.   Franzini, R. M.; Randolph, C., Chemical Space of DNA-Encoded Libraries. *J. Med. Chem.* **2016**, 59, 6629-44.

5.   Favalli, N.; Bassi, G.; Scheuermann, J.; Neri, D., DNA-encoded chemical libraries–achievements and remaining challenges. *FEBS Lett.* **2018**, 592, 2168-2180.

6.   Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, 23, 101681.

7.   Brenner, S.; Lerner, R. A., Encoded combinatorial chemistry. *Proc Natl Acad Sci U S A* **1992**, 89, 5381-3.

8.   Goodnow Jr, R. A., *A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery.* John Wiley & Sons: 2014.

9.   Satz, A. L., What do you get from DNA-encoded libraries? *ACS medicinal chemistry letters* **2018**, 9, 408-410.

10.   Franzini, R. M.; Neri, D.; Scheuermann, J., DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. *Acc. Chem. Res.* **2014**, 47, 1247-55.

11.   Madsen, D.; Azevedo, C.; Micco, I.; Petersen, L. K.; Hansen, N. J. V., An overview of DNA-encoded libraries: A versatile tool for drug discovery. *Progress in medicinal chemistry* **2020**, 59, 181-249.

12.   Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S., DNA Encoded Libraries: A Visitor's Guide. *Isr. J. Chem.* **2020**, 60, 268-280.

13.   Kontijevskis, A., Mapping of drug-like chemical universe with reduced complexity molecular frameworks. *J. Chem. Inf. Model.* **2017**, 57, 680-699.

14.   Martín, A.; Nicolaou, C. A.; Toledo, M. A., Navigating the DNA encoded libraries chemical space. *Commun. Chem.* **2020**, 3, 127.

15.   Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M., ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, 47, D930-D940.

16.   Bishop, C. M.; Svensen, M.; Williams, C. K. I., GTM: The generative topographic mapping. *Neural Comput.* **1998**, 10, 215-234.

17.   Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A., Chemography: Searching for Hidden Treasures. *J Chem Inf Model* **2021**, 61, 179-188.

18.   Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A., Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J Chem Inf Model* **2019**, 59, 564-572.

19.   Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W.; Tomkinson, N. P., Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015**, 20, 11-7.

20.   Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, 29, 855-68.

21.   Sidorov, P.; Viira, B.; Davioud-Charvet, E.; Maran, U.; Marcou, G.; Horvath, D.; Varnek, A., QSAR modeling and chemical space analysis of antimalarial compounds. *J. Comput. Aided Mol. Des.* **2017**, 31, 441-451.

22.   Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A., SynthI: a new open-source tool for synthon-based library design *ChemRxiv. Cambridge: Cambridge Open Engage;* **2021**, doi: 10.33774/chemrxiv-2021-v53hl-v2. This content is a preprint and has not been peer-reviewed.

23.   *LillyMol: Eli Lilly Computational Chemistry and Chemoinformatics Group Toolkit. https://github.com/EliLillyCo/LillyMol* 2020.

24.   Horvath, D.; Marcou, G.; Varnek, A., Generative topographic mapping in drug design. *Drug Discov Today Technol* **2019**, 32-33, 99-107.

25.   Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided Mol. Des.* **2015**, 29, 1087-108.

26.   Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A., Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL

Antiviral Compound Set. *J Chem Inf Model* **2016**, 56, 1438-54.

27.   eMolecules,   Inc. *https://www.emolecules.com/*.

28.   *Enamine, Ltd. https://enamine.net/*.

29.   *ChemAxon. JChem, Version 20.8.3, ChemAxon, Ltd: Budapest, Hungary* **2020**.

30.   *Virtual   Screening   Web   Server. http://infochim.u-strasbg.fr/webserv/VSEngine.html*,   December 2020.

31.   Zambaldo, C.; Geigle, S. N.; Satz, A. L., High-Throughput Solid-Phase Building Block Synthesis for DNA-Encoded Libraries. *Org. Lett.* **2019**, 21, 9353-9357.

32.   Satz, A. L.; Cai, J.; Chen, Y.; Goodnow, R.; Gruber, F.; Kowalczyk, A.; Petersen, A.; Naderi-Oboodi, G.; Orzechowski, L.; Strebel, Q., DNA Compatible Multistep Synthesis and Applications to DNA Encoded Libraries. *Bioconjug Chem* **2015**, 26, 1623-32.

33.   Landrum, G., *RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org*.

34.   Lipinski, C. A., Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, 44, 235-49.

35.   Gleeson, M. P., Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, 51, 817-34.

36.   Eidam, O.; Satz, A. L., Analysis of the productivity of DNA encoded libraries. *MedChemComm* **2016**, 7, 1323-1331.

37.   Wu, Z.; Graybill, T. L.; Zeng, X.; Platchek, M.; Zhang, J.; Bodmer, V. Q.; Wisnoski, D. D.; Deng, J.; Coppo, F. T.; Yao, G.; Tamburino, A.; Scavello, G.; Franklin, G. J.; Mataruse, S.; Bedard, K. L.; Ding, Y.; Chai, J.; Summerfield, J.; Centrella, P. A.; Messer, J. A.; Pope, A. J.; Israel, D. I., Cell-Based Selection Expands the Utility of DNA-Encoded Small-Molecule Library Technology to Cell Surface Drug Targets: Identification of Novel Antagonists of the NK3 Tachykinin Receptor. *ACS Combinatorial Science* **2015**, 17, 722-731.

Summary

In this project, around 2 500 DELs of different sizes (from 1M to 100M) were designed using commercially available BBs. A representative subset of 1M compounds for each library was generated, standardized, and projected onto the first universal map. The resulting 2,5B DEL chemical space was analyzed in terms of physicochemical properties and different MedChem rules compliance. It was also compared to biologically relevant compounds from ChEMBL. It appears that there are several ChEMBL-specific regions – zones that are not occupied by any of the DELs. They are populated by complex NPs, like steroids, macrolides, peptides, nucleotides, etc. Thus, in general, DEL technology gives access to the biologically relevant chemical space with a quite expected exception of complex NPs.

However, in a screening campaign, only one DEL will be used. Thus, a 'golden' DEL (or set of a few complementary DELs) that provides the highest coverage of ChEMBL chemical space should be found. With the help of GTM in general, and responsibility patterns (RPs) in particular, it was shown that more than half of DELs could separately cover less than 30% of ChEMBL, and only 90 libraries cover 50-60% of ChEMBL. Considering this rather low value, the possibility of usage of the set of complementary joint DELs (pool of libraries) was investigated. In this case, several DELs should be synthesized separately, followed by their combination in one single vessel for common affinity screening. Each new complementary DEL was identified in a way so that it covers the maximal portion of the ChEMBL chemical space that was not covered in the previous steps. As a result, in the case of 3 combined complementary DELs, ChEMBL coverage increased up to 72%, while simultaneous usage of 5 DELs provides 82% coverage.

This study can be considered as a seminal study of the ultra-large chemical space of DELs. The generated DEL compounds and preliminary results obtained in this project open possibilities to various computational studies that would be highly important for the scientists working in the field of DEL.

4.4 Building blocks

There are two main approaches in the screening library design: i) cherry-picking compounds from extensive screening libraries and ii) rational selection of BBs[124] required for final compounds synthesis. In the first case, various structure- and ligand-based VS methodologies are routinely applied in order to fish out the most suitable compounds for a particular task in mind. However, this approach is limited to the expensive commercial screening collections (analyzed in Chapters 4.1 and 0) available to any potential buyer and biased by the supplier design strategies. In case of a limited budget or if a certain level of novelty and exclusivity is desired, the second approach becomes the best option.

As soon as the quality and diversity of screening compounds unavoidably depend on the BBs used for their synthesis, their rational selection can significantly benefit the drug design process by preliminary focusing on substructures and properties that will ensure desirable activity and ADMETox profile of the potential drug candidates.[16] Even though this fact is widely recognized by medicinal chemists, the number of scientific reports targeting quality analysis of the existing purchasable building blocks (PBB) and potential strategies for the corresponding libraries enhancement is significantly lower than the same for the commercially available screening compounds.

This fact can be explained by several challenges in chemoinformatics treatment of BB structures. From one point of view, the nature of BB is determined by the

protected and unprotected functional groups it contains. They define the list of reactions BB can participate in and partners it can react with. However, in the medicinal chemistry context, functional groups are far less interesting than the increments introduced by BBs to the final molecule. One BB, used under different conditions, can contribute differently to the final molecule and thus be associated with more than one such increment. Similarly, the same increment can be introduced by different BBs.

Up to date, there was no openly available software that would allow BBs analysis in a medicinal chemistry context and compare them with fragments derived from reference molecules. Therefore, the new toolkit called SynthI has been developed to empower chemoinformatics analysis of BBs. Moreover, in the end, its functionality went beyond simple BBs analysis up to the focused library design.

4.4.1 SynthI: a new open-source tool for synthon-based library design and building blocks analysis

Introduction

The rational design of screening libraries is crucial for successful drug discovery, and chemoinformaticians have played a highly important role in its rapid development[125]. Most of the existing technics of computational library design are based on the generation, rational selection, and reassembling of favorable structural motifs to generate members of the new library[126]. Over the last decades, various methodologies that differ mostly in a set of rules applied for fragments generation and recombination were reported. However, the absence of a direct link between the chemical space of the retrosynthetically generated fragments and the pool of available reagents makes such approaches appear as rather theoretical and reality-disconnected.

Therefore, in this work, we have developed a new open-source toolkit for library design called Synthons Interpreter or Synth. It combines the RECAP-like fragmentation approach with a synthons-based way of reagents representation. Synthons are increments of the BB that will be added to the final compound upon a particular chemical reaction. In SynthI, synthons are used as a unified representation of BBs and fragments – they are generated not only from reagents but also as a result of pseudo-retrosynthetic[127] fragmentation of larger molecules of interest. Their distinctive feature is the presence of special markings at the former position of the leaving groups (or bond disconnection if derived from compound fragmentation). The type of the mark defines the type of the reaction center – electrophile, nucleophile, radical etc.

SynthI: a new open-source tool for synthon-based library design

Yuliana Zabolotna[1], Dmitriy M.Volochnyuk[3,6], Sergey V.Ryabukhin[4,6,] Kostiantyn Gavrylenko[5,6], Dragos Horvath[1], Olga Klimchuk [1], Olexandr Oksiuta[3,7], Gilles Marcou[1], Alexandre Varnek[1,2] *

___

Abstract: Most of the existing computational tools for library design are focused on the generation, rational selection, and combination of promising structural motifs to form members of the new library. However, the absence of a direct link between the chemical space of the retrosynthetically generated fragments and the pool of available reagents makes such approaches appear as rather theoretical and reality-disconnected. In this context, here we present Synthons Interpreter (*SynthI*), a new open-source toolkit for library design that allows merging those two chemical spaces into a single synthons space. Here synthons are defined as actual fragments with valid valences and special labels, specifying the position and the nature of reactive centers. They can be issued from either the "break-up" of reference compounds according to 38 retrosynthetic rules or real reagents, after leaving groups withdrawal or transformation. Such an approach not only enables the design of synthetically accessible libraries and analogs generation but also facilitates reagents (building blocks) analysis in the medicinal chemistry context. SynthI code is publicly available at https://github.com/Laboratoire-de-Chemoinformatique/SynthI.

Keywords: library design, synthons, fragmentation, enumeration, building blocks, retrosynthesis

___

INTRODUCTION

The rational design of chemical libraries for activity screening is crucial for successful drug discovery and

1. University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France) *e-mail: varnek@unistra.fr

2.   Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan

3. Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine

4. The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine

5. Research-And-Education ChemBioCenter, National Taras Shevchenko University of Kyiv, Chervonotkatska str., 61, 03022 Kiev, Ukraine

6. Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine

7. Chemspace, Kyiv, Ukraine.

chemoinformaticians have played a highly important role in its rapid development[1] Various computational methods evolved over time to allow chemical data manipulations, structure transformations, de novo generation etc.[2] With such a diversity of existing approaches, the main challenge in modern library design is a trade-off between the theory-inspired novelty introduced by chemoinformaticians and practical considerations of experimentalists.[3] The ability of medicinal chemists to consider both factors is influenced by the availability of the easy-to-use computational tools that provide solutions to the most frequent library design problems while still retaining some level of flexibility embodied in the variety of user-tunable parameters.

Most of the existing technics of de novo library design are based onx the generation, rational selection, and combination of promising structural motifs to generate members of the new library[4]. The first task is

usually achieved by the fragmentation of relevant compounds (for example known ligands of a particular biological target).[5] The resulting fragments or their subset can then be reassembled forming a new library with desired properties. Over the last decades, various methodologies that differ mostly in a set of rules applied for fragment generation and recombination were reported. The most prominent openly available fragmentation method is the retrosynthetic combinatorial analysis procedure (RECAP)[6]. Proposed twenty years ago, it was the first of its kind pseudo-retrosynthetic tool, that applied 11 reaction rules in order to break chemical bonds that can be easily formed via combinatorial chemistry. This methodology together with its latter extension called BRICS[4] has gained extreme popularity and has been used successfully in different drug discovery projects and implemented in several chemoinformatics toolkits, like ChemAxon[7], OpenEye[8], and RDKit[9].

The limitations inherent to the rather small set of reaction rules behind RECAP have been discussed previously, as opposed to the hundreds of automatically extracted reaction schemes introduced in more complex tools for library design and retrosynthetic analysis, like AiZynthFinder[10], Chematica[11], ICSYNTH[12] etc. It is usually claimed that such tools are covering the scope of known chemical reactions more comprehensively. On the one hand, they indeed reflect up-to-date synthesis expertise, but at the same time, they include some sophisticated protocols pertaining to synthetic creativity, rather than an optimal solution for everyday routine problems. Considering how uncertain is the success of the drug design campaign at its early stages, investing more time and resources in the synthesis of the initial screening libraries does not seem very efficient. Therefore, medicinal chemists traditionally use only a tiny fraction of the reactions that allow faster advancement in drug discovery projects, saving complex elaborated procedures for optimization of confirmed leads[13-16].

This tendency is advocated in a recent study, showing that molecular quality, comprising molecular complexity, diversity, and novelty, is typically not related to the type of chemical reactions used to produce screening compounds (excepting targets for which only natural product-like ligands are known).[17] Their diversity, complexity, and novelty are more influenced by the quality of the selected building blocks (BBs). In this context, the absence of the direct link between the chemical space of the generated fragments and the pool of available BBs makes tools like RECAP and BRICS appear as rather theoretical, reality-disconnected approaches, distant from down-to-earth practical library design based on the reagents present in the laboratory drawers.[18] Some methodologies of library de-novo designs considering BBs availability have been previously reported, including both commercial/proprietary software[19-21] and methodologies used mostly by the authoring academic group[22].

Here we describe a new open-source toolkit for synthons-based library design, called Synthons Interpreter (SynthI). In chemoinformatics synthons were first introduced by R.D.Cramer et al.[23] in 2007 as structures with one or more open valences each having a defined reactivity. In this work, synthons are defined differently: the open valence at the connection/disconnection point is complemented by hydrogen atom(s) and a special label determining its reactivity is assigned. The label is associated with those reagent classes (in total there are almost 150 mono- bi- and trifunctional subclasses) that can produce a given synthon (see Table 1). Their chemical validity allows to treat synthons as any other chemical structures: to assess different properties using machine-learning models, to evaluate similarity, and to visualize their chemical space. In the unified scheme presented here synthons can be transparently issued from either the "break-up" of reference compounds according to 38 pseudo-retrosynthetic rules, or from real reagents, after leaving/protective groups removal or any other transformations required to generate the moiety inherited by the reaction product. As a result, SynthI can be used for several tasks: i) analysis of the available BBs collections; ii) global enumeration of all compatible synthons combinations based on the selected reactions and available BBs; iii) detection of BBs producing synthons that are needed to synthesize desired compounds; iv) synthons-based focused library design – a combination of synthons identical or analogous to those obtained via pseudo-retrosynthetic fragmentation of active compounds.

IMPLEMENTATION

General description of SynthI

SynthI is a python3 RDkit-based[9] (2021) library that generates synthons from larger molecules via fragmentation or from small reagents via functional group transformations. Being a knowledge-based tool, SynthI is based on the extensive library of SMARTS, defining each reagent class and SMIRKS that specify the reaction rules for synthon generation from BBs,

pseudo-retrosynthetic bond disconnections, or synthon recombination. SynthI consists of four modules (**Figure 1**), each being responsible for a particular task. In the following chapters, you can find a detailed description of each of them.
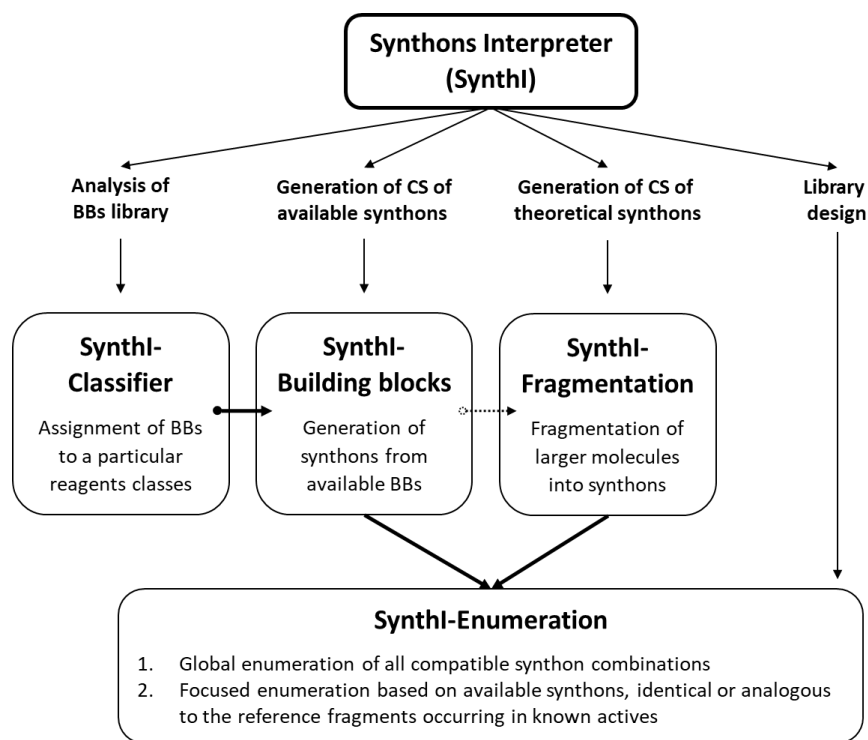

Figure 1. SynthI functionality: analysis of BBs libraries, achieved with SynthI-Classifier; generation of chemical space (CS) of available synthons from the BBs after their classification - SynthI-BBs; generation theoretical synthons CS via fragmentation of larger compounds (with or without the use of available synthons library for prioritizing the fragmentation schemes resulting in a higher portion of available synthons) – SynthI-Fragmentation; library design via global or focused enumeration – SynthI-Enumeration.

SyntI-BBClassifier

The first step in BB processing is a selection – a binary decision-making algorithm returning whether a given molecule may or may not qualify as a reagent of a specified class in a specified reaction. This involves three key aspects:

- Detection of the required characteristic functional group[s] characterizing the envisaged reagent class, which can straightforwardly be achieved by SMARTS pattern matching.
- Analysis of the chemical context in which the characteristic functional group is placed, and which modulates its reactivity. This is a weak point of the procedure because these effects are often long-range (conjugation, inductive effects), geometry-dependent (steric effects, intramolecular hydrogen bonds) and, of course, overlapping (several substituents inducing conflicting and not always additive effects). In absence of a robust global model of chemical

reactivity, SMARTS encoding of the most often seen and impactful structural patterns associated to a loss of functional group reactivity is the only practical solution so far.

- Detection of unprotected competing or cross-reacting functional groups, likely to trigger secondary reactions leading to a mixture of products. For example, in order to be effectively used as an aldehyde reagent, BB should not contain structural moieties of acylators, alkylators, unprotected amino groups, thiols, isocyanates, metalorganics, etc. These may also be provided as a list of SMARTS patterns.

The full list of SMARTS for the BBs classification is provided in *SMARTSlib.json* and *SynthI_AllSmartsFromClassifier.xlsx* files on GitHub page. In total, 22 monofunctional BB classes were considered, like acyl halides, boronics, ketones, primary amines etc. Almost each of them incorporates subclasses, totaling up to 100. For example, class "Alcohols" includes three subclasses that would have different reactivity – "Heterols", "Aliphatic alcohols"

and "Phenols". In addition, there are 28 bifunctional and 19 trifunctional classes. All of them concern only reagents for coupling reactions as soon as the given version of SynthI does not include heterocyclization reactions. From the library design point of view, their usage would lead to the destruction of the privileged scaffolds that contribute significantly to the exhibited activity. Therefore, in the first implementation of SynthI heterocyclization reactions were not taken into consideration. For more detailed retrosynthesis, however, those reactions are highly important, therefore we are currently working on the implementation of the SynthI-Heterocyclization module, that would allow the user to select whether they want to include cycle bonds disconnection.

SynthI-BBs

The same BB can be assigned to several classes followed by the generation of synthons, corresponding to each class using SynthI-BBs module. In each synthon, the special labels are placed at the former position of the leaving groups (**Figure 2** and **Table 1**). They define the type of the bond disconnection and reaction center (RC) – electrophile, nucleophile, radical, etc. The full list of unique synthons generated from the user-provided BBs library produces a chemical space of available synthons. In the case of a compound, containing protective groups it is up to the user to decide whether to keep protected synthons or not (*keepPG* option). The list of all synthons generated from each BB class is provided in *SynthI_BB_classes_and_respectiveSynthons.xls, which can be found* on GitHub page of the project.
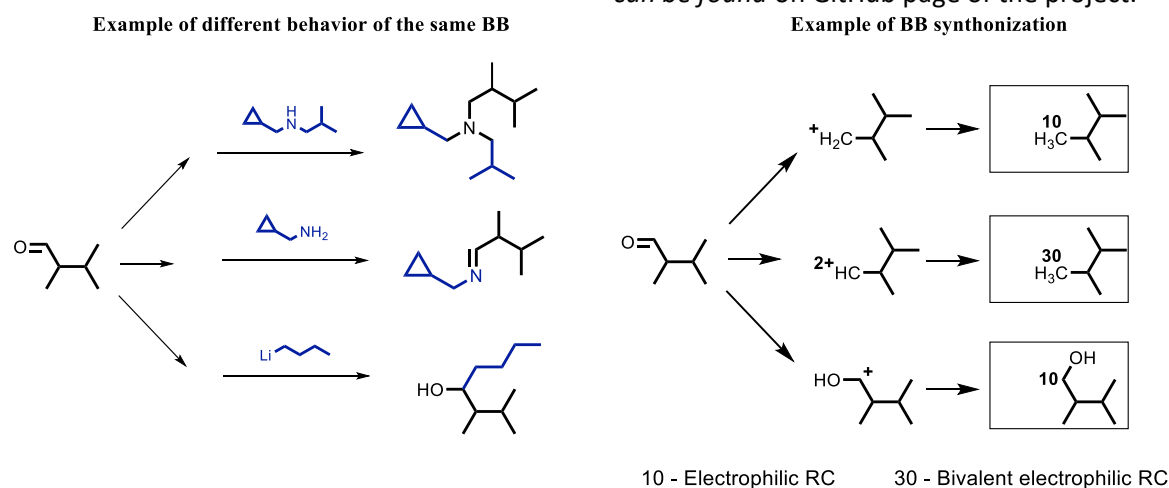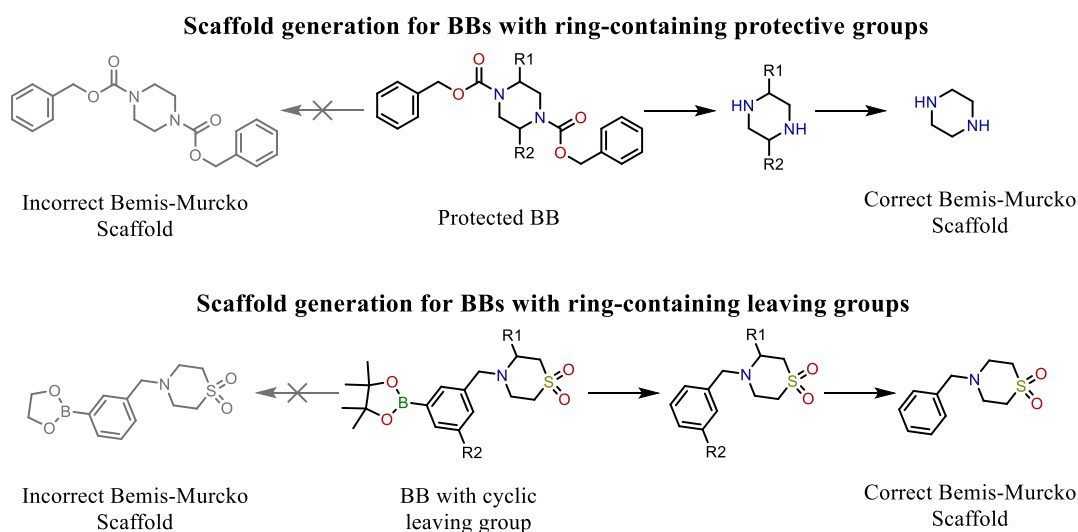

Figure 2. Example of different behavior of the same BB (here - aldehyde) and generation of corresponding synthons. Labels on the synthons define the nature of the reaction center (RC).

Scheme 1. Scaffold generation in BBs analysis. Ring-containing protective and leaving groups should be removed before generating a scaffold.

Table 1. Synthons labels and examples of corresponding reagents.

Synthon Lable	Examples of Synthon	Nature of the reaction center (RC)	Example of corresponding reagent classes
$AH_n{:}10$	(structures)	Electrophilic	Acyl, aryl and alkyl halides, sulfonylhalides, anhydrides, acides, aminoacids, esters, alcohols, aldehydes, ketones, Weinreb amides, acylated azides, iso(thio)cyanates, oxiranes
$AH_n{:}20$	(structures)	Nucleophilic	Alcohols, thiols, amines, amides, NH-azoles, hydrazines, hydrazides, hydroxylamines, oximes, esters, element organics, metal organics, ketones, aryl and allyl sulphones, alkenes for Heck couplings
$CH_n{:}30$	(structures)	Bivalent electrophilic	Aldehydes, ketones
$AH_n{:}40$	(structures)	Bivalent nucleophilic	Ketones, primary amines, hydrazines, hydroxylamines, reagents for olefination (Jullia-Kocienski, Wittig, Horner-Wadsworth-Emmons)
$CH_3{:}50$	(structures)	Bivalent neutral	Terminal alkenes (for metathesis)
$CH_n{:}60$	(structures)	Electrophilic radical	Minisci CH-partners, Michael acceptors
$CH_n{:}70$	(structures)	Nucleophilic radical	BF3 and MIDA boronates, oxalate alkyl esters, NOPhtal alkyl esters, sulphinates
$CH_n{:}21$	(structures)	Boronics-derived nucleophilic	Boronic reagents
NH:11	(structures)	Electrophilic nitrogen	Benzoyl O-acylated hydroxilamines

Scaffold generation for BBs

The most common approach for the structural analysis of any compound library is to generate scaffolds[24] - cyclic molecular cores without side chains - and count the frequency of their occurrence in the compound collection[25]. For the analysis of reagent libraries, BB structures need to be preprocessed prior to the scaffolds generation by removing any ring-containing moieties that are not parts that will not be kept in the reaction product and thus are irrelevant in BB analysis (**Scheme 1**). It includes some protective (benzyl (Bnz),

benzyl carbamate (Cbz) and fluorenylmethyloxycarbonyl (Fmoc)) and leaving groups (boronics, oxiranes). Based on such preprocessing, SynthI allows to generate relevant BBs scaffolds, count their occurrence in the provided collection of BBs, and even construct cumulative scaffold frequency plot.

SynthI-Fragmentation

The chemical space of theoretically relevant synthons can be generated via pseudo-retrosynthetic bond disconnection of the relevant compounds (e.g. ligands

of a particular target) implemented in SynthI-Fragmentation. It is based on the most common combinatorial reactions, expressed via SMIRKS. Previously, 11 RECAP bond cleavage rules were proposed based on the "commonly used" combinatorial chemistry. However, after more than 20 years these rules needed to be revised in accordance with modern synthetic techniques. In addition, in RECAP and BRICS for each type of bonds there was only one disconnection rule. However, the same bond can be formed by different reagents via reactions that can

have completely different mechanisms. For example, N-alkylation of lactams can be performed via nucleophilic substitution of alkyl halides or via Chan-Lam coupling with boronic acids (Scheme 2). In this context, in order to be able to link the chemical space of available synthons, generated from provided BBs library, to the synthons resulted from fragmentation, several rules of disconnection are needed for the same bond type.
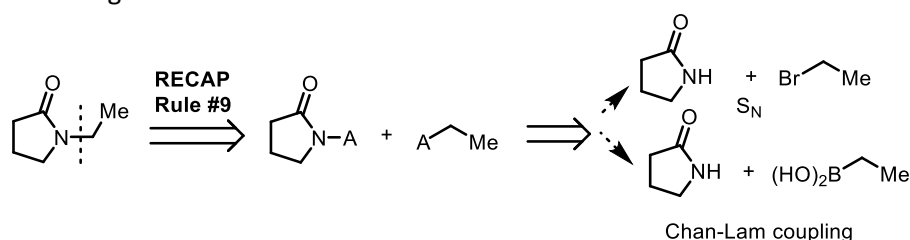
Scheme 2. Example of RECAP disconnection of the bond that can be formed via different reactions.

The reaction rules behind SynthI were collected based on the analysis of current literature and our experience in medicinal chemistry synthesis. It included various reactions, leading to:
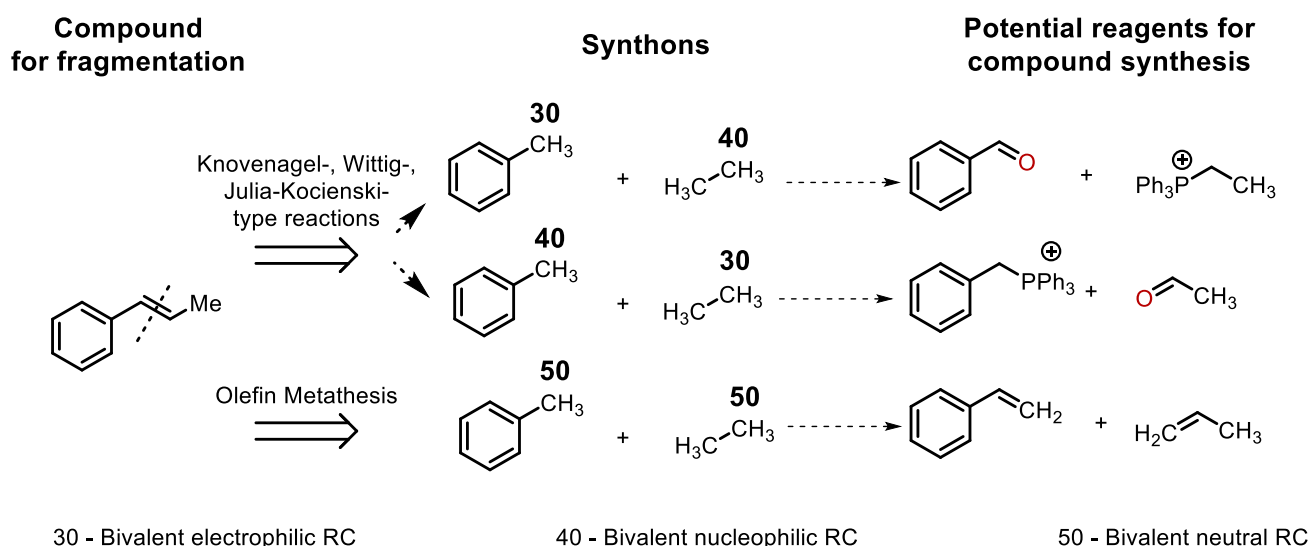
- several ways of disconnection of the same strategic bonds that were already considered in RECAP and/or BRICS (Buchwald-Hartwig amination[26], Cu-mediated C-N/O coupling[27], umpolung cross-coupling[28], Chan–Evans–Lam coupling[29], olefin metathesis[30], non-classical carbonyl olefination (like Julia-Kocienski)[31, 32], C-H activation[33], sulfonyl fluorides chemistry[34], Suzuki C_{Ar}-C_{Ar} cross-coupling, novel methods for C_{Ar}-C_{sp3} couplings).

- disconnection of the new strategic bonds absent in the previous implementation (Heck C_{Ar}-C_{sp2}, Sonogashira C_{Ar}-C_{sp} and Suzuki C_{sp2}-C_{sp2} couplings, imines, oximes, hydrazones and semicarbazones synthesis, sulphinic acid salts alkylation and their Cu-catalyzed arylation)

Also, the set of new radical chemistry, as well as new methods of late-stage functionalization (Baran diversinates[35], Minisci-type reaction[36]), were included in SynthI. These new reactions dramatically changed modern retrosynthetic thinking of the medicinal chemist[14, 37], and the new more effective conditions for such reactions still actively investigating[38].

In total SynthI contains 13 broad reaction types for the bond disconnections and 37 subtypes, that may lead to different synthons. For example, for the "Olefination" type, there are two subtypes – "Knoevenagel-, Wittig-, Julia-Kocienski- type reactions" and "Olefin Metathesis". The first one is the example of polar bond disconnection resulting in bivalent electrophilic and nucleophilic synthons, while the second one produces neutral biradicals (**Scheme 3**). Obtained synthons can be traced back to the potential BBs for compound synthesis. The full list of reaction rules with some examples is available in the Supporting Information.

SynthI-Fragmentation allows one to select a subset of reactions, but in this study, all of them are used. After each cut, the combination of synthons from which molecule can be synthesized is stored. If more than one bond in a molecule can be disconnected, then the hierarchy of all possible disconnections and resulting synthons combinations are stored. Given the list of "available" synthons provided by the available BBs, fragmentation schemes predominantly returning fragments listed amongst these available synthons are obviously preferable. The availability rate is herein defined as the percentage of heavy atoms of the fragmented compound that can be provided by available synthons:

$$Availability\ rate = \frac{\sum heavy\ atoms\ coming\ from\ available\ synthons}{Total\ number\ of\ heavy\ atoms\ in\ a\ molecule} * 100\%$$

| Compound for fragmentation | Synthons | Potential reagents for compound synthesis |

30 - Bivalent electrophilic RC   40 - Bivalent nucleophilic RC   50 - Bivalent neutral RC

Scheme 3. Example of SynthI reaction type with two subtypes representing different mechanisms of the same bond formation/disconnection. Labels on the synthons define the nature of the reaction center (RC).

Based on this value, the optimal pathway can be selected to be written into the summary output file (see SI). One may also navigate the disconnection hierarchy using several built-in functions. More details on the usage of SynthI and tutorial can be found on the GitHub page (https://github.com/Laboratoire-de-Chemoinformatique/SynthI).

SynthI-Enumeration

This last module applies the list of the abovementioned reaction rules in order to generate the full combinatorial library of all compounds that can be synthesized using a given set of synthons (**Figure 3**). Users can control the maximum number of synthons that can be combined together. As well as the list of reactions for enumeration. If the maximal number of synthons has been reached but some open RCs were left this product will be discarded.
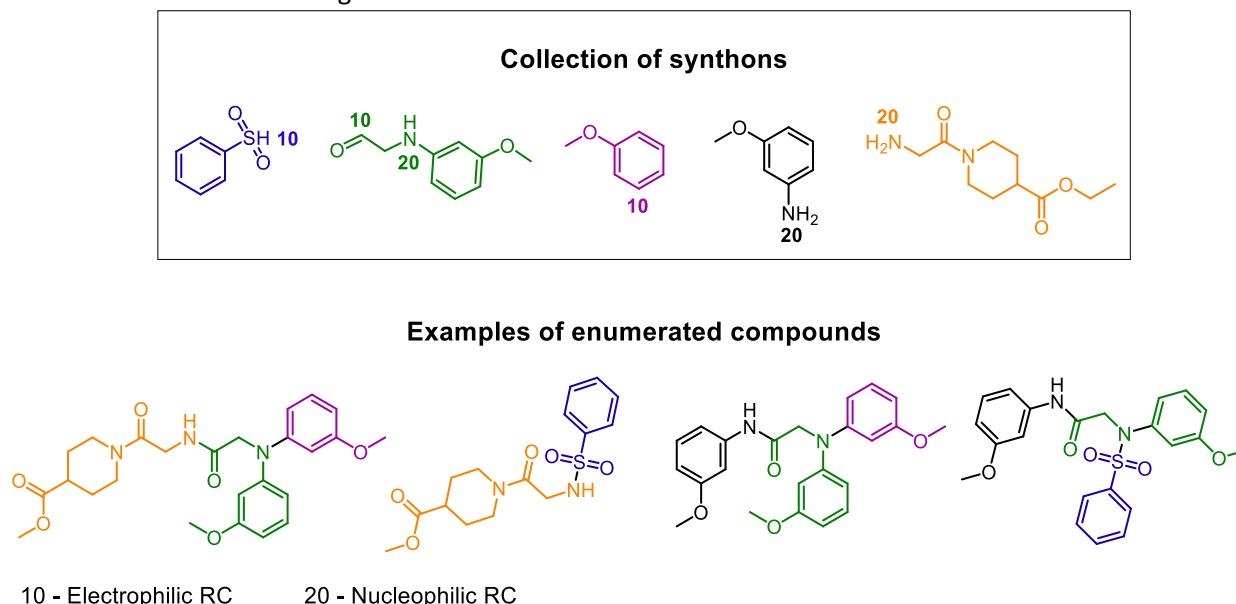

Figure 3. Example of library enumeration using a user-provided collection of synthons.

SynthI-Enumeration also allows to generate a focused library of the synthesizable analogs of the provided compound. The input molecule is first fragmented up to the smallest synthons. Their availability is checked using the BBs synthons library. The same library is used for the search of the analogs

of generated synthons - synthons containing the same types of RCs (but not necessarily in the same positions), the same number of rings, and matching the constraints, adopted from the positional analogs scanning (PAS) strategy for lead optimization[39]. According to the latter, analogs should be a substructure/superstructure of the original compound (in our case synthon) and differ from it only in the absence/presence of one functional group: CH_3, F, NH_2, OH or be a result of C_{Ar}->N_{Ar} or N_{Ar}->C_{Ar} replacements. These rules have been changed slightly to soften the criteria for synthons selection in order to enable producing more comprehensive focused libraries. Thus, the structural isomers were also considered analogs. In addition, there is a possibility for the user to specify the synthons similarity threshold that will be applied independently of the previous filters for the search of additional analogs of the original synthon via similarity approach. The rules concerning RC types and number of rings are used for all analogs selection including sililarity. The Tanimoto coefficient is calculated with RDKit using Morgan fingerprints (radius=2, nBits=2048) as descriptors.

With *strictAvailabilityMode* only synthons that were found in the available BBs or have available analogs are selected for library generation. If one of the required synthons does not have any direct or analogous correspondence in the provided BB library, easily synthesizable analogs for the input molecule can not be generated. Otherwise, unavailable synthons will be also used for focused library design. The new library generation is based only on the reaction according to which compound was fragmented. The number of combined synthons is fixed to the number of synthons obtained via molecule fragmentation in a selected synthetic path.

DATA FOR CASE STUDY

As a source of available BBs, the library of 201 675 in-stock reagents provided by Enamine was used. 79 drugs, recently approved by FDA have been used as a dataset for fragmentation and analogs generation. The full list together with fragmentation results can be found in Supporting Information.

RESULTS AND DISCUSSION

The weak spot of the RECAP-like tools is their potentially low propensity to propose the exact same fragments that are provided by real-world BBs ready to use in the laboratory. This gap can be bridged by

introducing an unified chemoinformatics formalism to handle the synthon chemical space of both RECAP fragments and BB-provided, "available" synthons. From one point of view, the nature of BB is determined by the protected and unprotected reactive functional groups it contains. They define the list of reactions BB can participate in, and partners it can react with. However, in the medicinal chemistry context, those leaving groups are less interesting than the structural, pharmacophoric or physico-chemical features that will be contributed by the BB to the final molecule. One BB, used under different conditions can contribute differently to the final molecule, while the same structural fragments can be introduced by different BBs (**Figure 2**). Using synthons as a unified representation, SynthI allows merging the chemical space of BBs (or rather structural increments that they bring to the final molecule) with a chemical space of fragments, obtained via pseudo-retrosynthetic bond disconnections. The herein-developed system of labels encodes the position and chemical nature of the reactive centers while preserving structure validity, allowing to treat synthons as actual compounds. This not only enables the design of synthetically accessible libraries but also facilitates BB analysis in the medicinal chemistry context.

BB classification, synthonization and scaffold analysis

Out of 201 675 BBs used in this work, 18 were not processed by RDKit and 25 414 reagents were not assigned to any classes implemented in the first version of SynthI (mostly reagents for heterocyclization like nitriles, oximes, etc.). For the remaining 176 261 BBs, 388 019 synthons were generated. In **Figure 4** one can see examples of BB classification and synthonization. Some of the BB classes, e.g. secondary amines, produce only one synthon per BB (**Figure 4A**). Others, like ketones, can result in numerous synthons depending on the reaction conditions (**Figure 4C**). An example of aminoesters synthonization with option *keepPG* is shown in *Figure 4E*.

The advantage of adopted synthon representation is that in SynthI synthons are neutral structures with valid valences. The RC position and nature are encoded via atom mapping, which does not change the synthon structure. This allows to analyze them as any other compounds. For example, it is possible to calculate their physicochemical properties and filter them according to the rule of two (Ro2). This rule has been introduced by Goldberg et al.[40] as a simple way of BBs prioritization for designing compounds with physical properties that are suitable for oral administration.

According to Ro2, increment that will be introduced to the molecule by BB should have such properties: MW ≤ 200, logP ≤ 2, H-bond donors ≤ 2, H-bond acceptors ≤ 4. SynthI allows filtration of synthons according to this rule at the stage of synthons library generation from available BBs, fragmentation (for the synthesability check) or analogs library enumeration (for control of the physical properties of generated compounds) (**Figure 5**).

10 - Electrophilic RC   20 - Nucleophilic RC   30 - Bivalent electrophilic RC   40 - Bivalent nucleophilic RC

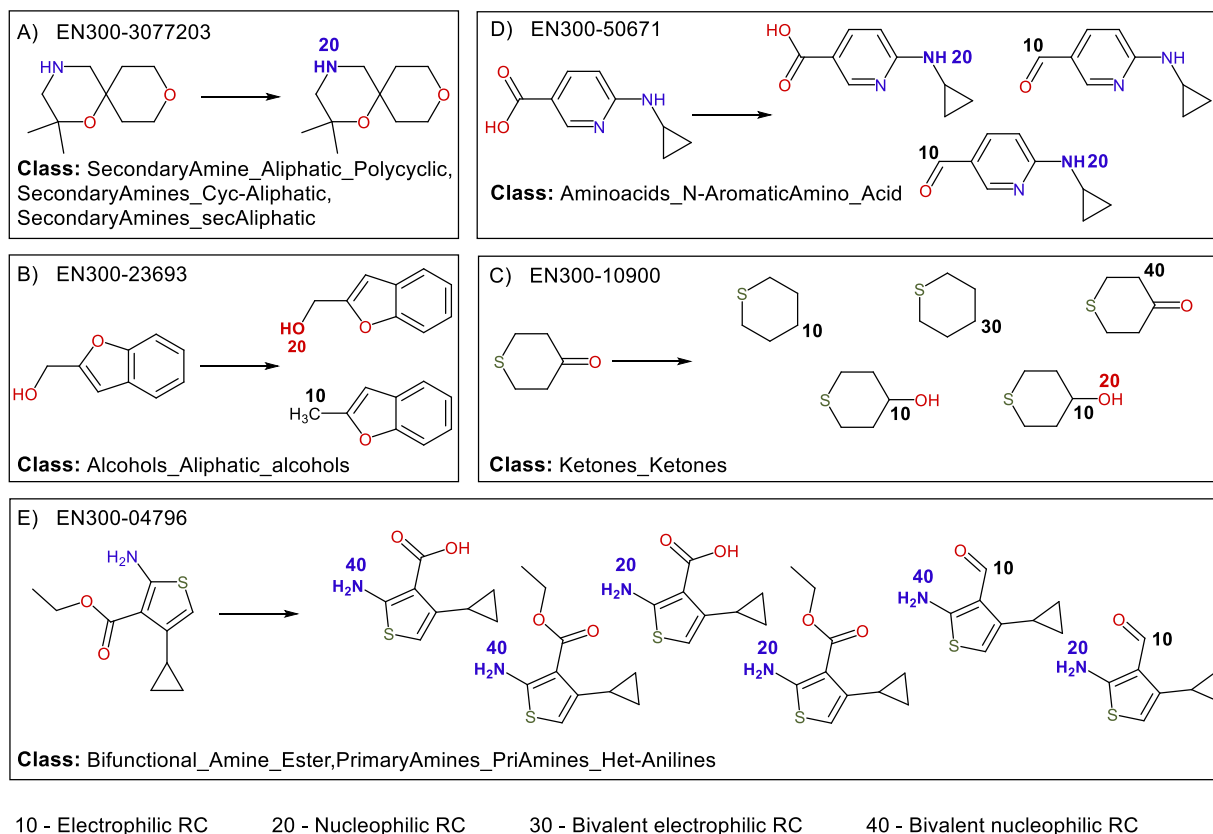**Figure 4.** Examples of BB classification and synthonization. Labels on the synthons define the nature of the reaction center (RC).

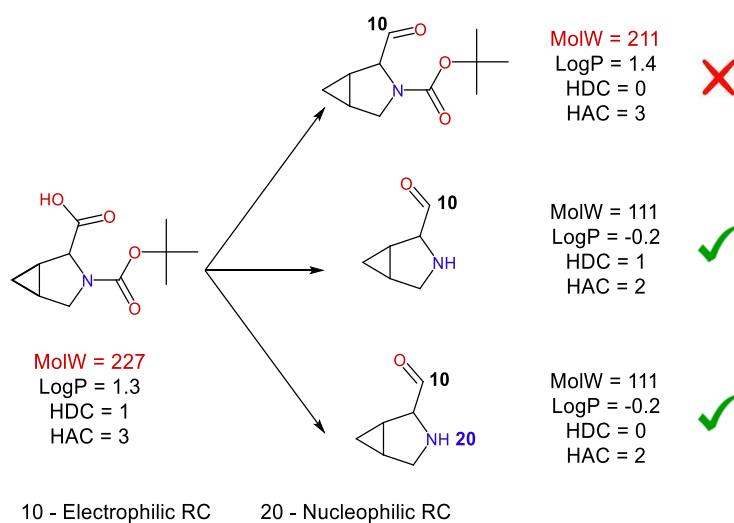10 - Electrophilic RC   20 - Nucleophilic RC

Figure 5. Ro2 synthons filtering for BB prioritization (MW ≤ 200, logP ≤ 2, H-bond donors ≤ 2, H-bond acceptors ≤ 4).

Scaffoldization of 200K Enamine BBs resulted in 19 820 scaffolds with the majority of them (12 272 or 62%) being singletons (occur only in one BB). As one can see in **Figure 6**, a very tiny fraction of scaffolds (<1%) covers almost 60% of BBs from the analyzed collection. The most frequent scaffolds are simple one-ring structures - benzene, pyridine, pyrazole, piperidine, pyrrolidine, cyclohexane, thiophene, and cyclopropane – and the diversity of BBs libraries is mostly gained via their side chains decorations.
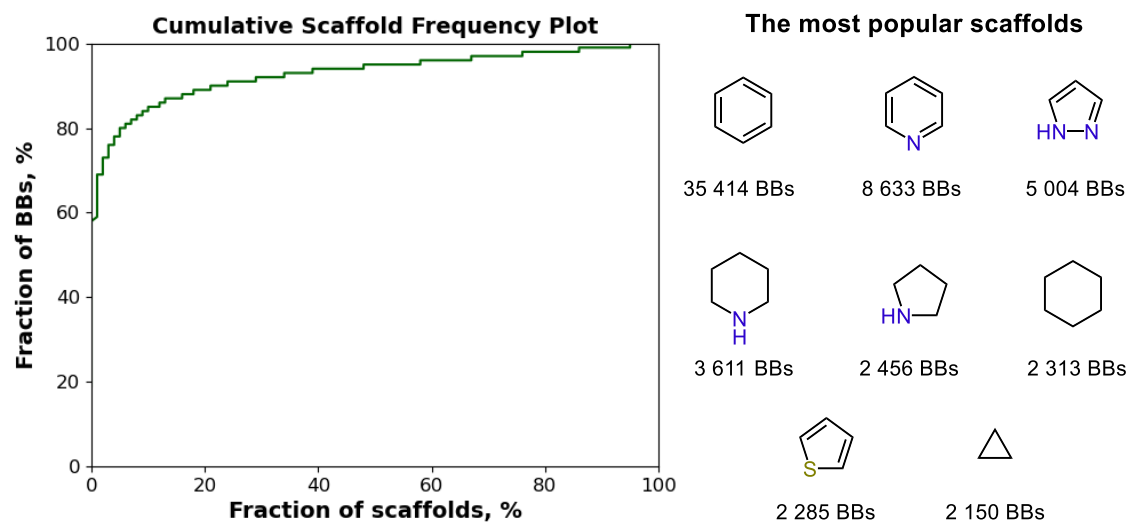

Figure 6. Scaffold analysis of the BBs library.

Fragmentation of FDA approved drugs

As a case study for SynthI-Fragmentation, 79 drugs FDA-approved in 2020 were used examples of compounds to be circumscribed by focused combinatorial libraries of analogues, using the above-processed available BBs. All molecules, except osilodrostat, were fragmented and the optimal set of 2-6 synthons were selected. Out of them, 8 molecules resulted in a set of synthons with a 100% availability rate (all required synthons were incarnated in existing BBs). In order to evaluate the accuracy of the proposed fragmentation schemes from the experimental synthesis perspective, it was compared to the published synthetic pathways (found using Reaxis®[41, 42] and SciFinder) for each of the case study drugs (see Supporting Information). For 24 drugs, SynthI fragmentation fits perfectly to the experimentally validated synthetic procedures. Fragmentation results for the other 18 drugs have minor discrepancies caused by the absence of heterocyclization and reduction/oxidation reactions. Heterocyclization reactions prevail in the synthesis of the remaining compounds and thus corresponding literature data for these compounds cannot be fairly compared to SynthI fragmentation results.

In **Scheme 4** one can see the hierarchy of synthons and reactions, resulted from the fragmentation of cenobamate. SynthI-Fragmentation produced four synthetic pathways, each including two stages. The optimal pathway consisted of consecutive application of S_N alkylation and O-acylation disconnection rules. Two out of three resulted synthons were found in the provided synthons library (availability rate = 72%). The synthetic pathway found in literature is highly similar to the one, proposed by SynthI[43]. The difference is in the usage of the 2-bromo-1-(2-chlorophenyl)ethanone as a precursor for 2-bromo-1-(2-chlorophenyl)ethanol and chlorosulfonyl isocyanate instead of trichloroacetyl isocyanate for the introduction of carbamate moiety.

Analog search case study

Exploring analogs of a reference molecule in terms of combinations of analogues of its constituent BBs is widely used for navigation of very large commercial and proprietary virtual libraries like WuXi Apptec, Enamine REAL (1.3B)[44], Enamine REAL space (29B)[45], Eli Lilly PLC (10^10)[46], BICLAIM by Boehringer Ingelheim (10^11)[47], Pfizer Global Virtual Library (10^14)[21] etc. All of them are based on the fixed internal collections of reagents and reactions, but with the help of SynthI, it becomes possible to navigate in a similar manner a customized non-combinatorial chemical space, defined by the user-selected reactions and BB collections.

With the help of SynthI, one can perform a retrosynthetic fragmentation of compounds of interest, search for the available BBs producing synthons that

are similar to the resulting fragments, and thereupon enumerate analogs of the initial compound. As a result of SynthI application with activated *strictAvailabilityMode* and additional similarity synthons selection option (with Tanimoto coefficient ≥ 0.5), analogs for 23 out of total 79 drug compounds were generated. The number of compounds in the

analog libraries varies significantly - from 4 compounds for cenobamate to almost 7M for fedratinib (see Supporting Information). The size of the analog libraries depends on the number of synthons resulted from initial compound fragmentation and the number of analogs synthons found in the Enamine collection.

10 - Electrophilic RC   20 - Nucleophilic RC   21 - Boronics-derived nucleophilic RC
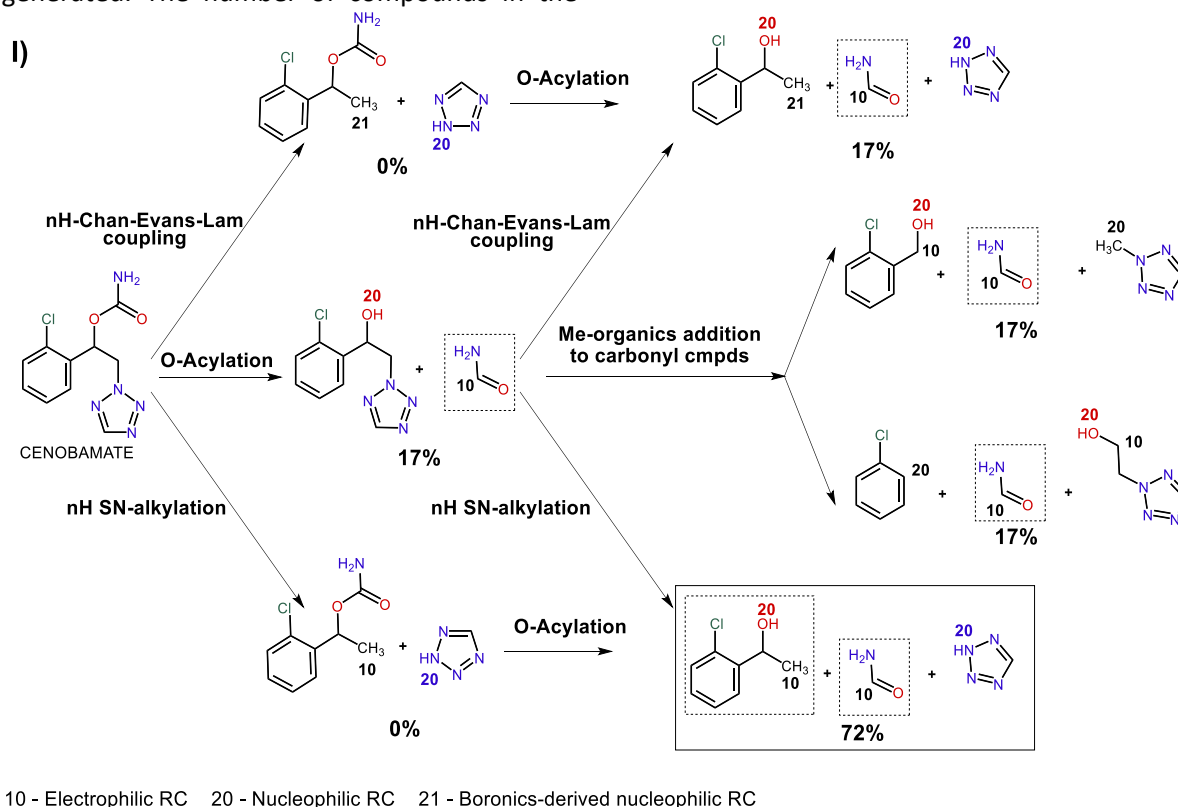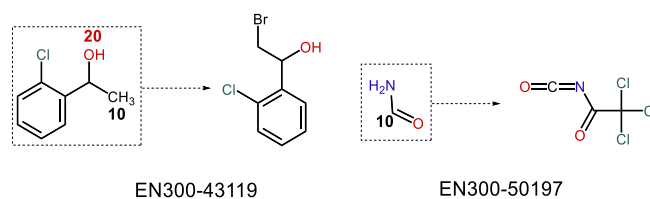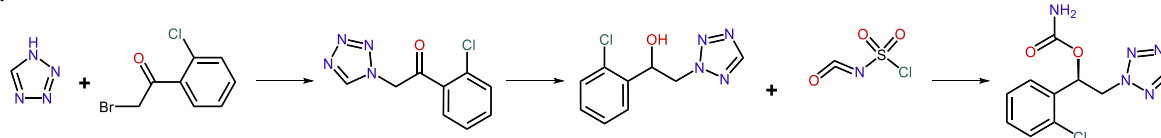
Scheme 4. (I) Example of SynthI fragmentation of cenobamate with the full synthetic hierarchy and experimentally validated synthesis of this compound. The number near the selected set of synthons corresponds to its Availability Rate, %. (II) Available BBs, their identifiers in Enamine catalog and related synthones (in dashed frames). (III) Synthesis of cenobamate reported in reference[43.]

In **Figure 7** one can see an example of the analog generation for solriamfetol. For this molecule, there are three possible fragmentation schemes, but only

one of them results in a set of synthons that are present as such or represented by close analogs in the available synthons library. As it was previously

explained in the methods, there are several sets of rules according to which two synthons may be considered analogs: i) they differ by simplest PAS modifications, ii) are isomers of each other or iii) have synthon similarity above a specified threshold (here Tanimoto coefficient ≥ 0.5). In **Figure 7** the examples of

synthon analogs for each of these categories are given. Solriamfetol analogs generated using them are also provided and as one can see, they are structurally very close to the starting drug, but still providing some level of diversity inside the focused solriamfetol library.
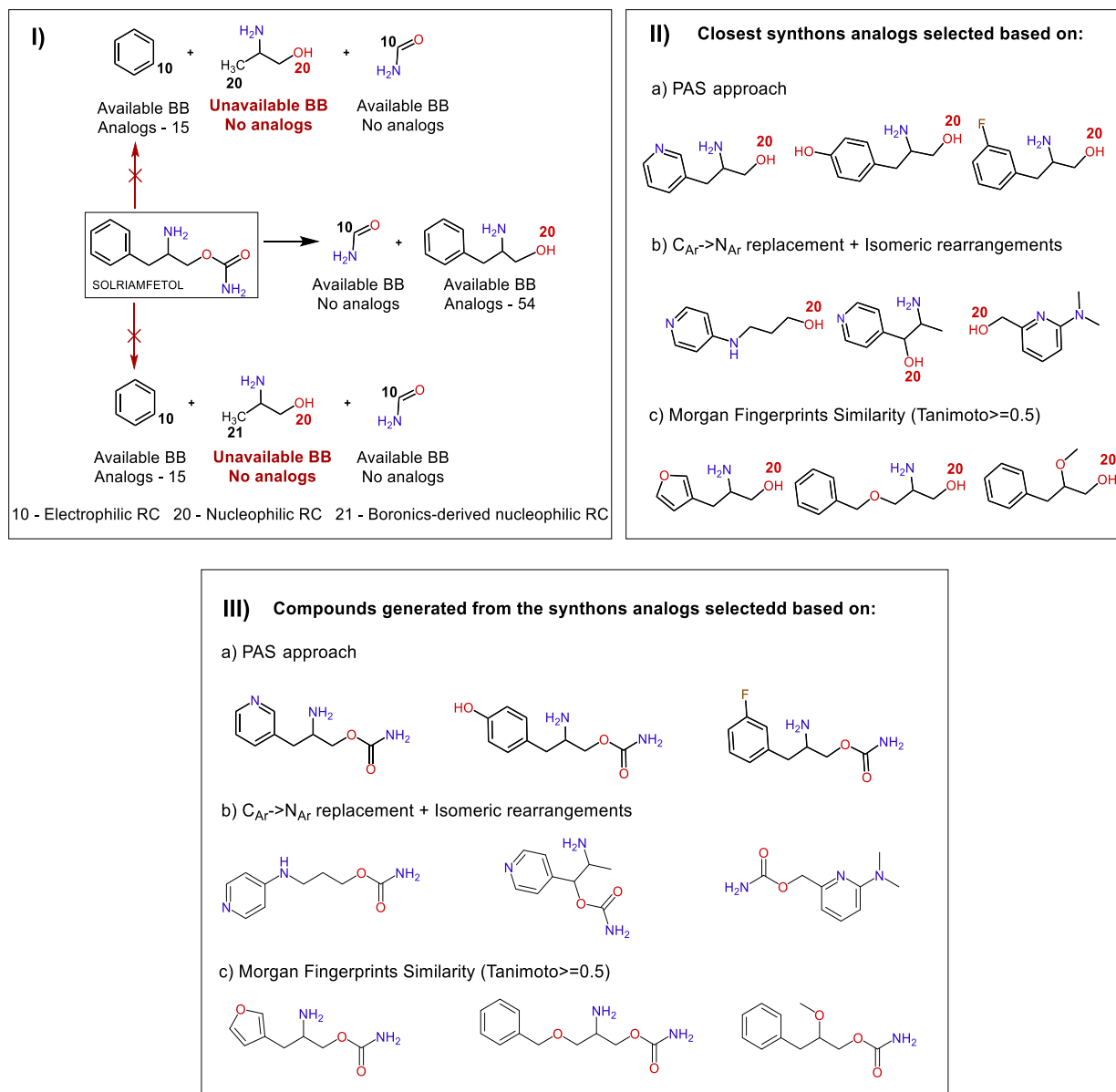
Figure 7. Synthons-based generation of solriamfetol analogs. (I) Solriamfetol fragmentation and synthetic pathways selection. (II) Selection of the closest synthon analogues based on (a) PAS approach, (b) CAr->NAr replacement + isomeric rearrangements and (c) Morgan Fingerprints Similarity (Tanimoto>=0.5). (III) Compounds generated from synthons selected at the step (II).

Considering that the similarity score is always a function of selected descriptors, for the unbiased analysis we need the reference library that would serve as some kind of internal "calibration" scale of the similarity score. In order to create such a library, the simplest PAS modifications (CH$_{Ar}$→F, CH$_{Ar}$→OH, CH$_{Ar}$→CH$_3$, CH$_{Ar}$→NH$_2$ and CH$_{Ar}$→N$_{Ar}$) of the chemical

structure of the reference compound (Molecule 1 **Figure 8**) was performed. Note that modifications were applied manually to the whole structure of the reference compound and not to the underlying fragments like it is done in SynthI. As a result, the reference focused library (RefLib) containing 53 analogs of Molecule 1 was obtained. These compounds

differ only by one atom from the reference molecule, thus their similarity to it can set up a "baseline" of what

to consider as similar compounds in the chosen descriptor space.

I) Similarity distribution initial molecule vs generated analogs

Library2 with additional synthons being selected by T≥0.5

Library1 generated with default setup

II) Initial molecule fragmentation

Analogs - 8   Analogs - 39   Analogs - 2   Analogs - 20
          Analogs - 16   Analogs - 73

III) SynthI-generated analogs

0.2   0.3   0.4   0.5

0.6   0.7   0.8
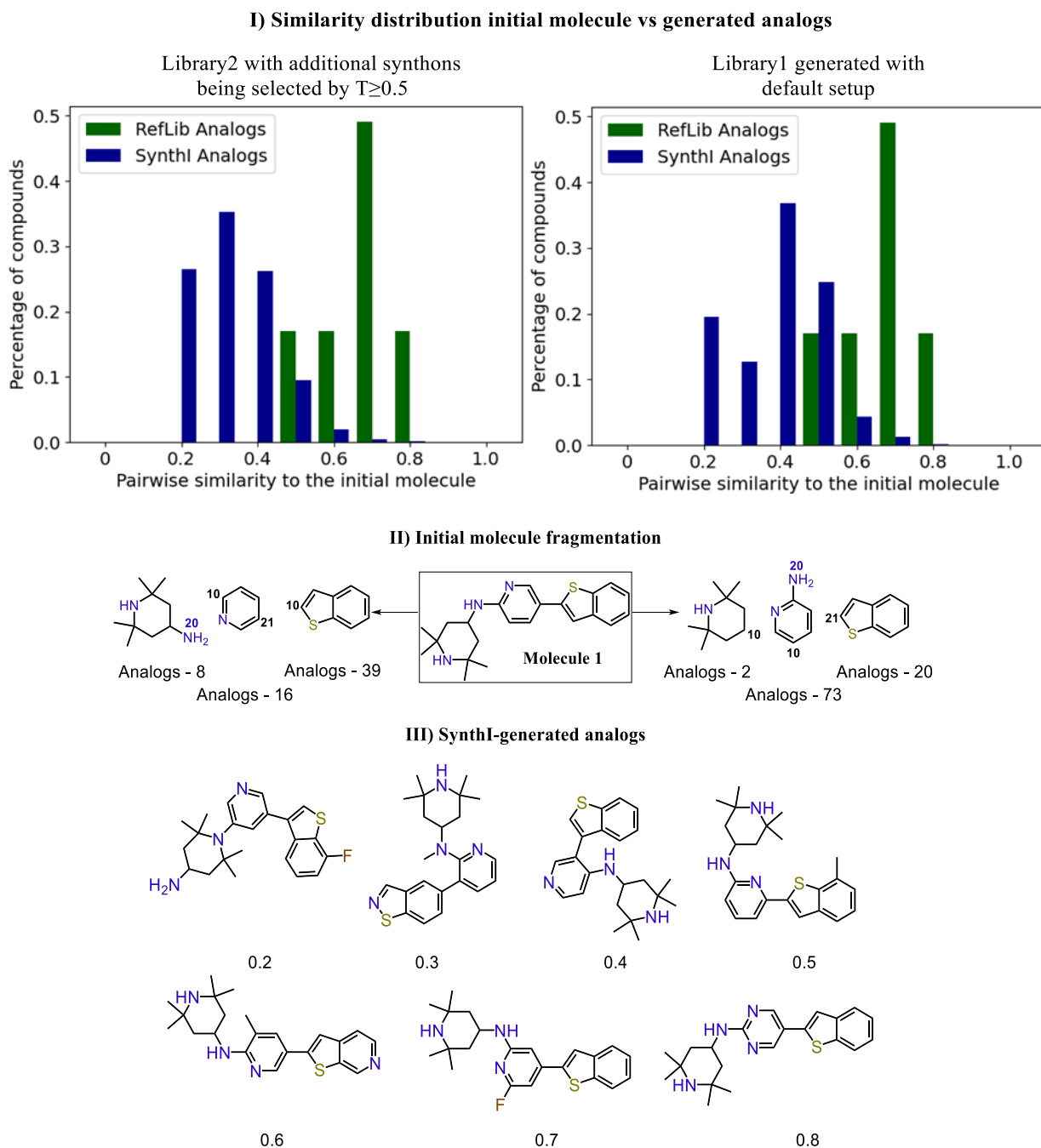
Figure 8. I) Comparison of the similarity distribution between the initial molecule and three analogs libraries (RefLib, SynthI-generated Library1 (default setup) and Library2 (additional synthons being selected by the T≥0.5). I) Fragmentation of the initial Molecule1 and number of analogs found for each synthon. III) Examples of generated analogs of Molecule1 with different similarities to the initial compound. The numbers correspond to pairwise Tanimoto similarity with Molecule 1**.**

From the other side, with the help of SynthI-Enumeration we have generated two libraries of analogs: i) Library1 - 2 593 compounds with a default SynthI setup and ii) Library2 - 8 928 compounds with

activated similarity synthons selection (additional synthons were selected as analogs if their similarity to one of the original synthons was higher than 0.5). Morgan Fingerprint similarity between Molecule 1 and

each member of these two libraries was compared to the same values for the 53 closest analogs from RefLib. As one can see from **Figure 8 (I)**, SynthI-generated compounds, especially from Library2, possess higher diversity with respect to Molecule 1 than analogs from RefLib. This is an expected and desired result, that follows from the adopted approach of the search of synthons analogs rather than direct analogs of the molecule. In the second case, only a single modification is allowed for the whole molecule, while in the first one this rule concerns each synthon, resulting in more diverse compounds.

Examples of analogs with different similarities to the initial molecules are given in **Figure 8 (III)**. As one can see, compounds with Tanimoto coefficient less than 0.5 are still quite similar to Molecule 1. Their distinctive feature is isomeric rearrangements in the position of substituents in the pyridine ring. Analogs with higher similarity mostly have pyridine substituted in the same positions as Molecule 1, which should increase not only structural but also shape similarity. Depending on the task in mind, the user can generate only the closest analogs with the default SynthI-Enumeration setup or also more diverse compounds by activating additional synthons selection with user-defined Tanimoto similarity threshold. This together with the ability to select reactions for bond disconnection/reassembling and BBs, provide a wide range of freedom for users.

CONCLUSIONS

In this work, a new open-source toolkit for library design, called Synthons Interpreter or SynthI, was developed. It connects the building blocks (BBs) and fragments, derived from the pseudo-retrosynthetic fragmentation of larger compounds, via synthons-based representation. It is based on 38 reaction rules for bond disconnection. Their application results in a set of synthons that thanks to the presence of the special labels can be traced back to around 150 types of BBs. A herein-developed system of labels encodes the position and chemical nature of the reactive centers while preserving structure validity, allowing to treat synthons as actual compounds. Such an approach not only enables the design of synthetically accessible libraries but also facilitates BBs analysis in the medicinal chemistry context.

Here, SynthI was tested on the Enamine in-stock BB library for reagent classification, filtration and scaffold analysis. The list of recently approved drugs was used for compound fragmentation. The synthetic pathways

for those compounds reported in the literature were compared to SynthI results, demonstrating its accuracy in almost all cases, except heterocyclization steps, that have not been implemented yet. The analogs libraries were also generated for some of the drugs. The distinctive feature of SynthI library design is its strong dependence on the available BBs. Synthons-based library design allows generating collections of synthesizable compounds, that are structurally similar to the initial molecule and yet diverse with respect to each other.

Supporting Information

Supporting information is available at https://github.com/Laboratoire-de-Chemoinformatique/SynthI.

REFERENCES

1. Zhou, J. Z. Chemoinformatics and Library Design. In *Chemical Library Design*, Zhou, J. Z., Ed.; Humana Press: Totowa, NJ, 2011, pp 27-52.
2. Lenci, E.; Trabocchi, A., Smart Design of Small-Molecule Libraries: When Organic Synthesis Meets Cheminformatics. *ChemBioChem* **2019**, 20, 1115-1123.
3. Jamois, E. A., Reagent-based and product-based computational approaches in library design. *Curr. Opin. Chem. Biol.* **2003**, 7, 326-330.
4. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M., On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, 3, 1503-1507.
5. Fechner, U.; Schneider, G., Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *J. Chem. Inf. Model.* **2006**, 46, 699-707.
6. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 511-522.
7. *ChemAxon, Ltd Budapest, Hungary,* http://www.chemaxon.com.
8. *OpenEye Scientific Software, Santa Fe, NM.* http://www.eyesopen.com.
9. Landrum, G., *RDKit: Open-Source Cheminformatics Software.* http://www.rdkit.org.
10. Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E., AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **2020**, 12, 70.

11.Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A., Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, 4, 522-532.

12.Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H., Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Organic Process Research & Development* **2015**, 19, 357-368.

13.Roughley, S. D.; Jordan, A. M., The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, 54, 3451-3479.

14.Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W., The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **2016**, 45, 546-576.

15.Cooper, T. W. J.; Campbell, I. B.; Macdonald, S. J. F., Factors Determining the Selection of Organic Reactions by Medicinal Chemists and the Use of These Reactions in Arrays (Small Focused Libraries). *Angew. Chem. Int. Ed.* **2010**, 49, 8082-8091.

16.Brown, D. G.; Boström, J., Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, 59, 4443-4458.

17.Tomberg, A.; Boström, J., Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discov. Today* **2020**, 25, 2174-2181.

18.Hartenfeller, M.; Renner, S.; Jacoby, E., Reaction-Driven De Novo Design: a Keystone for Automated Design of Target Family-Oriented Libraries. *De novo Molecular Design* **2013**, 245-266.

19.Yasri, A.; Berthelot, D.; Gijsen, H.; Thielemans, T.; Marichal, P.; Engels, M.; Hoflack, J., REALISIS: A Medicinal Chemistry-Oriented Reagent Selection, Library Design, and Profiling Platform. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2199-2206.

20.Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J., SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, 46, 2765-2773.

21.Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A., Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Combinatorial Science* **2012**, 14, 579-589.

22.Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G., DOGS: Reaction-Driven de novo Design of Bioactive Compounds. *PLOS Computational Biology* **2012**, 8, e1002380.

23.Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B., AllChem: generating and searching 1020 synthetically accessible structures. *J. Comput. Aided Mol. Des.* **2007**, 21, 341-350.

24.Bemis, G. W.; Murcko, M. A., The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.

25.Langdon, S. R.; Brown, N.; Blagg, J., Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, 51, 2174-2185.

26.Ruiz-Castillo, P.; Buchwald, S. L., Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chem. Rev.* **2016**, 116, 12564-12649.

27.Evano, G.; Blanchard, N.; Toumi, M., Copper-Mediated Coupling Reactions and Their Applications in Natural Products and Designed Biomolecules Synthesis. *Chem. Rev.* **2008**, 108, 3054-3131.

28.Korch, K. M.; Watson, D. A., Cross-Coupling of Heteroatomic Electrophiles. *Chem. Rev.* **2019**, 119, 8192-8228.

29.West, M. J.; Fyfe, J. W. B.; Vantourout, J. C.; Watson, A. J. B., Mechanistic Development and Recent Applications of the Chan–Lam Amination. *Chem. Rev.* **2019**, 119, 12491-12523.

30.Hughes, D.; Wheeler, P.; Ene, D., Olefin Metathesis in Drug Discovery and Development—Examples from Recent Patent Literature. *Organic Process Research & Development* **2017**, 21, 1938-1962.

31.Korotchenko, V. N.; Nenajdenko, V. G.; Balenkova, E. S.; Shastin, A. V., Olefination of carbonyl compounds: modern and classical methods. *Russian Chemical Reviews* **2004**, 73, 957-989.

32.Zajc, B.; Kumar, R., Synthesis of Fluoroolefins via Julia-Kocienski Olefination. *Synthesis* **2010**, 2010, 1822-1836.

33.Caro-Diaz, E. J. E.; Urbano, M.; Buzard, D. J.; Jones, R. M., C–H activation reactions as useful tools for medicinal chemists. *Bioorg Med Chem Lett* **2016**, 26, 5378-5383.

34.Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Pipko, S. E.; Konovets, A. I.; Sadkova, I. V.; Tolmachev, A., Sulfonyl Fluorides as Alternative to Sulfonyl Chlorides in Parallel Synthesis of Aliphatic
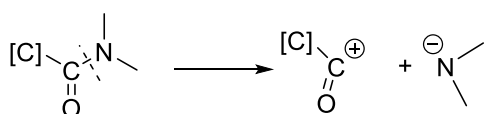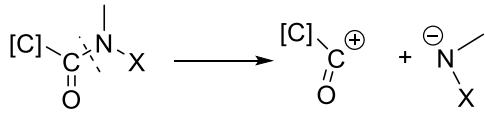
Sulfonamides. *ACS Combinatorial Science* **2014**, 16, 192-197.

35. Kuttruff, C. A.; Haile, M.; Kraml, J.; Tautermann, C. S., Late-Stage Functionalization of Drug-Like Molecules Using Diversinates. *ChemMedChem* **2018**, 13, 983-987.

36. Proctor, R. S. J.; Phipps, R. J., Recent Advances in Minisci-Type Reactions. *Angew. Chem. Int. Ed.* **2019**, 58, 13666-13699.

37. Smith, J. M.; Harwood, S. J.; Baran, P. S., Radical Retrosynthesis. *Acc. Chem. Res.* **2018**, 51, 1807-1817.

38. Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A., Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **2018**, 10, 383-394.

39. Pennington, L. D.; Aquila, B. M.; Choi, Y.; Valiulin, R. A.; Muegge, I., Positional Analogue Scanning: An Effective Strategy for Multiparameter Optimization in Drug Design. *J. Med. Chem.* **2020**, 63, 8956-8976.

40. Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P., Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015**, 20, 11-17.

41. Goodman, J., Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, 49, 2897-2898.

42. Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; American Chemical Society: 2014; Vol. 1164, Chapter 8, pp 127-148.

43. Moo, Y. N., Ryune; LEE, Dae, Won; LEE, Ju, Young; KIM, Hui, Ho; LEE, Dong, Ho; , Method for preparation of carbamic acid (r)-1-aryl-2-tetrazolyl-ethyl ester. *World Intellectual Property Organization* **2010**, Patent number: WO2010/150946; A1.

44. Shivanyuk, A.; Ryabukhin, S. V.; Bogolubsky, A. V.; Tolmachev, A., Enamine REAL database: making chemical diversity real. *Chimica Oggi-Chemistry Today* **2007**, 58-59.

45. Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, 23, 101681.

46. Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J., The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, 56, 1253-1266.

47. Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H., Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, 49, 270-279.

Summary

The Synthons Interpreter, or SynthI, is a new open-source toolkit for library design and BB analysis. It uses the synthons-based representation to connect BBs and fragments formed via pseudo-retrosynthetic fragmentation of bigger molecules according to 38 bond disconnection reaction rules (**Table 7**). Their application produces a set of synthons that can be traced back to 150 different types of BBs, thanks to the presence of specific labels. Synthons can be treated as genuine compounds thanks to a labeling scheme developed here. Its main advantage is that it encodes the position and chemical nature of the reactive centers while maintaining structure validity. This approach not only enables the design of synthetically accessible libraries but also allows MedChem relevant analysis of BBs.

SynthI was used to classify, filter, and analyze Enamine in-stock BB library. The list of recently approved drugs was used as a case study for compound fragmentation. The comparison of SynthI-Fragmentation with literature reported experimentally validated synthetic pathways demonstrated that they go into the correspondence in most cases, except when heterocyclizations are prevailing reactions (these reactions have not been implemented into the first release of SynthI).

Table 7. SynthI reaction rules specification.

R1 – N-acylation	
R1.1 - Amine acylation	
R1.2 – N-Acylation of RN-X compounds	*(hydrazides, sulfonylacetamides, substituted acetyl isocyanides, N-hydroxyamides, N-Acetylguanidines)*

R1.3 - N-Acylation by O=C(+)-X reagents (except isocyanates - R1.4)	(*carbamates, carbamide halogenides, substituted sulfanylformamide*)
R1.4 - Amine acylation by isocyanates or analogues	
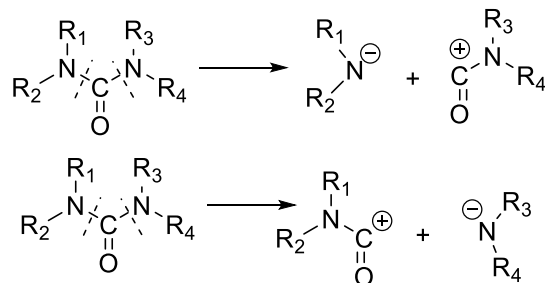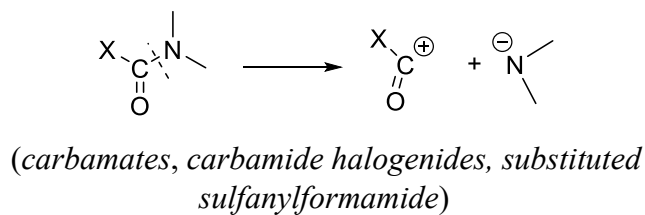
R2 - O-acylation

R2.1 Alcohol/Phenol acylation	
R2.2 O-Acylation by O=C(+)-X reagents	
R2.3 O-Acylation of O-X compounds	
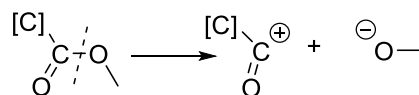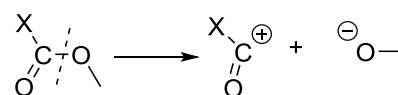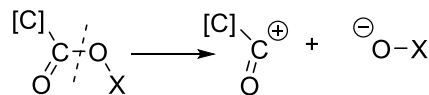
R3 Amine_alkylation_arylation

R3.1 - SN alkylation of amines;	$\ce{N-[C_{sp}^3]} \Longrightarrow \ce{N^-} + [C_{sp}^3]^+ $
R3.2 - Buchwald-Hartwig amination(BHA), Cu-mediated C-N coupling;	$\ce{N-[C_{sp}^2]} \Longrightarrow \ce{N^-} + [C_{sp}^2]^+ $ aromatic
R3.3 Umpolung cross-coupling	$\ce{N-[C_{sp}^2]} \Longrightarrow \ce{N^+} + [C_{sp}^2]^- $ aromatic
R3.4 Tertiary amines alkylation arylation	$-\overset{+}{N}-[C_{sp}^3] \longrightarrow -\ce{N^-} + [C_{sp}^3]^+ $
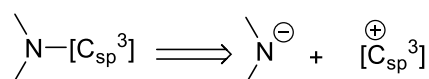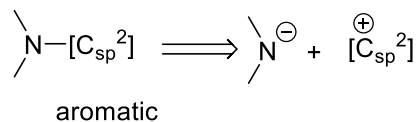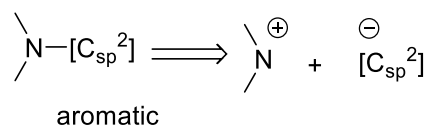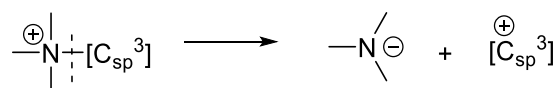
R4 - O-alkylation_arylation
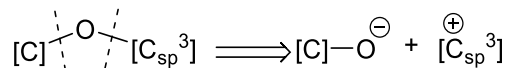
R4.1 - SN alkylation	$[C]\!\!-\!\!O\!\!-\!\![C_{sp}^{3}] \Longrightarrow [C]\!\!-\!\!O^{\ominus} \;+\; [\overset{\oplus}{C}_{sp}^{3}]$
R4.2 - Cu-mediated C-O coupling	$[C]\!\!-\!\!O\!\!-\!\![C_{sp}^{2}] \Longrightarrow [C]\!\!-\!\!O^{\ominus} \;+\; [\overset{\oplus}{C}_{sp}^{2}]$ aromatic
R4.3 - Chan-Evans-Lam coupling	$[C]\!\!-\!\!O\!\!-\!\![C_{sp}^{2}] \Longrightarrow [C]\!\!-\!\!O^{\ominus} \;+\; [\overset{\ominus}{C}_{sp}^{2}]$ aromatic only Boron cmpds
R4.4 -N-O-alkylation	$[N]\!\!-\!\!O\!\!-\!\!C(Alk) \Longrightarrow [N]\!\!-\!\!O^{\ominus} \;+\; \overset{\oplus}{C}(Alk)$
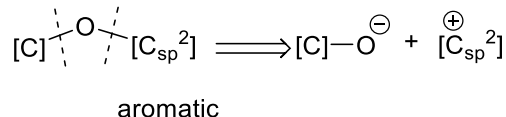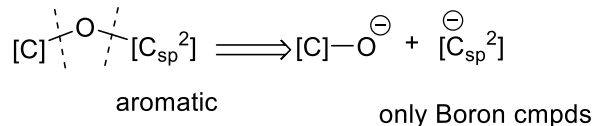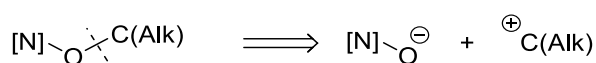
R5 - Alkylation_arylation_of_NH-heterocycles

R5.1 - SN alkylation;	triazole-$N\!-\![C_{sp}^{3}] \Longrightarrow$ triazole-$N^{\ominus} \;+\; [\overset{\oplus}{C}_{sp}^{3}]$
R5.2 - Chan-Evans-Lam coupling	triazole-$N\!-\![C_{sp}^{2,\,sp3}] \Longrightarrow$ triazole-$N^{\ominus} \;+\; [\overset{\ominus}{C}_{sp}^{2,\,sp3}]$ aromatic or vinyl only Boron cmpds
R5.3 - Cu-mediated C-N coupling	triazole-$N\!-\![C_{sp}^{2}] \Longrightarrow$ triazole-$N^{\ominus} \;+\; [\overset{\oplus}{C}_{sp}^{2}]$ aromatic or vinyl
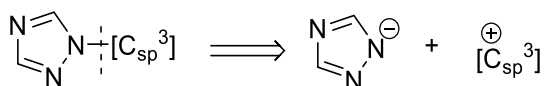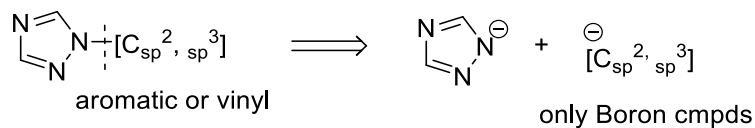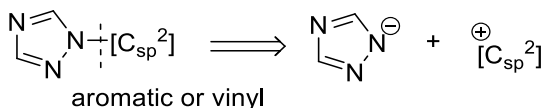
R6 - Alkylation_arylation_of_NH-lactam
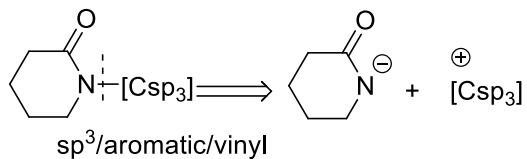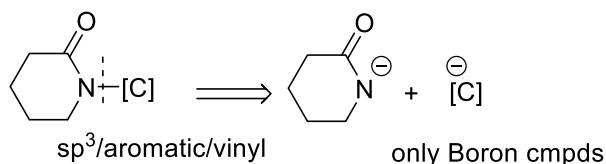
R6.1 - SN alkylation	lactam-$N\!-\![Csp_{3}] \Longrightarrow$ lactam-$N^{\ominus} \;+\; [\overset{\oplus}{C}sp_{3}]$ sp^{3}/aromatic/vinyl
R6.2 - Chan-Evans-Lam coupling	lactam-$N\!-\![C] \Longrightarrow$ lactam-$N^{\ominus} \;+\; [\overset{\ominus}{C}]$ sp^{3}/aromatic/vinyl only Boron cmpds
R6.3 - Cu-mediated C-N coupling	lactam-$N\!-\!Ar \Longrightarrow$ lactam-$N^{\ominus} \;+\; \oplus Ar$
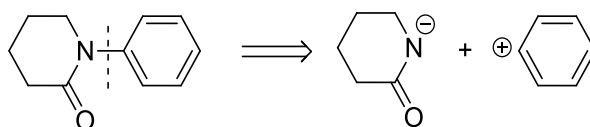
R7 – Amines sulphoacylation	
R7.1- S-alkylation arylation	$[C]\diagdown_S\diagup[C] \implies [C]\diagdown_S{}^{\ominus} + {}^{\oplus}C(Alk, (Het)Ar)$
R7.2 - Simple alkylation of sulphinic acid salts; R7.3 - Cu-catalyzed arylation of sulphinic acid salts	$[C]\diagdown_{\underset{O}{\overset{O}{S}}}\diagup[C] \implies [C]\diagdown_{\underset{O}{\overset{O}{S}}}{}^{\ominus} + {}^{\oplus}C(Alk, (Het)Ar)$
R8 - Amine_sulphoacylation	$\diagdown N-\underset{O}{\overset{O}{S}}-R \implies \diagdown N{}^{\ominus} + {}^{\oplus}\underset{O}{\overset{O}{S}}-R$
R9 - Condensation_of_Y-NH2_with_carbonyl_compounds	$[C]\diagdown_{\underset{N}{\overset{[C,H]}{C}}}-[C,N,O] \implies [C]\diagdown_{\overset{[C,H]}{C}{}^{\oplus}} + {}^{\ominus}N{-}[C,N,O]$
R10 - Metal organics C-C bong assembling	
R10.1 - Addition of Li, Mg, Zn organics to aldehydes and ketones	$[H,C]\diagdown_{\underset{[C]}{\overset{OH}{C}}{-}[H,C]} \implies [H,C]\diagdown_{\overset{OH}{C}{}^{\oplus}}{}[H,C] + {}^{\ominus}[C]$
R10.2 - Acylation of Li, Mg, Zn organics	$[H,C]\diagdown_{\underset{[C]}{\overset{O}{C}}} \implies [H,C]\diagdown_{\overset{O}{C}{}^{\oplus}} + {}^{\ominus}[C]$
R11.1 - Knovenagel-, Wittig-, Julia-Kocienski- type reactions,	$\diagdown C \vdots C \diagup \longrightarrow \underset{R_2}{\overset{R_1}{C}}{}^{\oplus}_{\oplus} + {}^{\ominus}\underset{R_4}{\overset{R_3}{C}}{}^{\ominus}$ $\diagdown C \vdots C \diagup \longrightarrow \underset{R_2}{\overset{R_1}{C}}{}^{\ominus}_{\ominus} + {}^{\oplus}\underset{R_4}{\overset{R_3}{C}}{}^{\oplus}$
R11.2 - Olefin Metathesis	$\underset{[H]}{\overset{[C]}{C}}{=}\underset{[C]}{\overset{[C,H]}{C}} \implies \underset{[H]}{\overset{[C]}{C}}: + :\underset{[C]}{\overset{[C,H]}{C}}$
R12 - C-C couplings	
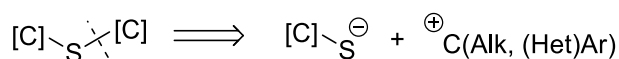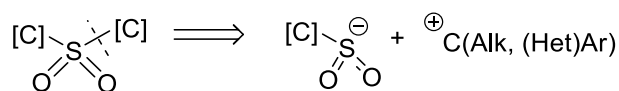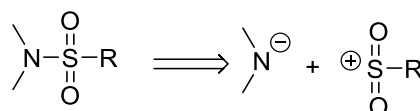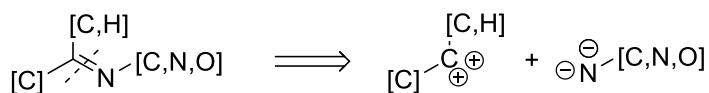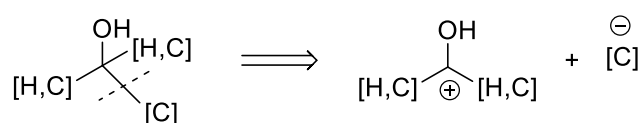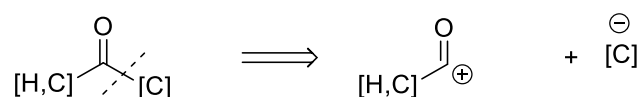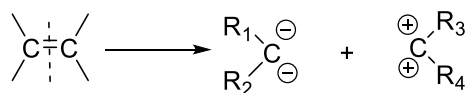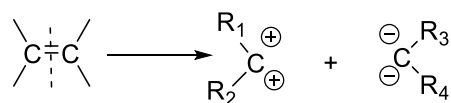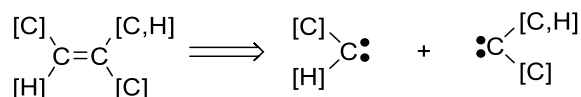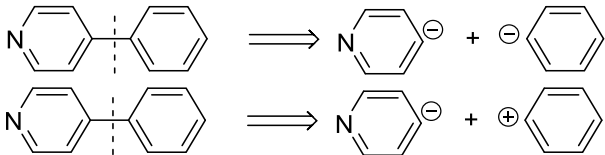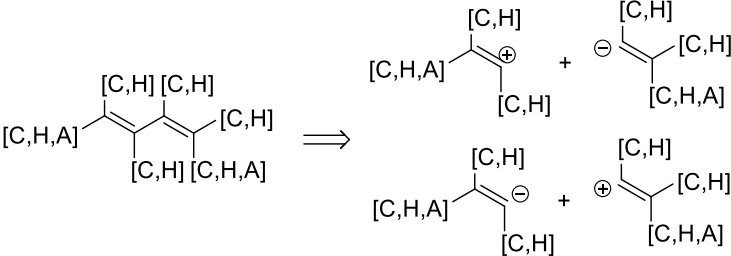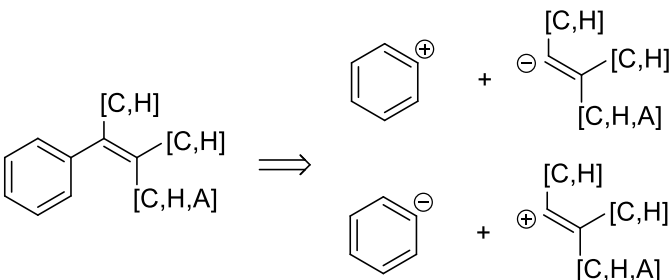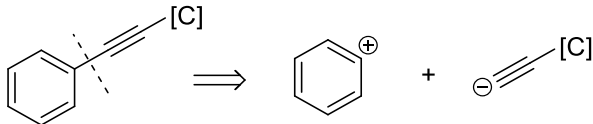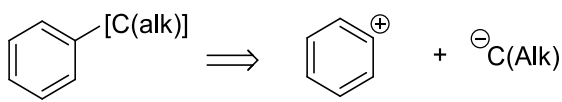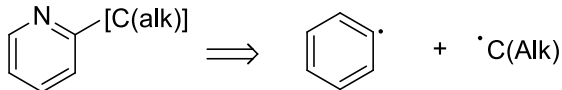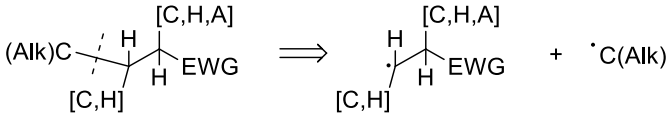
R12.1 - Suzuki cross-coupling C(Ar)- C(Ar)	
R12.2 - Suzuki coupling C(sp2) - C(sp2)	
R12.3 - Heck and Suzuki coupling C(Ar) - C(sp2)	
R12.4 - Sonogashira coupling C(Ar) - C(sp)	
R12.5 - Novel methods for C(Ar)-C(sp3) coupling	

R13 - Radical_reactions

R13.1 - Minisci reaction and Baran diversinates C(Ar)-C(sp3)	
R13.2 - Giese reaction C(sp3) - C(sp3)	EWG = CN, COR, CONH2, CONHR, CONR2, COOR, NO2, SO2R

4.4.2 A close-up look at the chemical space of commercially available building blocks for medicinal chemistry

Introduction

With the development of SynthI, analysis of the BBs libraries became more straightforward. The unified synthons representation allows not only to analyze BBs with MedChem bias but also to compare them to fragments obtained from the reference library. If ChEMBL library is used as a reference, such comparison allows to evaluate the biological relevance of purchasable BBs (PBBs) and their ability to face medicinal chemistry needs.

Thus, in this work, we present the first detailed analysis of the PBBs chemical space. The availability, rule-of-two-defined quality, diversity, and biological relevance of the main classes of BB were examined. The diversity of synthons has been analyzed using ISIDA fragment descriptors[89] that consider labeled connection points in synthons (former locations of the leaving groups). Thanks to that, it becomes possible to distinguish between BB that structurally differ only in terms of leaving groups and reactive center position. These descriptors were also used to define the chemical space of BB. For its visualization, a new universal map of synthons (synthons-uMap) was constructed by optimizing map performance in class separation for the different types of reactive centers present in synthons.

A close-up look at the chemical space of commercially available building blocks for medicinal chemistry

Yuliana Zabolotna[1], Dmitriy M.Volochnyuk[3,6], Sergey V.Ryabukhin[4,6], Dragos Horvath[1], Kostiantyn Gavrylenko[5,6], Gilles Marcou[1], Yurii S.Moroz[5,7], Oleksandr Oksiuta[3,7], Alexandre Varnek[2,1]*

Abstract: The ability to efficiently synthesize desired compounds can be a limiting factor for chemical space exploration in drug discovery. This ability is conditioned not only by the existence of well-studied synthetic protocols but also by the availability of corresponding reagents, so-called building blocks (BB). In this work, we present a detailed analysis of the chemical space of 400K purchasable BB. The chemical space was defined by corresponding synthons – fragments contributed to the final molecules upon reaction. They allow an analysis of BB physicochemical properties and diversity, unbiased by the leaving and protective groups in actual reagents. The main classes of BB were analyzed in terms of their availability, rule-of-two-defined quality, and diversity. Available BBs were eventually compared to a reference set of biologically relevant synthons derived from ChEMBL fragmentation, in order to illustrate how well they cover the actual medicinal chemistry needs. This was performed on a newly constructed universal generative topographic map of synthon chemical space, allowing to visualize both libraries and analyze their overlapping and library-specific regions.

Keywords: building blocks, synthons, library design, chemical space analysis, GTM

[1] University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France) *e-mail: **varnek@unistra.fr**

[2] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan

[3] Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine

[4] The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine

[5] Research-And-Education ChemBioCenter, National Taras Shevchenko University of Kyiv, Chervonotkatska str., 61, 03022 Kiev, Ukraine

[6] Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine

[7] Chemspace, Kyiv, Ukraine.

INTRODUCTION

The success of drug discovery strongly depends on the quality of the screening compounds. Starting molecules may be derived from natural sources or synthesized by organic chemists. Even though natural products have been evolutionarily selected to efficiently bind to biological macromolecules, they may not be easy to extract and purify on a large industrial scale. The pursuit of structural diversity with easily obtainable compounds led to the mutually dependent symbiotic relationships between drug discovery and organic synthesis[1].

Over the past decades, the chemical market has evolved to meet medicinal chemistry demands, with new compounds having medChem relevant physiochemical properties – low molecular weight and lipophilicity, high Fsp3, etc[2]. At the same time, it is well known that chemotype distribution in the

commercially available libraries of screening compounds is highly unbalanced towards synthetically accessible benzenesulfonamides, anilids and other amides, etc[3]. Beyond the immediately available "on shelf" collections, "tangible libraries" of easily accessible (but not yet produced) molecules were proposed[4]. They have emerged as the result of the stock enhancement campaigns, directed towards the overall improvement of collections' quality and novelty. However, they still tend to sample already overpopulated areas of the chemical space.[3] That means that current strategies of the commercial library enhancement do not provide a uniform chemical space sampling and thus there is an urgent need for their improvement.

One of the most efficient ways to do that consists of an early quality control via monitoring properties and novelty of used building blocks (BB) - reagents that participate in the synthesis of the final screening molecules. Usage of the medicinally relevant BB can significantly improve the quality of the designed compounds by preliminary focusing on substructures and properties that will ensure desirable activity and ADMETox profile of the potential drug candidates. Moreover, the introduction of the new BB will allow to explore underrepresented regions of the chemical space, potentially accessing diverse properties and bioactivities.

Even though this fact is widely recognized by medicinal chemists, the number of scientific reports, targeting quality analysis of the existing purchasable building blocks (PBB) and potential strategies for the corresponding libraries enhancement, is significantly lower than the ones concerning commercially available screening compounds. Within the last two decades, the latter has been evaluated in numerous medicinal chemistry publications[2, 3, 5-11]. At the same time, there are only a few works dedicated to BB used in medicinal chemistry.

Based on the AstraZeneca (AZ) five-year 'long strategic reagent initiative' F.W. Goldberg et al.[12] outlined general design principles for novel BB in order to maximize their impact on drug discovery projects. Besides, they listed the most popular types of BB, chosen by medicinal chemists from AstraZeneca for different drug design campaigns. In another study, Hartenfeller et al.[13] investigated the biological relevance of the chemical space spanned by 58 of the most popular organic chemistry reactions, based on a subset of the readily available BB (≈26 000). They have concluded, that established synthetic resources are well suited to cover selected biologically relevant compounds. However, the chosen reference subset was limited to only ≈62 000 compounds from GVK-BIO[14], Drug Bank[15] and TIMBAL[16], which might fall short as a comprehensive representation of all known biologically active compounds.

Moreover, the analysis of all PBB was beyond the scope of both mentioned papers. To our best knowledge, the only report of such analysis is a price-focused study of almost one million PBB from 121 vendors, published by T.Kalliokoski[17]. In this work he analyzed the availability of the 13 types of BB, reporting a number of reagents available for purchase under a specific range of price up to $150/g. However, even though all these reports provide an important insight about the PBB libraries and some of the medicinal chemistry relevant properties, those articles, each being published at least five years ago, can hardly characterize the current state of the quickly growing chemical space of the PBB.

Therefore, in this work, we present the analysis of the to-date PBB set, addressing the availability of the most popular classes of BB, their diversity, and their ability to face current medicinal chemistry needs in the synthesis of biologically relevant compounds. As a source of PBB in-stock database of the biggest BB aggregator, eMolecules Inc.[18] has been used. For BB analysis, we have employed the previously reported freely available python library – Synthons Interpreter (SynthI) – knowledge-based reaction toolkit for the library analysis and design[19]. It allows examining BB not as individual chemical entities but as a set of synthons – fragments obtained after leaving groups removal/transformation with a system of labels that encodes position and type of reactive center (RC). They define the substructure that will contribute to the final molecule upon different reactions (except heterocyclization, omitted in this analysis). The same tool has been used for fragmenting compounds from ChEMBL[20] in order to detect synthons and, if available, corresponding BB required for the synthesis of the biologically relevant molecules from this database.

The diversity of synthons has been analyzed using marked-atom ISIDA fragment descriptors[21] that consider the marked connection points in synthons (former locations of the leaving groups). Thanks to that, it becomes possible to distinguish between BB that structurally differ only in terms of leaving groups and RC placement. These descriptors were also used to define the chemical space of BB, which was visualized via Generative Topographic Mapping (GTM)[22]. This non-linear visualization method has proven multiple times to be effective in the analysis of large chemical

databases[3, 23-27]. However, it is the first time it was used to map the space of synthons.

DATA

The 489 781 building blocks, provided by eMolecules, Inc.[18], have been used as a source of readily available PBB. Unique chemical structures within Tier 1 or 2 (corresponding to shipments within times of 5 and 10 days respectively) were selected to represent in-stock compounds.

ChEMBL (version 26) served as a reference dataset for biologically relevant molecules. 1 950 765 compounds have been standardized according to the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics (infochimie.u-strasbg.fr/webserv/VSEngine.html), using the ChemAxon Standardizer. That included:

- dearomatization and final aromatization (heterocycles like pyridone were not aromatized);
- conversion to canonical SMILES;
- salts and mixture removal; neutralization of all species, except nitrogen (IV);
- major tautomer generation
- stereochemical information removal.

Stereochemical information has been ignored because used ISIDA descriptors[21] would not capture it, anyway. Remaining after standardization, 1 721 155 unique ChEMBL compounds were then fragmented in order to obtain biologically relevant synthons. The resulting synthons, as well as synthons generated from eMolecules library, were standardized according to the same procedure.

METHODS

Synthons Interpreter (SynthI)

Considering that a single BB can contribute different structural motifs to the molecule, depending on the synthesis conditions and reaction partners, it is not useful to analyze primary chemical structures of the BB in the context of their usage in medicinal chemistry. Different protective and leaving groups can constitute a large (sometimes the largest) part of the reagent. Synthons, by contrast represent the substructure of a BB that will be inherited by the product, annotated by marks on the atoms that will connect to partner synthons. In our previous work, we have developed a python library - Synthons Interpreter or SynthI, for synthon generation from either BBs or drug-like

products, by RECAP-based fragmentation.[19] It consists of four modules, three of which were used in this work:

1. **SynthI-Classifier** consists of the library of smarts identifying structural motifs required and respectively forbidden in BB suitable as particular class of reagents required by the considered set of chemical reactions. For now, this set only includes coupling reactions (no heterocyclizations). These involve 22 generic monofunctional reagent classes, like acyl halides, boronics, ketones, primary amines, etc. These can be further subdivided into about 100 finer subclasses of significantly diverging reactivities. For example, class "Alcohols" includes three subclasses of reactivity – "Heterols", "Aliphatic alcohols" and "Phenols". In addition, there are 28 bifunctional and 19 trifunctional classes.

2. **SynthI-BB** allows to generate exhaustively the most probable synthons from a given BB – a process herein referred to as "synthonization". The position of the functional groups, as well as the formal type of the resulting fragment (electrophilic, nucleophilic, radical, etc.), is encoded as synthon SMILES with class-specific numeric marks on the "connecting" atoms with formal free valences (allowing to be coupled to partner synthons). There are 9 types of reactive centers (RC) that can appear in synthons:

- **electrophilic** (produced by acyl and aryl halides, acids, aldehydes, ketones, etc.);
- **nucleophilic** (alcohols, thiols, amines, metal organics, hydrazines, hydrazides etc.);
- **bivalent electrophilic** (aldehydes and ketones);
- **bivalent nucleophilic** (primary amines, hydroxylamines, reagents for olefination, etc.);
- **bivalent neutral** (terminal alkenes for metathesis);
- **electrophilic radical** (Minisci CH-partners, Michael acceptors);
- **nucleophilic radical** (BF3 and MIDA boronates, NOPhtal alkyl esters, sulphinates, etc.);
- **boronics-derived nucleophilic** (boronic reagents);
- **electrophilic nitrogen** (benzoyl O-acylated hydroxilamines).

The resulting synthons, represented by ISIDA descriptors, were used to define chemical space of commercially available BBs. The type of ISIDA fragments was selected during synthons-uMap optimization.

3. **SynthI-Fragmentation** was used in order to evaluate the ability of current PBBs space to face medicinal chemistry needs via ChEMBL molecules fragmentation. ChEMBL database has been chosen as the best representation of the biologically relevant chemical space. In SynthI-Fragmentation, the algorithm fragments molecules in all possible ways according to the specified list of reactions and then select the most optimal fragmentation scheme in a way to maximize number of synthons that correspond to at least one BB from the user-provided library (in our case PBB from eMolcules library). Parts of the molecules not covered by PBB synthons were broken down to the smallest possible synthons. They can be used as inspiration for enhancement of PBB collections.

Synthon quality assessment

According to the "rule of two" (Ro2)[12], good quality BB for medicinal chemistry could be defined as those that typically do not add more than 200 Da in MW, 2 units of clogP, 2 H-bond donors and 4 H-bond acceptors. Therefore, the synthons, as a fragments of BB that will be added to the final molecule, were filtered according to this rule and the number of BB compliant to it was assessed for each BB class

Diversity analysis

Diversity analysis of different types of reagents was also performed in synthon ISIDA descriptor space. It was done by calculating pairwise Tanimoto distance for all synthons within a selected reagent class, followed by the creation of the frequency plot for each of the diversity values. Note that a same introduced fragment may stem from distinct sytnhons, with RCs at different positions. The corresponding synthons will have distinct ISIDA descriptors in spite of being based on a same molecular graph, due to the marked-atom mechanism. Two synthons contributing the same fragment and having the RC at the same position, but of different type (allegedly different reaction mechanisms) have however identical ISIDA descriptors (they capture the label position, not its actual value). Such synthons are distinct options covering the same medChem need – their existence is practically important because they allow for altenrative synthetic pathways, but they are indeed redundant from a structural point of view.

GTM

In chemoinformatics, chemical space can be defined by the N-dimensional molecular descriptor vector, where N is typically very large (10^2-10^4) for vectors designed to capture significant chemical information. The most intuitive way to analyze such a complex space is to reduce its dimensionality by projection of a human-readable 2D map. Generative topographic mapping (GTM) was first proposed by Bishop in 1998 [22] and appears as one of the most efficient methods of dimensionality reduction[28]. It performs non-linear projections of compounds from the initial multidimensional descriptor space to a 2D latent space - a manifold defined by a set of radial basis functions (RBF). The shape and position of each point of the manifold in the N-dimensional space are determined during its training – unsupervised fitting to the "frameset" items - molecules used to probe the chemical space of interest. Afterward, the manifold is unfolded back to the planar form – square grid 2D map.

Once trained, the manifold can host not only compounds of the "frameset" but also any external molecules, under the condition that in the multidimensional space they are residing close to the manifold (log Likelihood applicability domain of GTM[29]). The distinctive feature and the main advantage of GTM is its probabilistic nature, ensured by RBFs. In GTM molecules are not assigned to a particular point on the map. Instead, each molecule is fuzzily projected over the whole map with larger probabilities ("responsibilities") for nodes, situated closer to this compound in the initial space. Such smooth projection enables the creation of GTM landscapes – 2D plots of cumulated responsibilities, colored by average values of different properties, e. g. density, biological activity, physicochemical property, assigned class, etc. One manifold can host multiple landscapes allowing the analysis of multiple libraries according to different properties and also be used as a basis for building QSAR models[25, 28-30].

Universal map of syntons (synthons-uMap)

The "universal" map of synthons (synthons-uMap) is the GTM that would simultaneously host different types of synthons (electrophiles, nucleophiles, radicals, etc.). It can be constructed by optimizing map performance in class separation for the different types of reactive centers present in synthons.

A fixed frame set of 15 255 randomly selected synthons has been used. It contained an approximately equal ratio of synthons obtained by eMolecules in-stock BB library synthonization and ChEMBL fragmentation in order to span the chemical space of

both PBB and biologically relevant BB. Seven scoring sets, 15 000 synthons each, were used to evaluate map performance in class separation for electrophiles, nucleophiles, bivalent nucleophiles, bivalent electrophiles, neutral biradicals, electrophilic radicals and boronic-derived nucleophiles (for Chen-Lam reaction and couplings). The map was optimized, in exploring its (hyper)parameter space by an evolutionary procedure as customarily employed to tune GTMs[23, 31, 32], however following a Pareto-front-driven multiobjective strategy. This approach considered the 6x7/2=21 synthons class separation performances, expressed as balanced accuracies as independent objectives, and the Pareto front of non-dominated maps were considered as "best" solutions (defining the pool of selected individuals that were allowed to produce offspring in the evolutionary strategy).

RESULTS AND DISCUSSION

Availability of the main reagent classes and their quality

406 141 reagents out of 391 378 BB from eMolecules library have been classified and synthonized. The remaining non-classified reagents are either used in heterocyclization reactions that are out of the scope of this analysis or contain conflicting or competing functionalities disqualifying them for combinatorial chemistry.

As a result, 798 643 synthons were generated. In *Figure 1* one can see the detailed analysis of the availability of the monofunctional reagents on the market. The expected leaders of the distribution are amines, acids and aryl halides. Their "excessive" availability can be explained by wider usage of combinatorial reactions that employ this reagents. Among all classes of compounds, approximately half of them pass the Ro2, and thus represent the means for drug-like libraries synthesis.
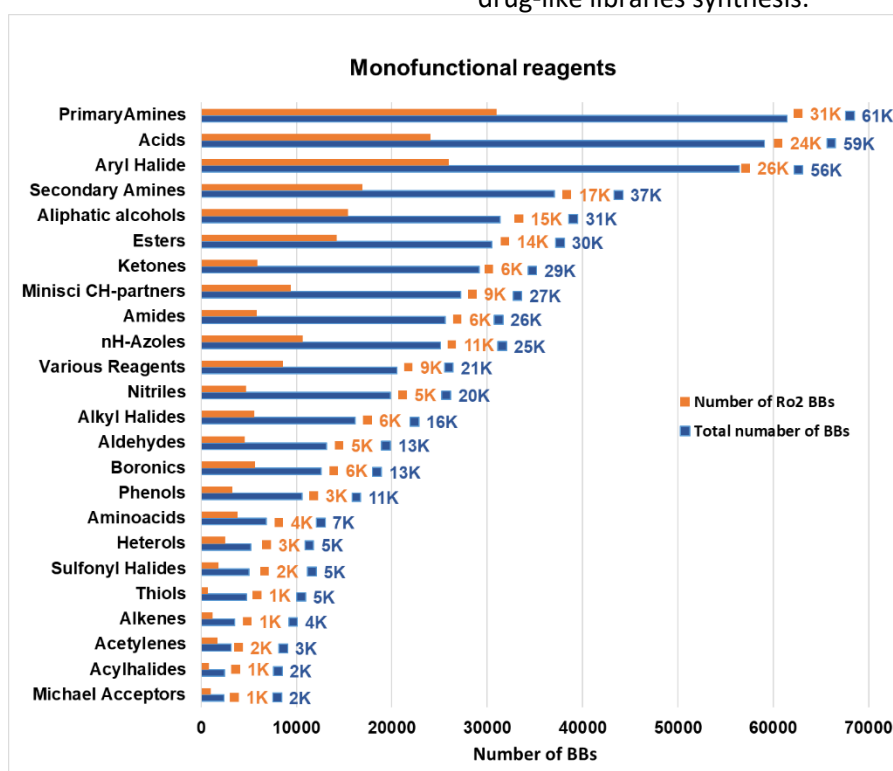
Figure 1. Monofunctional commercially available reagents: total number and number of high-quality Ro2 compliant reagents.
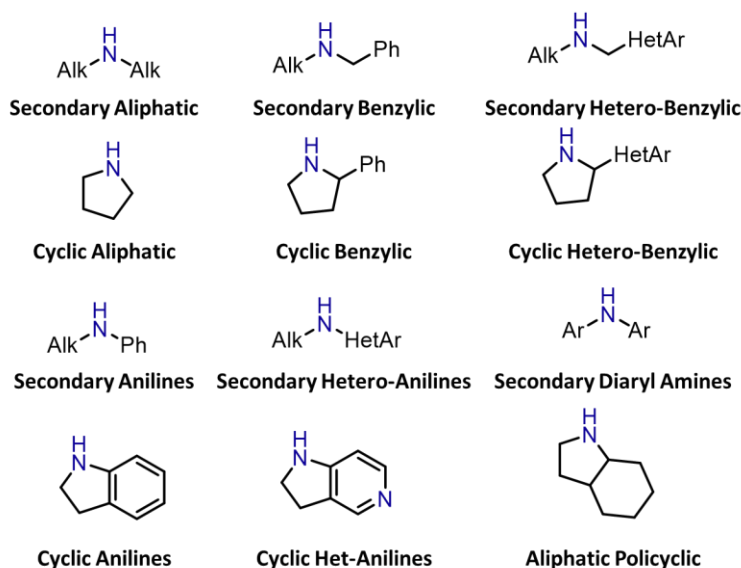
Figure 2. The schematic representation of different topologies for secondary amines.
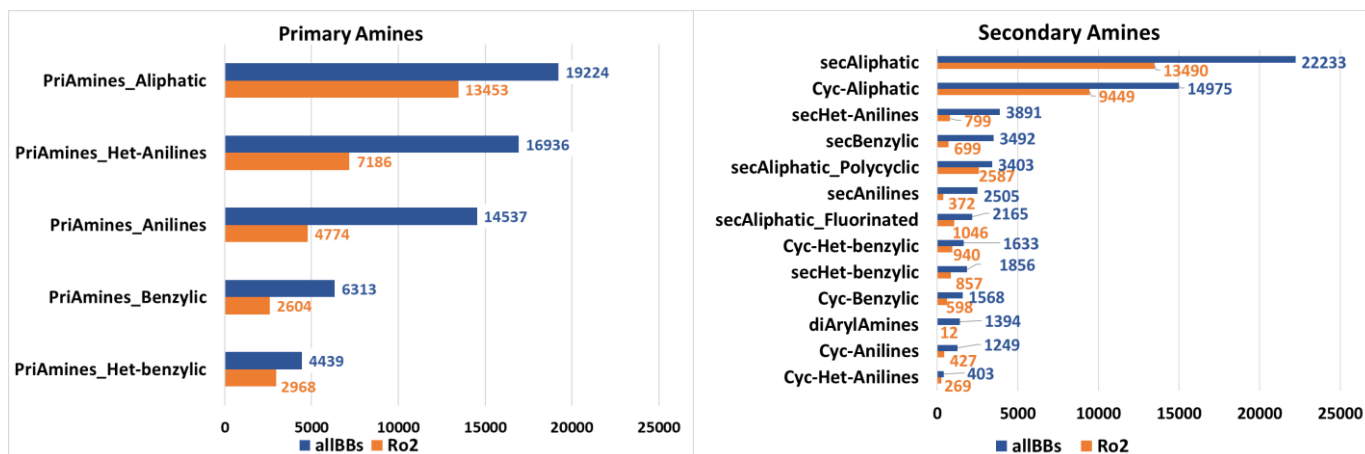
Figure 3. Availability of primary and secondary amines: total number and number of high-quality Ro2 compliant molecules.

Amines

Despite the strong development of modern organic synthesis ,medicinal chemists traditionally use only a tiny fraction of the available reactions, especially in compound library and analogs synthesis[33]. The general criteria for the ideal MedChem reactions were formulated in 2010 by GSK[34] and have not changed significantly over the last decade. Among them, there are requirements for reproducible chemical transformations, applicable to structurally diverse substrates, tolerance for the range of functionalities, simple equipment and reasonable cost. The reactions that fulfill these criteria, such as amides and sulfonamides formation, alkylations (including reductive amination), SNAr/Buchwald and C(Ar)-C(Ar) Suzuki couplings, will be always attractive to medicinal chemists. The majority of such reactions use primary and secondary amines as coupling partners, which explain their leading position on the market.

For more detailed analysis, primary amines have been split into several groups depending on the position of the functional group – aliphatic, benzylic, heterobenzilic amines, anilines and hetero-anilines. Secondary amines, however, can have even more different topologies (**Figure 2**).In both cases, aliphatic amines (cyclic and acyclic) are the most popular (**Figure 3**), which can be explained with current medicinal chemistry demand for the high Fsp3 compounds[35, 36]. Next are the derivatives of hetero-anilines and anilines, which allow one-stage introduction of new aromatic cycles.

Carboxylic acids

The second place on the market is taken by acids – the main coupling partners of amines. A recent study

from AZ indicates that amide couplings sum up to one-third of all the reactions in their electronic notebooks[37]. As one can see in **Figure 4,** similar to amines, among carboxylic acids the aliphatic counterparts are dominant. They are followed by heteroaromatic and benzoic acids. It should be noted that the homologs of heteroaromatic and benzoic acid - corresponding (hetero)aryl acetic acids - are significantly less present (from 7 to 10 times). It goes in accordance with the observation of AZ made in 2011, that synthetically this type of acids is much less accessible[38]. Indeed, out of 148 compounds proposed for the synthesis in AZ, only 17 were successfully made.

Figure 4. Commercially available carboxylic acids: total number and number of high-quality Ro2 compliant molecules.

Arylation reagents

The leading position of the aryl halides can be explained by the active development of the Pd-mediated Csp2-Csp2 and N-Csp2 couplings[39]. According to Boström's analysis, the Suzuki Csp2-Csp2 coupling is the second most popular transformation after the amide bond formation[33]. The same was later confirmed by Elli Lilly's analysis of the reactions performed using their ASL robotic synthesis system[40] and AbbVie's high-throughput chemistry department[41]. The high reproducibility of Csp2-Csp2 coupling together with its modern improvement made this reaction suitable for automation. In 2015 Burke designed a generalized automated process for the C-C couplings, by analogy with well-known automated peptide synthesis based on amide bond creation[42]. Despite such great achievements in Suzuki couplings the commercial accessibility of organoboron building blocks still significantly lower in comparison with (hetero)aromatic electrophiles (**Figure 1**).

Buchwald-Hartwig (BH) amination is also very popular. The power of this reaction lies in the ability to couple two fragments with minimal addition of rotatable bonds in the final structure. However, its success rate still hardly exceeds 45% due to the lack of a general catalytic system for diverse substrates. Besides, the reactivity in BH amination for the significant portion of available amines has not been experimentally validated yet and is hard to predict. At the same time, the active development of high-throughput experimentation (HTE) chemistry [43, 44] as well as machine learning approaches[45] significantly accelerates the identification of effective catalytic systems and the scope of their application.

The alternative well studied metal-free transition - "classical" SN_{Ar} amination cannot compete with the BH reaction. It appears that among all aryl halides only a limited fraction bears activated halogen atoms suitable for non-catalytic amination **(Figure 5)**. Interestingly, in the case of (hetero)aromatic chlorides, almost all of them (13 305 out of 14 697) bear activated chlorine atoms likely to undergo SN_{Ar} reactions. It could be explained by the fact that early conditions for the Suzuki coupling were inapplicable for the aromatic chlorides. However, the opposite situation is observed for aromatic bromides, which are convenient partners for the Suzuki couplings. Indeed, only 10% of aryl bromides are suitable for metal-free amination (3 664 out of 34 586). The number of bromides for the SN_{Ar} reaction is comparable with hetero(aromatic) compounds bearing an active fluorine atom (3 361), but the number of identified iodides (957) is significantly smaller.
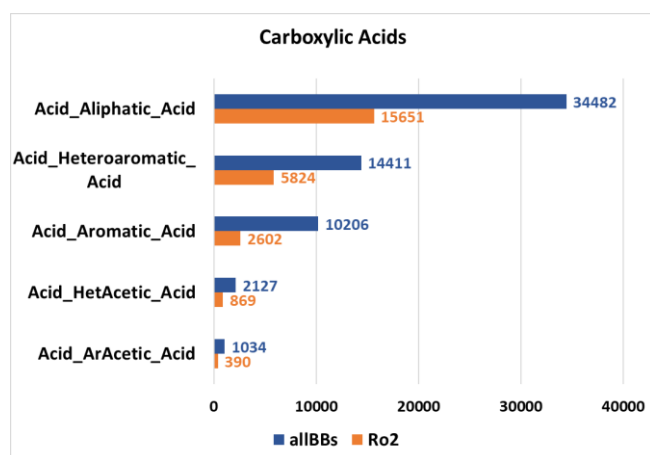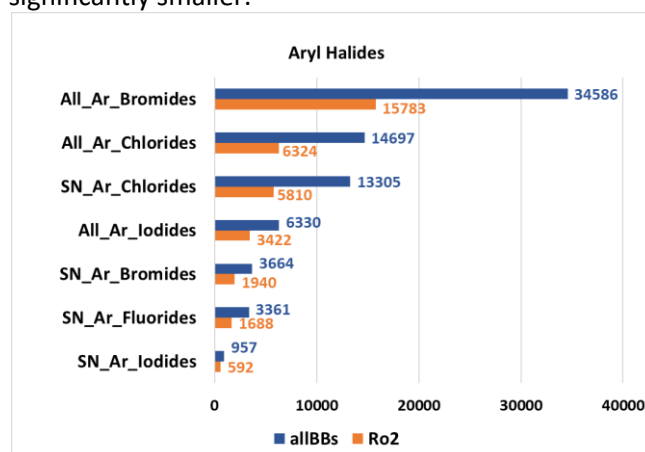

Figure 5. Commercially available aryl halides: total number and number of high-quality Ro2 compliant molecules.

Alkylation agents

The C(sp3)-N bond creation is also very popular and sum up to 10.6% of all reactions, performed in

industrial medicinal chemistry departments according to Vernalis statistics[46]. The alkylation or reductive amination is regularly used for that aim. Among these two reactions, the reductive amination is slightly more preferable[41, 47], because it is more selective and allows avoiding a significant number of by-products observed during alkylation. Nevertheless, this approach has its limitations, caused by the low diversity of the commercial carbonyl compounds. In the case of aldehydes (**Figure** 6), the most popular reagents are aromatic and heteroaromatic ones, generating benzylic type synthons. Aliphatic aldehydes are less represented, especially (het)aryl acetic ones due to their extremely low stability and high rate of self-condensation. Ketones are better represented in commercial catalogs, but there is still a lack of the most interesting for MedChem cyclic ketones (only 7 197 from which only 2 447 pass Ro2).
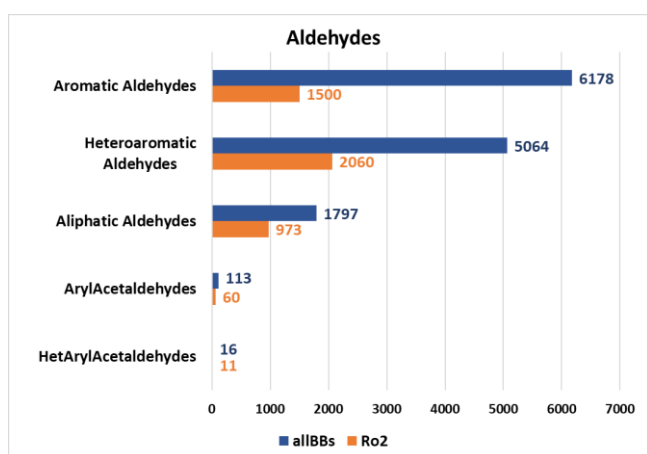
Figure 6. Commercially available aldehydes: total number and number of high-quality Ro2 compliant molecules.

Expanding the space of the synthons for alkylation could be achieved by commercially available alkyl halides. In ***Figure 7*** one can see that alkyl chlorides and

bromides are preferred over iodides and primary alkyl halides are significantly more accessible than secondary ones. (Hetero)benzylic primary alkyl halides (4 305) are less present in comparison with the corresponding aldehydes (11 242). Even higher difference is observed while comparing secondary halides (2 445 in total) and ketones (29 152). This can be explained by the lower shelf-life time of alkyl halides. Indeed, many of them are obtained from corresponding alcohols prior to synthesis. Moreover, nowadays efficient methods for the in situ alkylating agent generation (including chlorides, bromides and iodides) were developed. For example, recently SO_2F_2-mediated in situ generation of 1° and 2° alkyl halides was proposed[48].

Other very efficient alkylating reagents - sulfonate esters, like mesylates, tosylate and triflate also have low shelf-life time. This makes their precursors, alcohols, more attractive for purchase and storage as latent alkylators. There are also ongoing attempts to develop direct methods for the alkylation of amines with alcohols. Among them, there are development of the advanced reaction conditions for the well-known Mitsunobu reaction[49] that allows basic amine usage[50] and a novel Ru-based catalyst system for hydrogen borrowing reaction, proposed by GSK in 2009[51]. Therefore, it is not surprising, that representation of this reagent class on the market is comparable with secondary amines. In ***Figure 7*** one can see a more detailed analysis of different aliphatic alcohols topologies. In contrast to alkyl halides, there are approximately the same number of primary and secondary alcohols with a slight excess of the latter, while the number of benzylic and heterobenzylic alcohols is comparable to corresponding alkyl chlorides.
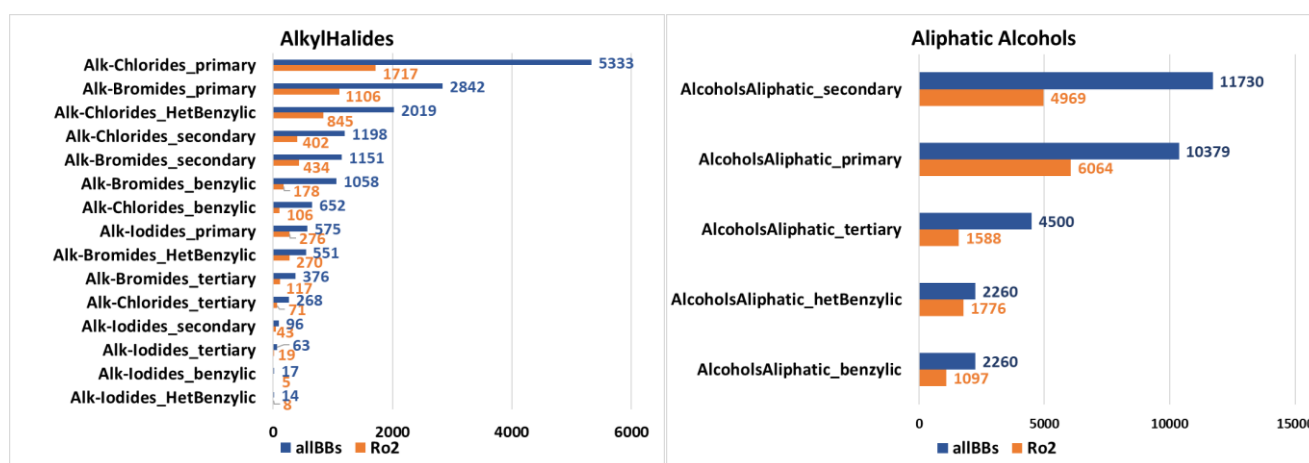
Figure 7. Availability of alkylation agents – alkyl halides and aliphatic alcohols: total number and number of high-quality Ro2 compliant molecules.

Sulfur-containing BB

Surprisingly, despite the high popularity of sulfonamides in medicinal chemistry, the number of available sulfonyl chlorides and fluorides is rather low. In *Figure 8* one can see that the leading position among them is taken by aryl sulfonyl chlorides, which can be explained by their higher stability in long-term storage in comparison to alkyl and heteroaryl sulfonyl chlorides, that can undergo SO2 extrusion. The seminal paper addressing this stability issue was published by Pfizer in 2006[52]. It was also shown that sulfonyl fluorides can become a convenient replacement of sulfonyl chlorides, as they are more thermodynamically stable, resistant to reduction and chemoselective towards sulfonylation products. However, such an approach has not gained attention until the introduction of Sulfur(VI) Fluoride Exchange (SuFEx) reaction for clic chemistry by Sharpless et al in 2014[53]. Since then numerous works have been published on synthesis and usage of SuFEx building blocks[54-58]. However, as one can see in *Figure 8*, the market did not have enough time to react to the newly emerged tendency and there are only limited number of such reagents available yet.

Another unexpected observation is that the total number of thiols on the market is rather low, even though S-alkylation is one of the most well-studied reactions in combinatorial chemistry. This can be explained by the complicated storage conditions, required for these reagents. Since thiols can easily undergo oxidation and form disulfides they should be stored in ampules with an inert atmosphere. The heteroaromatic thiols are the most populated group (*Figure 8*), as a result of their additional stability gained via thione-thiol tautomerism.

Other reagents

The above-mentioned tendency in the late-stage combinatorial reactions popularity is indirectly proved by recent statistics published by Pfizer. In a course of its Quick Building Blocks program out of all BBs they have used 29% of acids amine – 21%; alcohol - 9%; aryl halide - 9%; mono-BOC diamine - 6% ; aniline - 5%; aldehyde - 4%; aryl boronic acid - 4% and sulfonyl chlorides only 3%.[59]. At the same time, there are also less represented classes of reactions and reagents that are widely used for larger BBs synthesis in the early stages of the synthetic pathway. In *Figure 9* one can see that among various reagent classes the most numerous are hydrazides and hydrazines. Iso(thio)cyanates, hydroxylamines and element-organics occupy the middle position. Among metallorganics, Grignard reagents expectedly are the most numerous class. Organozinc BBs account for two times fewer compounds and there are only 6 Li-containing reagents.
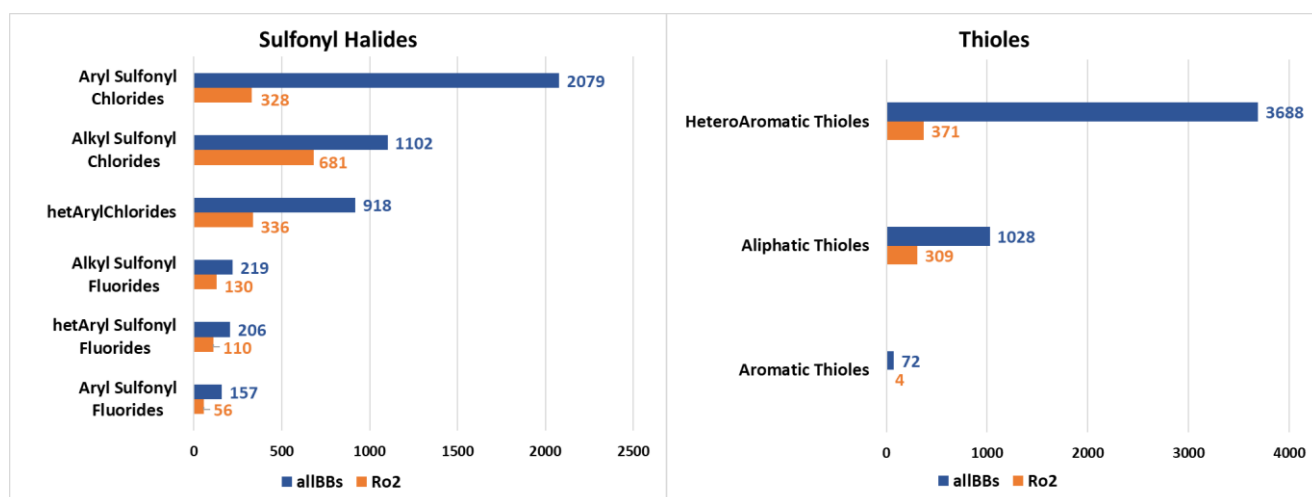
Figure 8. Availability of sulfonyl halides and thiols: total number and number of high-quality Ro2 compliant molecules.
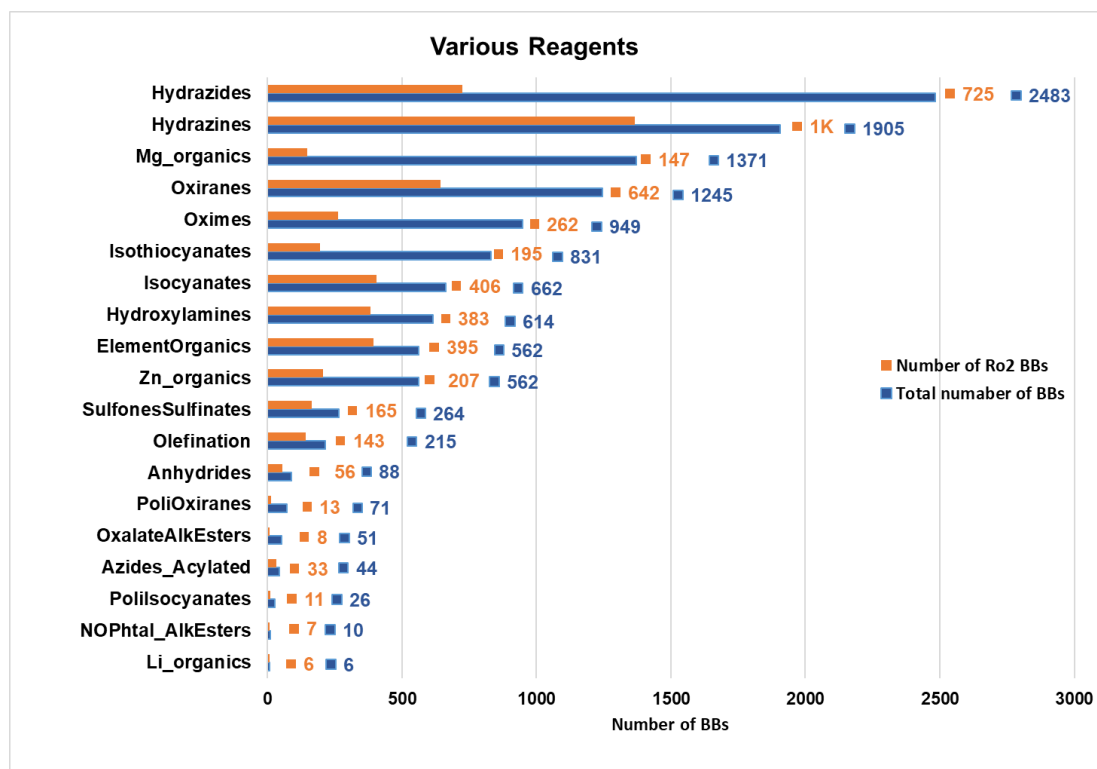
Figure 9. Commercially available various reagents: total number and number of high-quality Ro2 compliant molecules.

Reagents for novel reactions

Some novel synthetic methodologies became more frequently used by medicinal chemistry. This was influenced by the emergence of new methodologies, instruments and techniques for elaborating and carrying out combinatorial reactions. Among them, we can highlight the automatic optimization of the reaction conditions at nanomolar scale[43], a robustness screen which allows to quickly determine the scope of the application of substrates with additional functional groups[60], new selective radical processes[61], photoredox catalysis[62], a new generation of click chemistry[63, 64], automated interactive cross-coupling[42], and late-stage functionalization[65]. However, the above-mentioned distribution of BBs hinders the development of such novel combinatorial reactions due to the poor representation of necessary reagents among the commercially available. One such example is a Minisci reaction, which efficacy is based on an effective in-situ generation of free radicals[66].

As pointed in **Figure 10**, there are many types of reagents that can be used as a source of free radicals. Nevertheless, only some of them, such as R–SO$_2$F, R-SO$_2$H salts, RCOONPhtal, R-BF$_3$K and R-BMIDA are useful for the combinatorial synthesis of compound libraries, which could fill the "white spots" in chemical space. So, despite sufficiently numerous (27 204 BBs) 'Minisci_CHpartners' (**Figure 1**) the implementation of

this reaction in a combinatorial manner is limited by scarcity of coupling partners – there are only 716 R-BF$_3$K$^+$ and R-BMIDA (418 of them are Ro2 compliant) and even less R–SO$_2$F (**Figure 8**) and RCOONPhtal esters (**Figure 9**). Thereby chemical companies could focus their attention on this problem, and direct their efforts towards BB catalogues enhancement with such needful reagents.
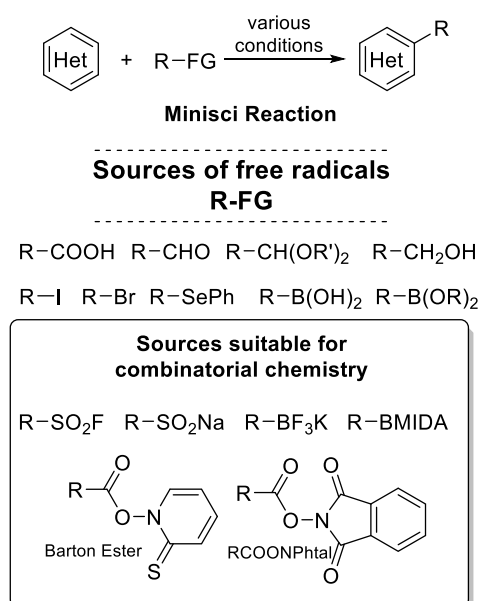
Figure 10. Representative types of reagents for Minisci Reaction.

Polyfunctional BB

Apart from the monofunctional reagents, appropriately protected bi- and trifunctional BB are required for optimal combinatorial library construction. Among the bifunctional ones, the absolute leaders are different derivatives of amino acids (*Figure 12*) due to the extreme popularity and automation of peptide synthesis. Other large classes are Boc-protected diamines and functional aryl halides. Such distribution reflects the same tendencies that have been observed and explained for monofunctional building blocks. Polyfunctional reagents are playing the role of molecular cores around which a diverse set of monofunctional partners allows the creation of large combinatorial libraries. Therefore, bi- and especially trifunctional BBs are crucial for the synthesis of DNA-encoded libraries (DEL), and thus their availability is affected by the popularity and efficiency of the reactions, adapted for this technology. Considering the rather recent development of DEL, a limited number of corresponding reagents on the market is understandable.

Medicinal chemists' highlights

Earlier in this analysis, the main focus was set on functional group types that define the BB that may be successfully used in a reaction. However, what is even more important for medicinal chemists is what structural moieties will be introduced and how these will influence the pharmacodynamics or pharmacokinetic properties of the synthesized compound. Considered motifs emerge from "breakthrough" approvals of a new drug containing unusual structural moieties. They include morpholine and piperazine bioisosters[67-69], unusual fluorine-containing aliphatic substituents,[70, 71] sulfoximines,[72] phosphine oxides[73], silicon-containing isosteres[74] and non-classical sp3-enriched benzene isosteres, such as bicyclo[1.1.1] pentanes, cubanes, etc.[75] In *Figure 11* one can see that there are only a limited number of BBs bearing such structural motifs. The distribution leader is morpholine and piperazine mimetics, oxetames, and sultames, while there are less than a hundred cubanes, disubstituted bicylo[2.1.1]hexanes, and silicon-containing BB.
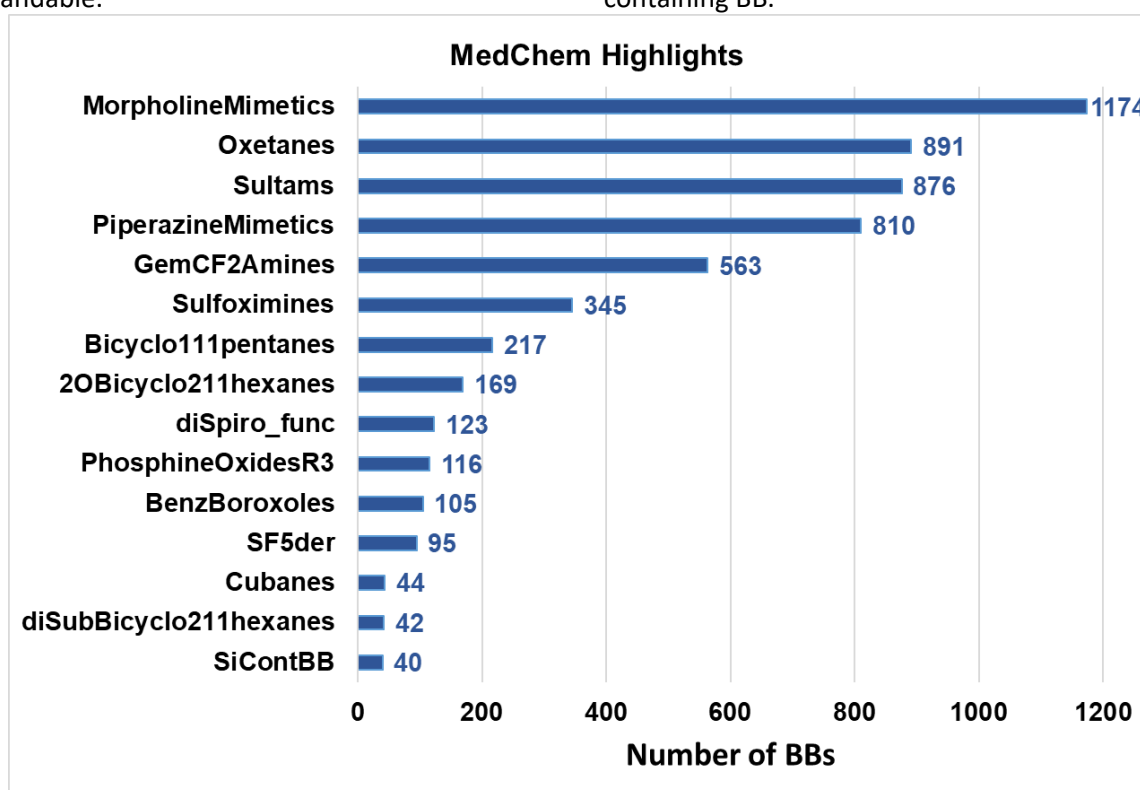
Figure 11. Commercially available reagents, containing highly attractive structural motifs
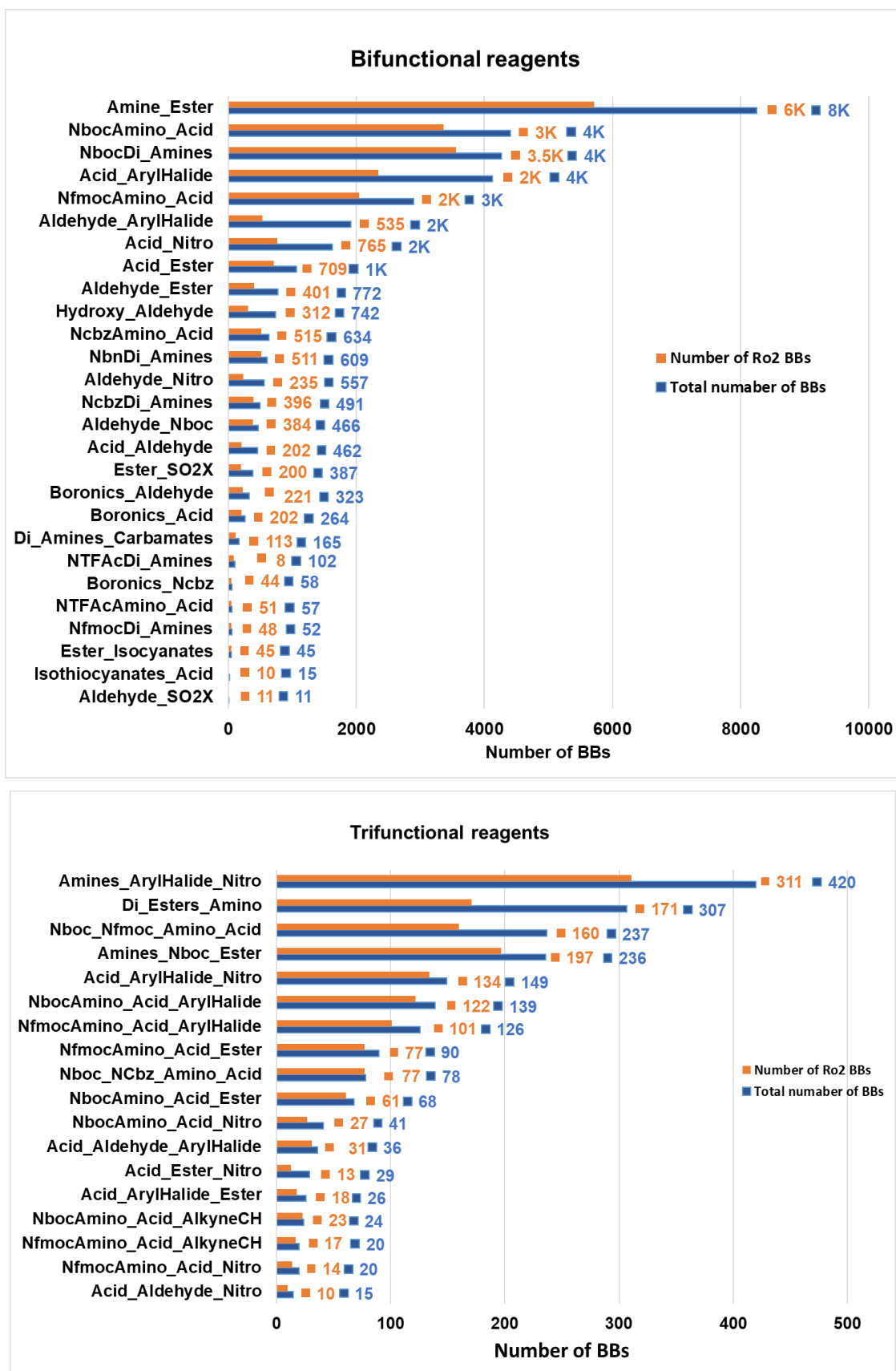
Figure 12. Polyfunctional commercially available reagents: total number and number of high-quality Ro2 compliant molecules.

.

The ability of the BB market to face current medicinal chemistry needs

For evaluation of the ability of PBBs to face medicinal chemistry needs, ChEMBL library, as a source of biologically relevant compounds, was fragmented using SynthI. In *Figure 13* one can see an example of such fragmentation. As a result, around 35% of ChEMBL molecules were fragmented into synthons that are all found in the eMolecules library. Around 5% of ChEMBL was not cut at all due to the small size of the molecules and lack of synthetically accessible acyclic bonds (heterocyclization was not taken into account). The remaining 60% of compounds have some but not all of synthons available – they include at least one synthon out of the scope of the eMolecules library.

For a more detailed analysis, electrophiles were further subdivided into acylating and sulfonylating agents, C-alkyl and C-aryl electrophiles. The nucleophiles were split into N-, O-, S-, C-alkyl and C-aryl nucleophiles. The populations of all synthon groups have been analyzed in *Figure 14.* In comparison with synthons generated from ChEMBL, the chemical market offers an abundance of reagents producing N-, O-nucleophiles, classical electrophiles, bivalent synthons and electrophilic radicals.

At the same time, there are several underrepresented synthons classes: all types of C-nucleophiles (Csp3-, Csp2- and C-boronics), S-nucleophiles, nucleophilic radical and N-electrophiles.

This goes in correspondence with conclusions derived in the previous chapter. However, synthons diversity for all the groups is higher for corresponding ChEMBL-specific synthons subsets (*Figure 15* and *Figure 16*), especially in the case of bivalent nucleophiles and electrophiles
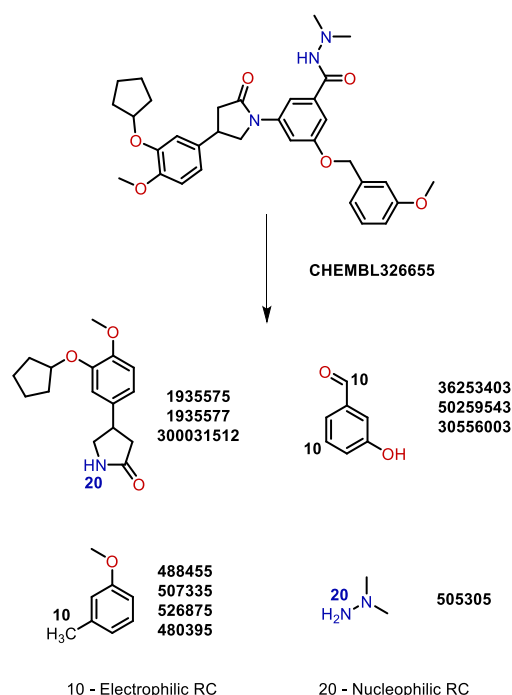
Figure 13. Example of ChEMBL molecule fragmentation towards commercially available synthons (eMolecules identifiers of corresponding BBs are provided).
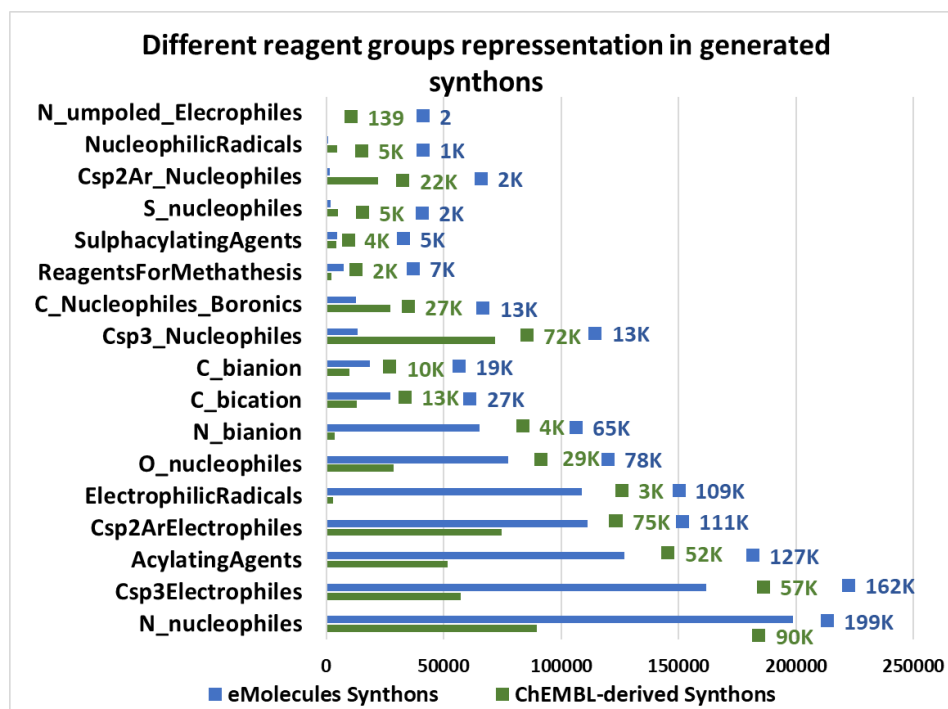
Figure 14. Comparison of the number of ChEMBL-specific and commercially available synthons
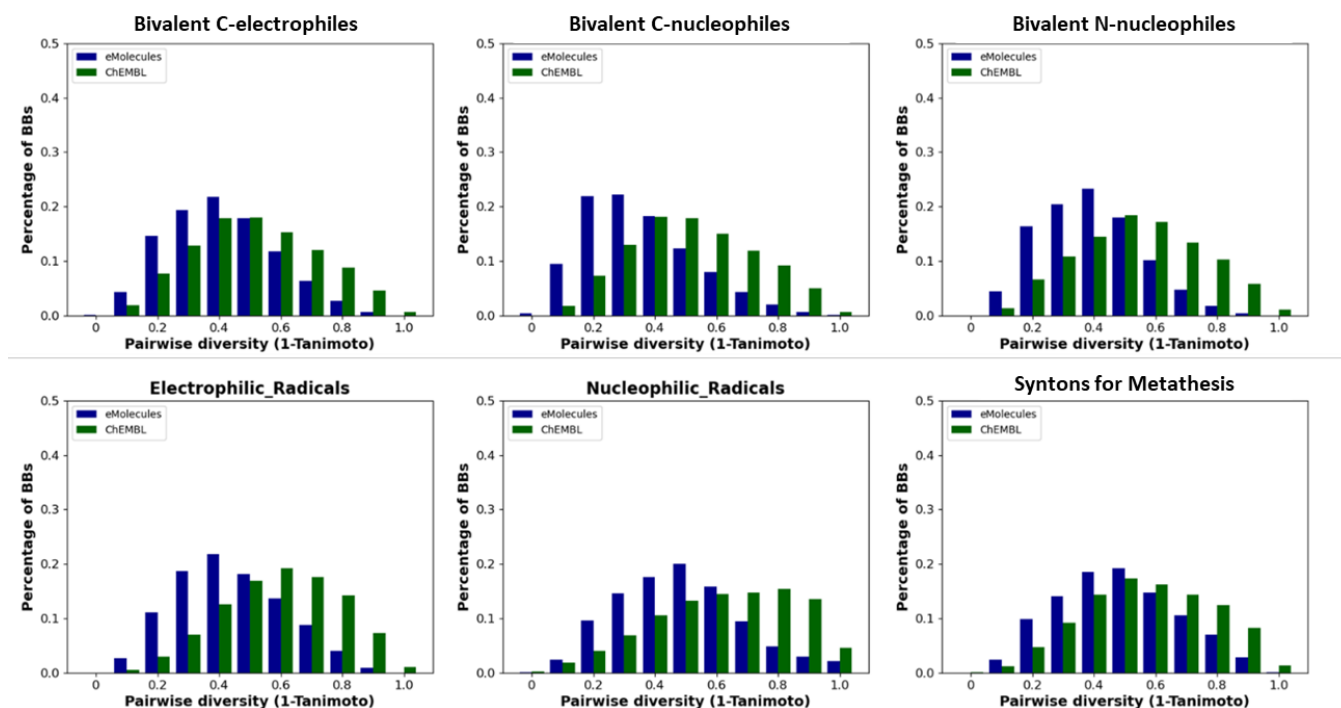
Figure 15. Relative diversity distribution for bivalent synthons classes.

GTM-based analysis of synthons

As a result of GA Pareto optimization, a synthons-uMap was selected out of thousands of evaluated options. This map is based on the atom-centered fragments of 1-2 atoms radius, that include both atoms and bond information. These descriptors are highly sensitive to the reactive center position, allowing to distinguish between synthons with different reactivity due to the inductive, mesomeric or steric effects. The manifold consists of a grid of 29*29 nodes coupled with 25*25 RBFs. This map provides synthon class separation with average balanced accuracy - BA of 0.9 (the lowest BA=0.79 for separation of C-nucleophiles from all other classes).

In **Figure 17(a)** one can see the density distribution for PBB-based synthons. Color code reflects the number of synthons in each point of the map – grey regions correspond to the minimally populated areas of the chemical space, while multicolored ones depict high-density picks. In agreement with previous synthon population analysis, the highest density is observed in the south-eastern part of the map. It corresponds to the primary N-monovalent and -bivalent nucleophiles produced by aliphatic amines and anilines (R8.1). Interestingly, primary hetero anilines form a separate cluster of slightly lower density further on the south (R8.2). At the same time, secondary N-nucleophiles are situated quite far from the primary ones in the central part of the map (R9). They are surrounded by acylation agents (R1) from one side and secondary aliphatic synthons with reactive center on the carbon atom from the other – mono- and bivalent C-electrophiles and bivalent C-nucleophilic synthons (R5). This is expected, as the ISIDA descriptors are sensitive to the position of the reactive center (marked atom) but not to the actual value of the atom label (encoding the type of intermediate). Therefore, C-electrophiles and C-nucleophiles (mono- or bivalent alike), can be found in the same region, but secondary (R5) and primary (R2) aliphatic synthons with reactive center on carbon atom are spatially separated. So are aliphatic (R2, R5) and aromatic (R6) synthons.

Similar to the primary N-nucleophiles in the regions R8.2 and R8.3, arylation agents (Csp2Ar-electrophiles, electrophilic radicals, Minisci CH-partners, aryl-boronics etc.) are split into two clusters with high density. The more crowded area is dominated by phenyl and α-pyridine synthons (R6.1). At the same time, the region with relatively moderate occupancy is populated by γ-heteroaryl synthons, usually with a higher number of heteroatoms (R6.2). The latter is neighboring the region R7, occupied by O-nucleophiles – aliphatic, benzylic alcohols, and phenols. Meanwhile, hetero-phenols and heteroaromatic thiols populate the area on the opposite part of the map (R3).
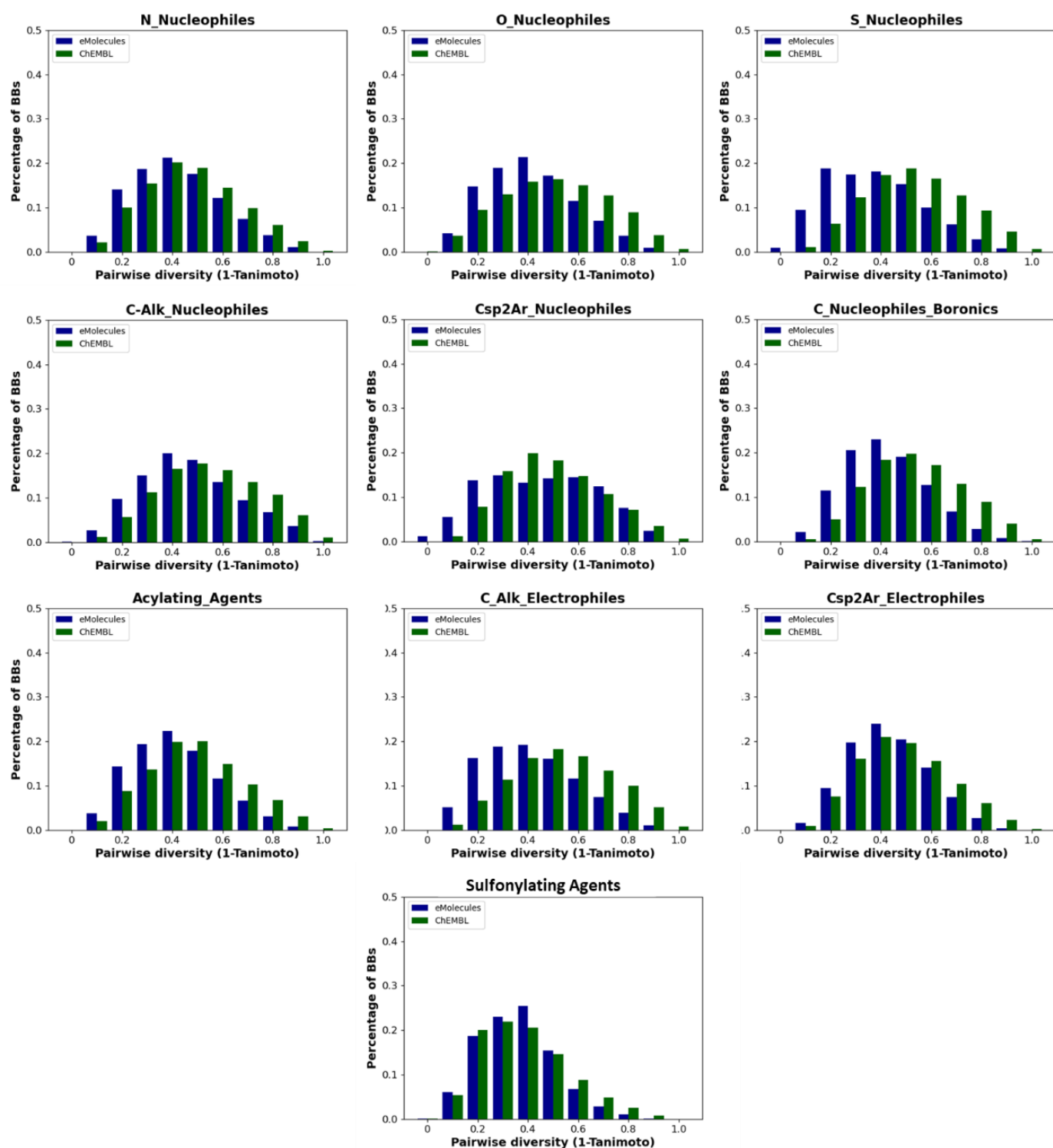
Figure 16. Relative diversity distribution for monovalent reagents' classes.
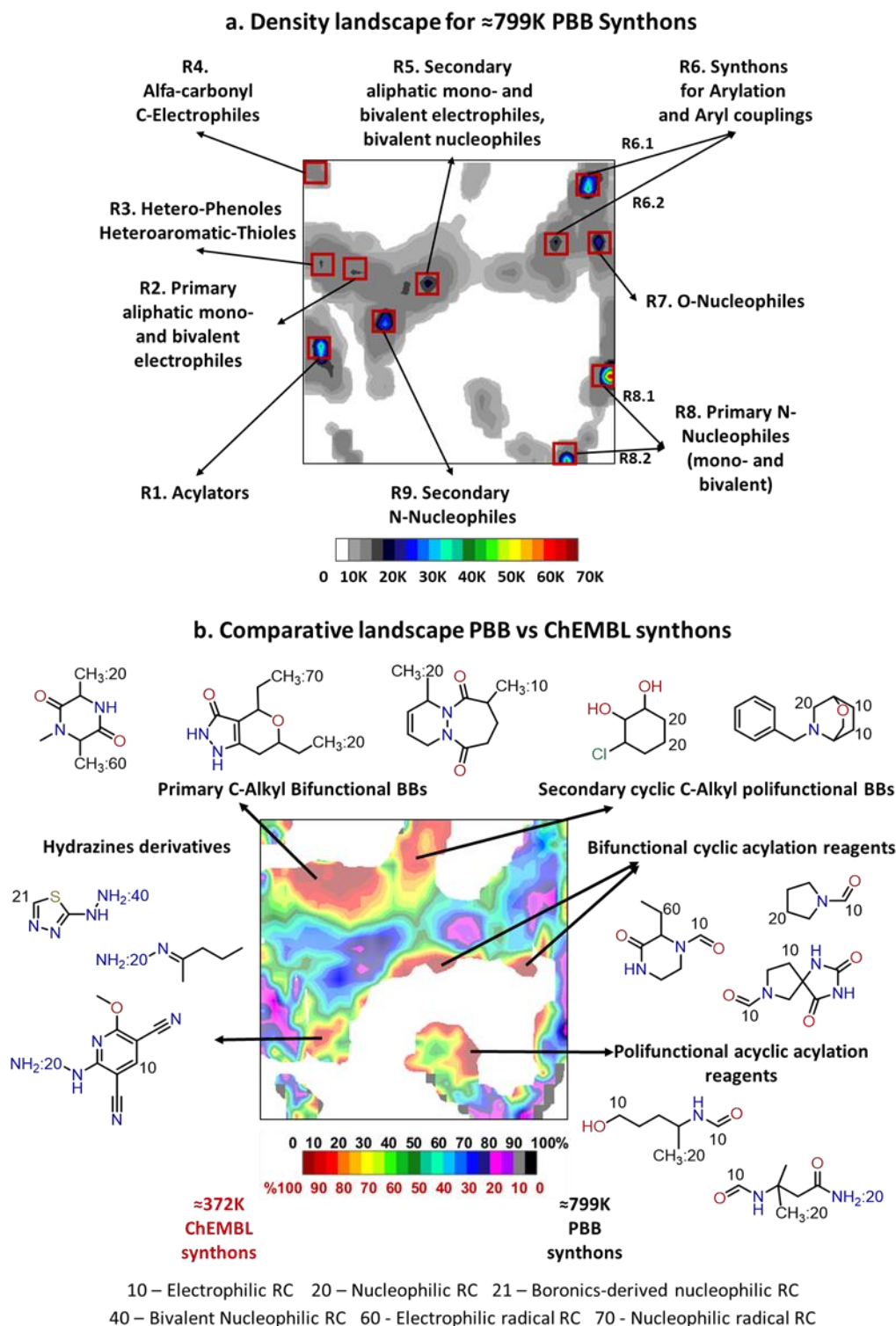
Figure 17. GTM analysis of synthons: a) density distribution of PBB synthons (color code reflects number of compounds in each point of the map); b) comparison of PBB synthons (black areas) and ChEMBL-derived synthons (red regions). ChEMBL-specific regions are profiled with examples of respective synthons.

a. Synthons classes with the highest libraries overlap

Synthons for Metathesis

ChEMBL – 2 288
PBB – 7 325

O-Nucleophiles

ChEMBL – 28 807
PBB – 77 539

Acylation Agents

ChEMBL – 25 976
PBB – 115 805

N-Nucleophiles

ChEMBL – 89 583
PBB – 198 664

b. Synthons classes with the highest excess of PBB but the lowest libraries overlap

Bivalent C-electrophiles

ChEMBL – 13 059
PBB – 27 321

Bivalent C-nucleophiles

ChEMBL – 10 042
PBB – 18 679

Csp2Ar-Electrophiles

ChEMBL – 98 540
PBB – 122 753

Bivalent N-nucleophiles

ChEMBL – 3 708
PBB – 65 445

C-Alk-Electrophiles

ChEMBL – 57 372
PBB – 161 856

Electrophilic radicals

ChEMBL – 2 901
PBB – 108 794

c. Synthons classes with the highest excess of ChEMBL BB

C-Alk-Nucleophiles

ChEMBL – 63 814
PBB – 13 251

Csp2Ar-Nucleophiles

ChEMBL – 31 181
PBB – 1 794

Nucleophilic radicals

ChEMBL – 4 715
PBB – 701

S-Nucleophiles

ChEMBL – 4 893
PBB – 1 890

C-Nucleophiles Boronics

ChEMBL – 27 271
PBB – 12 774

d. Synthons class with balanced libraries representation

Sulfonylation Agents

ChEMBL – 4 433
PBB – 4 588

ChEMBL synthons

0 10 20 30 40 50 60 70 80 90 100%
%100 90 80 70 60 50 40 30 20 10 0

PBB synthons

Figure 18. Synthons classes-based comparison of PBBs synthons (black areas) and ChEMBL-derived synthons (red regions).

The fact that in the same zone of the map one can find structurally similar or even identical synthons differing only in the nature of the reactive center is actually an advantage. The map can thus be used to search alternative synthesis ways, in situations where a same structural moiety can be provided by building blocks of radically different reactivity, applicable in distinct synthetic paths. For example, bivalent C-electrophiles and C-nucleophiles, intermediates in Knovenagel-, Wittig-, Julia-Kocienski- types of olefination, are occupying the same areas as reagents for metathesis – another reaction for double C=C bond formation.

Figure 17(b) shows the comparative landscape featuring PBB synthons (black color) versus synthons obtained via ChEMBL fragmentation (red color). All colors in between correspond to the mixed regions of different compositions (see the scale). It appears

that even though the number of PBB synthons is more than two times higher than the number of ChEMBL-derived synthons, there are still large ChEMBL-specific areas of the chemical space (red regions). These zones mostly correspond to the polyvalent synthons which, as it has been discussed earlier, are underrepresented on the market. In addition, the majority of synthons residing in ChEMBL-specific regions contain heterocycles, but heterocyclization processes were excluded from this analysis.

In order to obtain a better understanding of the chemical space of different synthons classes, 16 comparative ChEMBL vs PBB landscapes for each group, analyzed above, were constructed (***Figure 18***). Their comparison shows that despite lower diversity (***Figure 15*** and ***Figure 16***) of PBB synthons in all categories, there are still four classes that largely cover the chemical space of the respective ChEMBL synthons (***Figure 18 (a)***). Among them, there are synthons for metathesis, O- and N-nucleophiles and acylation agents. In all these classes there is a significant abundance of PBB synthons over ChEMBL-derived ones.

However, the high number of synthons does not always guarantee better coverage of biologically relevant synthons space. Indeed, bivalent electrophiles and nucleophiles, C-electrophiles and electrophilic radicals are also more numerous within the PBB synthons, but the overlap between commercially available and biologically relevant synthons is the smallest for these subsets (***Figure 18 (b)***). There are large areas exclusively occupied by representatives of only one library, which means that abundance of such synthons on the market still leaves room for improvement of the quality and structural diversity of corresponding BBs. Mostly it concerns areas that were associated with polyfunctional synthons containing more than one RC (***Figure 17(b)***).

The trends outlined in ***Figure 14*** are clearly seen in the comparative landscapes – there is a significantly higher portion of red areas for C- and S-nucleophilic synthons (***Figure 18 (c)***). Interestingly, even in the case of equivalently represented PBB and ChEMBL sulfonylation agents, there are still areas of biologically relevant synthons space not covered by PBBs.

CONCLUSIONS

In this work, commercially available BBs, provided by eMolecules, were analyzed in terms of purchasability, quality, diversity and ability to face current medicinal chemistry needs. The latter was achieved by fragmenting biologically relevant molecules from ChEMBL database with the help of SynthI – a knowledge-based reaction toolkit for library design and analysis. The resulting synthons were compared to those generated from PBB. This lead to a detailed comprehensive analysis of PBB in a medicinal chemistry context.

It was shown that the most represented classes of BBs – amines, acids, aryl halides and aliphatic alcohols – mirror the popularity of the respective reactions – amide formation, Pd-mediated couplings, Buchwald-Hartwig amination, alkylation etc. However, the existence of well-studied reactions is not the only factor defining reagent availability on the market. Indeed, sulfonate esters, secondary and (hetero)benzylic primary alkyl halides are far less present compared to other alkylation agents - alcohols, ketones and aldehydes respectively - due to their lower shelf-life time. The low number of S-nucleophiles can be explained by complicated storage conditions, while the lack of SuFEx reagents and polyfunctional BBs – by the relative youth of the efficient methodologies involving these reagents.

It was also noted that reported distribution of BB can limit the development of novel combinatorial techniques (nanomolar scale, robustness screen, photoredox catalysis, new generation of click chemistry, automated interactive cross-coupling, and late-stage functionalization). These are disfavored by the poor representation of necessary reagents, e.g. R–SO$_2$F, R-SO$_2$H salts, RCOONPhtal, R-BF$_3$K and R-BMIDA, SuFEx and polyfunctional BBs for DEL design.

Comparison of PBB- with ChEMBL-derived synthons reveals that the internal diversity among members of the same synthons class is significantly better for ChEMBL-derived synthons. It was shown that there is a lack of C- and S-nucleophiles and nucleophilic radicals, while O- and N-nucleophiles and electrophilic reagents are overrepresented on the market. GTM analysis allowed to identify that only in the case of four synthons classes PBB synthons cover largely ChEMBL-derived synthons chemical space: synthons for metathesis, acylation agents, O- and N-nucleophiles. For the other groups, even for those with high PBB synthons excess, there

are plenty of ChEMBL-specific areas of chemical space without any purchasable counterparts. Most of these areas correspond to the underrepresented on the market polyfunctional BBs.

All of these findings lead to the conclusion that there are plenty of possibilities for BBs libraries improvement – starting with enlargement of underrepresented BBs classes subsets and finishing with improving diversity and biological relevance of PBBs.

ACKNOWLEDGEMENTS

References

1. Grygorenko, O. O.; Volochnyuk, D. M.; Ryabukhin, S. V.; Judd, D. B. The Symbiotic Relationship Between Drug Discovery and Organic Chemistry. *Chemistry* **2020,** 26, 1196-1237.

2. Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of commercially available compounds for HTS. *Drug Discov. Today* **2019,** 24, 390-402.

3. Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A. Chemography: Searching for Hidden Treasures. *J Chem Inf Model* **2021,** 61, 179-188.

4. Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019,** 62, 1116-1124.

5. Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004,** 44, 643-51.

6. Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J Chem Inf Model* **2010,** 50, 470-9.

7. Lucas, X.; Gruning, B. A.; Bleher, S.; Gunther, S. The purchasable chemical space: a detailed picture. *J Chem Inf Model* **2015,** 55, 915-24.

8. Petrova, T.; Chuprina, A.; Parkesh, R.; Pushechnikov, A. Structural enrichment of HTS compounds from available commercial libraries. *MedChemComm* **2012,** 3, 571-579.

9. Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005,** 29, 55-67.

10. Verheij, H. J. Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Divers.* **2006,** 10, 377-88.

11. Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T. Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J Cheminform* **2017,** 9, 25.

12. Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015,** 20, 11-7.

13. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J Chem Inf Model* **2012,** 52, 1167-78.

14. GVK Biosciences Private Limited. In Plot No. 28 A, IDA Nacharam, Hyderabad 500076, India.

15. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006,** 34, D668-72.

16. Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug. Des.* **2009,** 74, 457-67.

17. Kalliokoski, T. Price-Focused Analysis of Commercially Available Building Blocks for Combinatorial Library Synthesis. *ACS Comb Sci* **2015,** 17, 600-7.

18. eMolecules, Inc. In https://www.emolecules.com/.

19. Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A. SynthI: a new open-source tool for synthon-based library design *ChemRxiv. Cambridge: Cambridge Open Engage;* **2021,** doi: 10.33774/chemrxiv-2021-v53hl-v2. This content is a preprint and has not been peer-reviewed.

20. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-

Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019,** 47, D930-D940.

21. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010,** 29, 855-68.

22. Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998,** 10, 215-234.

23. Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J. L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018,** 13, 540-554.

24. Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. Diversifying chemical libraries with generative topographic mapping. *J. Comput. Aided Mol. Des.* **2020,** 34, 805-815.

25. Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping in drug design. *Drug Discov Today Technol* **2019,** 32-33, 99-107.

26. Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: a New Look at the Natural Product Chemical Space. *Mol Inform, doi:10.1002/minf.202100068* **2021**.

27. Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J Chem Inf Model* **2015,** 55, 84-94.

28. Kireeva, N.; Baskin, II; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012,** 31, 301-12.

29. Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **2015,** 34, 348-56.

30. Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-task generative topographic mapping in virtual screening. *J. Comput. Aided Mol. Des.* **2019,** 33, 331-343.

31. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J Chem Inf Model* **2019,** 59, 564-572.

32. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided Mol. Des.* **2015,** 29, 1087-108.

33. Brown, D. G.; Bostrom, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016,** 59, 4443-58.

34. Cooper, T. W.; Campbell, I. B.; Macdonald, S. J. Factors determining the selection of organic reactions by medicinal chemists and the use of these reactions in arrays (small focused libraries). *Angew. Chem. Int. Ed. Engl.* **2010,** 49, 8082-91.

35. Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009,** 52, 6752-6.

36. Lovering, F. Escape from Flatland 2: complexity and promiscuity. *MedChemComm* **2013,** 4, 515-519.

37. Tomberg, A.; Bostrom, J. Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discov. Today* **2020,** 25, 2174-2181.

38. Ward, R. A.; Kettle, J. G. Systematic enumeration of heteroaromatic ring systems as reagents for use in medicinal chemistry. *J. Med. Chem.* **2011,** 54, 4670-7.

39. Buskes, M. J.; Blanco, M. J. Impact of Cross-Coupling Reactions in Drug Discovery and Development. *Molecules* **2020,** 25.

40. Nicolaou, C. A.; Watson, I. A.; LeMasters, M.; Masquelin, T.; Wang, J. Context Aware Data-Driven Retrosynthetic Analysis. *J Chem Inf Model* **2020,** 60, 2728-2738.

41. Vasudevan, A.; Bogdan, A. R.; Koolman, H. F.; Wang, Y.; Djuric, S. W. Chapter One - Enabling Chemistry Technologies and Parallel Synthesis—Accelerators of Drug Discovery Programmes. In *Progress in Medicinal Chemistry*, Witty, D. R.; Cox, B., Eds. Elsevier: 2017; Vol. 56, pp 1-35.

42. Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Synthesis of many different types of organic small molecules using one automated process. *Science* **2015,** 347, 1221-6.

43. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L. C.; Schneeweis, J.; Berritt, S.; Shi, Z. C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Organic chemistry. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015,** 347, 49-53.

44. Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **2018,** 359, 429-434.

45. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **2018,** 360, 186-190.

46. Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011,** 54, 3451-79.

47. Afanasyev, O. I.; Kuchuk, E.; Usanov, D. L.; Chusov, D. Reductive Amination in the Synthesis of Pharmaceuticals. *Chem. Rev.* **2019,** 119, 11857-11911.

48. Epifanov, M.; Mo, J. Y.; Dubois, R.; Yu, H.; Sammis, G. M. One-Pot Deoxygenation and Substitution of Alcohols Mediated by Sulfuryl Fluoride. *J. Org. Chem.* **2021,** 86, 3768-3777.

49. Figlus, M.; Wellaway, N.; Cooper, A. W.; Sollis, S. L.; Hartley, R. C. Synthesis of arrays using low molecular weight MPEG-assisted Mitsunobu reaction. *ACS Comb Sci* **2011,** 13, 280-5.

50. Huang, H.; Kang, J. Y. Mitsunobu Reaction Using Basic Amines as Pronucleophiles. *J. Org. Chem.* **2017,** 82, 6604-6614.

51. Hamid, M. H.; Allen, C. L.; Lamb, G. W.; Maxwell, A. C.; Maytum, H. C.; Watson, A. J.; Williams, J. M. Ruthenium-catalyzed N-alkylation of amines and sulfonamides using borrowing hydrogen methodology. *J. Am. Chem. Soc.* **2009,** 131, 1766-74.

52. Wright, S. W.; Hallstrom, K. N. A convenient preparation of heteroaryl sulfonamides and sulfonyl fluorides from heteroaryl thiols. *J. Org. Chem.* **2006,** 71, 1080-4.

53. Dong, J.; Krasnova, L.; Finn, M. G.; Sharpless, K. B. Sulfur(VI) fluoride exchange (SuFEx): another good reaction for click chemistry. *Angew. Chem. Int. Ed. Engl.* **2014,** 53, 9430-48.

54. Grygorenko, O. O.; Biitseva, A. V.; Zhersh, S. Amino sulfonic acids, peptidosulfonamides and other related compounds. *Tetrahedron* **2018,** 74, 1355-1421.

55. Mykhalchuk, V. L.; Yarmolchuk, V. S.; Doroschuk, R. O.; Tolmachev, A. A.; Grygorenko, O. O. [3+2] Cycloaddition of an Azomethyne Ylide and Vinyl Sulfonyl Fluorides — an Approach to Pyrrolidine-3-sulfonyl Fluorides. *Eur. J. Org. Chem.* **2018,** 2018, 2870-2876.

56. Sokolov, A.; Golovach, S.; Kozlinsky, I.; Dolia, K.; Tolmachev, A. A.; Kuchkovska, Y.; Grygorenko, O. O. Diastereoselective Synthesis of Cyclic sp3-Enriched cis-β-Alkoxysulfonyl Chlorides. *Synthesis* **2019,** 51, 848-858.

57. Tolmachova, K. A.; Moroz, Y. S.; Konovets, A.; Platonov, M. O.; Vasylchenko, O. V.; Borysko, P.; Zozulya, S.; Gryniukova, A.; Bogolubsky, A. V.; Pipko, S.; Mykhailiuk, P. K.; Brovarets, V. S.; Grygorenko, O. O. (Chlorosulfonyl)benzenesulfonyl Fluorides-Versatile Building Blocks for Combinatorial Chemistry: Design, Synthesis and Evaluation of a Covalent Inhibitor Library. *ACS Comb Sci* **2018,** 20, 672-680.

58. Kokhan, S. O.; Valter, Y. B.; Tymtsunik, A. V.; Komarov, I. V.; Grygorenko, O. O. 3-Carboxy-/3-Aminobicyclo[1.1.1]pentane-Derived Sulfonamides and Sulfonyl Fluorides - Advanced Bifunctional Reagents for Organic Synthesis and Drug Discovery. *Eur. J. Org. Chem.* **2020,** 2020, 2210-2216.

59. Helal, C. J.; Bundesmann, M.; Hammond, S.; Holmstrom, M.; Klug-McLeod, J.; Lefker, B. A.; McLeod, D.; Subramanyam, C.; Zakaryants, O.; Sakata, S. Quick Building Blocks (QBB): An Innovative and Efficient Business Model To Speed Medicinal Chemistry Analog Synthesis. *ACS Med Chem Lett* **2019,** 10, 1104-1109.

60. Collins, K. D.; Glorius, F. A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* **2013,** 5, 597-601.

61. Fujiwara, Y.; Dixon, J. A.; O'Hara, F.; Funder, E. D.; Dixon, D. D.; Rodriguez, R. A.; Baxter, R. D.; Herle, B.; Sach, N.; Collins, M. R.; Ishihara, Y.; Baran, P. S. Practical and innate carbon-hydrogen functionalization of heterocycles. *Nature* **2012,** 492, 95-9.

62. Crisenza, G. E. M.; Melchiorre, P. Chemistry glows green with photoredox catalysis. *Nat. Commun.* **2020,** 11, 803.

63. Meng, G.; Guo, T.; Ma, T.; Zhang, J.; Shen, Y.; Sharpless, K. B.; Dong, J. Modular click chemistry libraries for functional screens using a diazotizing reagent. *Nature* **2019,** 574, 86-89.

64. Barrow, A. S.; Smedley, C. J.; Zheng, Q.; Li, S.; Dong, J.; Moses, J. E. The growing applications of SuFEx click chemistry. *Chem. Soc. Rev.* **2019,** 48, 4731-4758.

65. Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **2016,** 45, 546-76.

66. Proctor, R. S. J.; Phipps, R. J. Recent Advances in Minisci-Type Reactions. *Angew. Chem. Int. Ed. Engl.* **2019,** 58, 13666-13699.

67. Zhang, G.; Howe, M.; Aldrich, C. C. Spirocyclic and Bicyclic 8-Nitrobenzothiazinones for Tuberculosis with Improved Physicochemical and

Pharmacokinetic Properties. *ACS Med Chem Lett* **2019,** 10, 348-351.

68. Manka, J. T.; Rodriguez, A. L.; Morrison, R. D.; Venable, D. F.; Cho, H. P.; Blobaum, A. L.; Daniels, J. S.; Niswender, C. M.; Conn, P. J.; Lindsley, C. W.; Emmitte, K. A. Octahydropyrrolo[3,4-c]pyrrole negative allosteric modulators of mGlu1. *Bioorg Med Chem Lett* **2013,** 23, 5091-6.

69. Grygorenko, O. O.; Radchenko, D. S.; Volochnyuk, D. M.; Tolmachev, A. A.; Komarov, I. V. Bicyclic conformationally restricted diamines. *Chem. Rev.* **2011,** 111, 5506-68.

70. Cahard, D.; Ma, J. A. *Emerging Fluorinated Motifs: Synthesis, Properties, and Applications*. 2020; p i-xiv.

71. Meanwell, N. A. Fluorine and Fluorinated Motifs in the Design and Application of Bioisosteres for Drug Design. *J. Med. Chem.* **2018,** 61, 5822-5880.

72. Mader, P.; Kattner, L. Sulfoximines as Rising Stars in Modern Drug Discovery? Current Status and Perspective on an Emerging Functional Group in Medicinal Chemistry. *J. Med. Chem.* **2020,** 63, 14243-14275.

73. Finkbeiner, P.; Hehn, J. P.; Gnamm, C. Phosphine Oxides from a Medicinal Chemist's Perspective: Physicochemical and in Vitro Parameters Relevant for Drug Discovery. *J. Med. Chem.* **2020,** 63, 7081-7107.

74. Ramesh, R.; Reddy, D. S. Quest for Novel Chemical Entities through Incorporation of Silicon in Drug Scaffolds. *J. Med. Chem.* **2018,** 61, 3779-3798.

75. Mykhailiuk, P. K. Saturated bioisosteres of benzene: where to go next? *Org Biomol Chem* **2019,** 17, 2839-2849.

————————

Summary

The chemical space of the commercially available BBs was studied here with the help of SynthI and GTM. The Purchasability, quality, diversity, and ability of PBB to face current medicinal chemistry needs have been analyzed. The latter was achieved by fragmenting biologically relevant molecules from the reference ChEMBL database. Comparing the resulted fragments with PBB-derived synthons allowed us to perform the first comprehensive analysis of PBBs in a medicinal chemistry context.

The representation of different classes of BBs in the chemical space of PBB was discussed as a function of:

i)   their usage in popular in medicinal chemistry reactions;

ii)   their stability and storage conditions;

iii)   'maturity' of the efficient methodologies involving these reagents.

Comparison of PBBs with ChEMBL-derived synthons reveals that only one-third of ChEMBL can be fully synthesized using commercially available BBs. Synthons chemical space was represented with the help of ISIDA descriptors. Their main advantage is that they are highly sensitive to the reactive center position, allowing to distinguish between synthons with different reactivity due to the inductive, mesomeric, or steric effects. The corresponding synthons will have distinct ISIDA descriptors despite being based on the same molecular graph due to the labels introduced into the synthons structure. Two synthons contributing the same fragment but having a different reactive center at the same position (envisaging different reaction mechanisms) have, however, identical ISIDA descriptors (they capture the label position, not its actual value). Such synthons are distinct options that provide the same contribution to the final compounds – their existence is practically important because they allow search for alternative synthetic pathways, but they are indeed redundant from a structural point of view. With the help of such representation of the chemical space of BB, the internal diversity among members of the same reagent classes was analyzed. It appears that it is significantly higher for ChEMBL-derived synthons. It was shown that there is a lack of C- and S-nucleophiles and nucleophilic radicals, while O- and N-nucleophiles and electrophilic reagents are overrepresented.

New synthons-uGTM were optimized herewith in a way to simultaneously host and efficiently separate different types of synthons (electrophiles, nucleophiles, radicals, etc.). It was constructed using both experimental (PBB-derived) and theoretical (ChEMBL-derived)

synthons, which should extend the scope of its applicability beyond the currently available reagents. This map enabled a detailed comparison of the chemical space of different synthons classes providing a better understanding of BB that medicinal chemists have at their disposal. It was shown that only in the case of four reagent classes – reagents for metathesis, acylation agents, O- and N-nucleophiles – PBBs cover largely ChEMBL-derived synthons chemical space. For other groups of BBs, even for those with high PBBs excess, there are plenty of ChEMBL-specific areas of chemical space without any PBBs counterparts. Most of these areas correspond to the underrepresented on the market polyfunctional BBs.

All of these findings lead to the conclusion that there are plenty of possibilities for BBs libraries improvement – starting with enlargement of underrepresented BBs classes subsets and finishing with improving diversity and biological relevance of PBBs.

4.5 Natural products

Introduction

Even though the drug discovery domain relies largely on organic chemistry to provide the pool of highly probable hits, natural products still remain the main source of inspiration for medicinal chemists. Numerous studies showed that natural products occupy parts of the chemical space, not explored by available screening collections, which makes them valuable components of screening libraries used in drug discovery[128]. Therefore the chemical space of NPs and NP-like[129, 130] compounds deserve a separate discussion.

In this Chapter, we report analysis of NPs from the COCONUT library. It included a new NP-uMap optimization, hierarchical zooming application, and comparison of genuine NPs to commercially available (ZINC) and biologically tested (ChEMBL) NP-like compounds. Moreover, NPs active against popular target families (kinases, proteases, other enzymes, ion channels, nuclear receptors, GPCRs, epigenetic targets, transporters), have been analyzed to find characteristic structural features unique for each of the ligand series.

Main terminology

NP-likeness – similarity of the given molecule to the structure space covered by natural products (NPs).

NP-liekeness score - a Bayesian measure which allows to determine how molecules are similar to the structural space covered by natural products as opposed to the structure space covered by synthetic molecules.

QED score - Quantitative Estimate of Druglikeness - a quantitative metric for assessing druglikeness with respect to the Ro5 complience. QED score values can range between zero (all properties unfavourable) and one (all properties favourable).

NP Navigator: a New Look at the Natural Product Chemical Space

Yuliana Zabolotna,[a] Peter Ertl,[b] Dragos Horvath,[a] Fanny Bonachera,[a] Gilles Marcou,[a] and Alexandre Varnek*[a, c]

Abstract: Natural products (NPs), being evolutionary selected over millions of years to bind to biological macromolecules, remained an important source of inspiration for medicinal chemists even after the advent of efficient drug discovery technologies such as combinatorial chemistry and high-throughput screening. Thus, there is a strong demand for efficient and user-friendly computational tools that allow to analyze large libraries of NPs. In this context, we introduce NP Navigator – a freely available intuitive online tool for visualization and navigation through the chemical space of NPs and NP-like molecules. It is based on the hierarchical ensemble of generative topographic maps, featuring NPs from the COlleCtion of Open NatUral producTs (COCONUT), bioactive compounds from ChEMBL and commercially available molecules from ZINC. NP Navigator allows to efficiently analyze different aspects of NPs - chemotype distribution, physicochemical properties, biological activity and commercial availability of NPs. The latter concerns not only purchasable NPs but also their close analogs that can be considered as synthetic mimetics of NPs or pseudo-NPs.

Keywords: chemoinformatics · natural products · chemical space · visualization · pseudo-NPs

1 Introduction

For centuries, natural products (NPs) were the only source of traditional medicines all over the world. Being evolutionary selected over millions of years to bind to biological macromolecules, they are able to selectively interact with many specific targets within the cell.[1] Therefore, NPs and their molecular frameworks remained an important source of inspiration for medicinal chemists even after the advent of efficient drug discovery technologies such as combinatorial chemistry[2] and high-throughput screening.[3] According to a comprehensive analysis, 6% of all small-molecule drugs approved between 1981 and 2014 are unaltered NPs, 26% are NP derivatives, and 32% are NP mimetics and/or contain an NP pharmacophore.[4]

Over the past 20 years, quite a large number of scientific reports exhaustively analyzed the chemical space of NPs in the medicinal chemistry context. Several studies were dedicated to the analysis of structural and physicochemical features of different libraries of NPs[5] as well as their comparison to drugs and synthetic combinatorial libraries.[6] In addition, several models were proposed for distinguishing between natural products and synthetic molecules.[7] All of these reports contributed to a better understanding of NP-distinctive features, like heteroatom composition, number of rings, degree of saturation etc. In numerous publications, it was shown that NPs occupy parts of the chemical space not explored by available screening collections, which makes them valuable components of screening libraries used in drug discovery and increases the impor-

tance of computational tools for navigation of NP chemical space.[8]

Different methods are suitable for this task and a lot of them have been already used to analyze libraries of compounds of natural origin.[9] Principal component analysis (PCA)[10] and scaffold trees[11] were most often used, but self-organizing maps,[12] generative topographic mapping (GTM)[13] and a new visualization method – tree maps (TMAP)[14] were also applied.

Most of the numerous articles in this field simply report static results of particular compound library analysis, not allowing readers to explore the chemical space of NPs by themselves. To our best knowledge, there are only three web-based open platforms providing users with a certain level of interactivity and exploration freedom. The first one is an interactive web portal associated to The Natural

[a] *Y. Zabolotna, D. Horvath, F. Bonachera, G. Marcou, A. Varnek*
University of Strasbourg,
Laboratory of Chemoinformatics,
4, rue B. Pascal, 67081 Strasbourg (France)
Phone: +33-368851560
E-mail: varnek@unistra.fr
[b] *P. Ertl*
Novartis Institutes for BioMedical Research
Novartis Campus, CH-4056, Basel, Switzerland
[c] *A. Varnek*
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University
Kita 21 Nishi 10,Sapporo, Kita-ku, 001-0021 Sapporo, Japan

Products Atlas – a database of microbial natural products that includes 24,594 compounds and associated data.[15] A similarity-based network is used to cluster and visualize these compounds providing the ability to browse and search through them. The second platform, provides TMAP visualization of the same database.[14b] The third one is called D-Peptide Builder. It is a peptide generator, that also allows to visualize chemical space of peptides from different libraries using PCA and t-SNE plots.[16] However, all of them are limited to just a few distinct compound classes, visualizing only particular segments of the chemical space of NPs (only up to ≈ 25 K NPs). Moreover, The Natural Product Atlas and D-Peptide Builder can be considered as simple database interfaces, that were not specifically designed for in-depth exploration, but rather for demonstrative purposes. For example, it is impossible to change "visualization perspective", i.e. display distribution of different properties that users may be interested in. D-Peptide Builder does not even allow to display chemical structures – only compound names appear on the plot. Last but not least, none of these three platforms allow to project user-defined molecules for comparison with the database content.

In this context, we present NP Navigator – a free, intuitive on-line tool for visualization and navigation through the chemical space of NPs and NP-like molecules. It is based on the hierarchical ensemble of generative topographic maps, featuring NPs from the COlleCtion of Open NatUral producTs (COCONUT),[5b,17] bioactive compounds from ChEMBL and commercially available molecules from ZINC.[18] Being a nonlinear probabilistic dimensionality reduction method,[19] GTM is well suited to power NP Navigator. It has already proven to be a successful approach for visualization and versatile analysis of large chemical libraries.[20] Hierarchical extension of GTM, combined with Maximum Common Substructure (MCS) detection[20b] allows to establish the link between the generalized visualization of the known chemical space of NPs/NP-like molecules and structural features of each separate compound.

As a result, NP Navigator allows to efficiently analyze different aspects of NPs - chemotype distribution, physico-chemical properties, (reported and/or predicted) biological activity and commercial availability of NPs. The latter concerns not only purchasable NPs but also their close analogs that can be considered as pseudo-NPs.[21] Users are welcome not only to browse through hundreds of thousands of compounds from ZINC, ChEMBL and COCONUT but also project a small dataset of external molecules that play the role of "chemical trackers" allowing to trace particular chemotypes in the NP chemical space and detect analogs of the compound of interest.

Web-based implementation of NP Navigator is freely accessible at the link - https://infochm.chimie.unistra.fr/npnav/chematlas_userspace.

2 Materials and Methods

2.1 Data Preparation

2.1.1 Natural Products

The COCONUT database v. 2020.4 is a free and open collection of more than 426,000 structures that were obtained by retrieving data from 53 sources and collecting additional data from the literature. However, molecules:

- with NP-likeness score < -0.5
- containing typical chemotypes privileged in synthetic compounds (polyhalogenated hydrocarbons, sulfonamides, thioureas etc.)

are not genuine NPs in our opinion, and were not considered in the present work.

The NP-likeness score threshold was selected based on the previous experience, in a way to remove some simple organic compounds, that usually would be considered as synthetic. Even though they still may naturally occur, they do not have the degree of complexity typically associated with the "NP" label. They happen to contain more structural motifs that are frequently found in synthetic molecules, rather than moieties common for NPs. For consistency reasons all datasets used in this work have been filtered according to the same threshold. NP-likeness score was calculated using RDKit-based implementation of the method described in the original article,[7b] which can be found in the GitHub repository https://github.com/rdkit/rdkit/tree/master/Contrib/NP_Score.

The remaining 254,024 compounds have been standardized according to the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics at the University of Strasbourg (infochimie.u-strasbg.fr/webserv/VSEngine.html) using the ChemAxon Standardizer.[22] That included:

- dearomatization and final aromatization (heterocycles like pyridone were not aromatized);
- conversion to canonical SMILES;
- salts and mixture removal; neutralization of all species, except nitrogen (IV);
- the major tautomer generation
- stereochemical information removal.

Stereochemical information has been ignored due to the fact that ISIDA descriptors,[23] used in this work, would not capture it, anyway. As a result, 253,893 unique "stereochemistry-agnostic" molecular graphs remained. Each unique entry was linked to all the molecular IDs of the one or more stereoisomeric forms under which it actually appears in COCONUT.

Some NPs are often glycosylated in nature, and it is debatable whether they should be best represented under their non-glycosylated form for analysis.[24] In this work, compounds were taken as in COCONUT.

2.1.2 In-Stock Commercially Available Compounds

9,218,095 In-Stock compounds of "standard" reactivity have been downloaded from the ZINC20 website in October 2020. After standardization and duplicate deletion 6,460,596 compounds remained. Only 586,235 of them have NP-likeness scores higher than −0.5. These compounds (further – NP-like ZINC dataset) were used to define NP-Like commercially available chemical space. Among them, 11 K compounds were found in COCONUT library and thus represent commercially available NPs.

2.1.3 Tangible Commercially Available Compounds

1.36 billion tangible compounds (not available for immediate purchase but might be synthesized upon request) were collected from the ZINC15 website in January 2019. After standardization, around 800 million stereochemistry-depleted tangible ZINC compounds remained, out of which 84,531,030 tangible NP-like compounds passed the NP-likeness >-0.5 filter.

2.1.4 Biologically Tested Compounds

ChEMBL (version 26)[25] served as a reference dataset for biologically tested molecules. 1,950,765 compounds have been collected in May 2020. After standardization, 1,721,155 unique compounds with known biological activities were filtered according to NP-likeness score resulting in 474,335 NP-like ChEMBL compounds.

The intersection of standardized ChEMBL and COCONUT returned 44,947 biologically tested NPs. Only 6,881 of them demonstrated dose-response activity on some target, with an activity value less than 10 μm – active NPs. They were further classified with respect to their target family as provided in ChEMBL:

- kinases;
- proteases;
- other enzymes;
- ion channels;
- nuclear receptors;
- GPCRs;
- epigenetic targets;
- transporters;
- others.

The full list of targets for each of the targets may be downloaded from the ChEMBL website where also an interactive browser is available (https://www.ebi.ac.uk/chembl/g/#browse/targets) allowing to see the whole target hierarchy.

2.2 ISIDA Descriptors

ISIDA property-labeled fragment descriptors encode molecular structures as counts of specific subgraphs. Nodes of these subgraphs, representing atoms, can be either labeled by element type (default) or by some local property/feature: pH-dependent pharmacophore type, electrostatic potential, force field type etc.[23] There are several fragmentation schemes - from classical atom pair and sequence counts to branched fragments or multiplets. Also, bond information may be represented or ignored, thus leaving a vast choice in terms of the level of resolution at which chemical information should be extracted into the descriptors.

In this work, we have generated more than 100 types of ISIDA descriptors, which were selected for the relatively low number of fragments they generate and previous success in chemical space analysis and activity modeling. The most suitable for NP chemical space exploration descriptor type was selected via evolutionary optimization described in the next chapters.

2.3 Generative Topographic Mapping

Generative topographic mapping (GTM) is a dimensionality reduction method originally described by Bishop.[19] The algorithm performs a non-linear projection from the initial N-dimensional space into a 2D latent space. In chemoinformatics the former is defined by the N-dimensional descriptor vectors assigned to each molecule of the dataset. The latent space resumes to a manifold, which is defined by a set of radial basis functions (RBF). The manifold is evaluated on sample points termed «nodes». At the training stage, the shape of the manifold is fitted to pass through the densest regions of the "frame set" (the pool of molecules used to probe the chemical space of interest). Then the nodes are folded back in 2D plane, as a squared grid.

By contrast to Self-Organizing Maps,[26] GTM assigns each molecule not to only one "winning" node but fuzzily distributes it over all nodes, with larger probabilities ("responsibilities") for near nodes. For each compound, responsibilities sum to one. Such a smooth projection supports the creation of GTM landscapes – 2D plots of cumulated compound responsibilities, colored by average values of different properties, e.g. density, biological activity, assigned class, etc. GTM landscapes can be used for chemical space analysis, library comparison or as a basis for building QSAR models.[27]

2.4 Universal NP Map: Concept and Construction

Universal GTMs have been introduced by Sidorov et al[28] and further developed by Casciuc et al.[29] They were defined as the "best compromise" maps, providing satisfac-

tory predictive performance with respect to very diverse biological properties. Seven universal maps of the ChEMBL chemical space, defined by ISIDA fragment descriptors, have been "evolved" by a genetic algorithm (GA)[30] in the map parameter space (including descriptor choice, grid size, manifold flexibility controls, etc, as key degrees of freedom). An average predictive performance over 236 biological activities was used as an objective function in a search for the best GTM parameters. These GTMs were proven to successfully serve as hosts for 618 (later extended to 749) activity landscapes associated with the respective target-specific structure-activity ChEMBL compound series. Later they were combined in a consensus model implemented as an on-line GTM-based Profiler (http://infochim.u-strasbg.fr/webserv/VSEngine.html).

Unfortunately, due to the limited number of NPs in ChEMBL, their applicability to NP chemical space analysis is not appropriate. A dedicated NP map was evolved as part of this work, albeit with a different, Pareto-front driven multiobjective strategy. A fixed frame set 16,025 randomly selected NPs was used. The maps were challenged to maximize:

(a) the pairwise separation of NPs assigned to different activity classes (vide supra): for each of the $9*(9-1)/2 = 36$ pairs (C_i,C_j) of distinct activity classes. The mutual separation of respective class members on the landscape is reported as a cross-validated balanced accuracy (BA) score and used as an objective function for best GTM parameters selection. Maps in which the compound sets significantly overlap will witness members of class C_i projecting amid a cluster of representatives of C_j during cross-validation, resulting in lower BA. By contrast, parameter choices defining maps in which members of C_i and C_j are projected on distinct areas of the manifold would not lead to such mispredictions and thus higher BA values will be obtained.

(b) the Shannon entropy of a large (24 K) random subset of NPs, normalized with respect to the maximal entropy achievable on a map of N nodes. Recall that the Shannon entropy of a mapped compound library is $S = -\sum_{i=1}^{N} f_i \ln f_i$, where f_i is the fraction of "compounds residing in node i" in terms of cumulated responsibilities (cumulated responsibility of node i by compound library size L). The "ideal" maximal entropy map providing the most homogeneous possible mapping would equally split the library over all its nodes, thus $f_i = 1/N$ and $S_{max} = -\sum_{i=1}^{N} \frac{1}{N} \ln \frac{1}{N} = \ln N$.

The entropy objective, equaling S/ln(N) becomes independent of map size and characterizes the homogeneity of the NP distribution over the landscape.

Unlike in the previous universal map strategy – where the initial 236 balanced accuracy objectives were "collapsed" into a single fitness score (their plain arithmetic average minus standard deviation) the present approach considered the above 36 (BAs) + 1 (S/lnN) as independent objectives, and the Pareto front of non-dominated maps

was considered as the current "breeding" population. A new "individual" obtained by standard genetic operators is evaluated by generating the map according to the parameter values encoded in its chromosome, required compounds are projected on it and the 37 objective scores are estimated. If another, previously discovered parameter configuration is known to have produced a map which is better than the "new born" one with respect to each of the 37 objectives, the newborn configuration is "dominated" and will be discarded. Otherwise, the configuration is better than the so-far found with respect to at least some of these objectives and is allowed to enter the current population.

2.6 Hierarchical GTM (HGTM)

While analyzing hundreds of thousands of compounds, map resolution may be insufficient for meaningful chemotype clustering. In such a case, a hierarchical zooming approach is required to improve class separation on the finer scale of zoomed maps. Hierarchical GTM (HGTM), a.k.a "Zooming"[31] is a technique that trains a new map on a set of compounds extracted from a given zone on the parent map, in order to further resolve compound clusters with degenerated responsibility patterns. This approach, combined with a maximum common substructure (MCS) detecting algorithm was previously implemented in AutoZoom[20b] – an in-house tool that has been developed for the chemotypes identification in the heavily populated zones of the map. First, it separates the map into small zones (3×3 nodes) and detecting "overcrowded" zones (of more than 1000 compounds). In this work, zone "residents" were counted as compounds for which the sum of responsibilities over the nodes in the particular zone is higher than 0.85. A pool of 10% of residents (but not less than 1000) was selected using the dissimilarity principle and used as a frame set for the new GTM manifold construction (with map parameters "borrowed" from the parent map). Successive zooming of all overcrowded zones was hierarchically performed until all are eventually broken up into clusters of less than 1000 compounds and then submitted to the MCS extraction, realized using ChemAxon's JChem engine.[22] Only MCS covering at least 30% of each of the molecules were reported. After the primary identification of the specific MCS, they were submitted as substructure search queries in order to verify whether they are genuinely absent from the entire subspace (and not only from the zones targeted by successive zooming)

3 Results and Discussion

3.1 Optimal NP-Umap

Figure S1 comparatively displays the residence areas of ChEMBL compounds versus COCONUT NPs on the seven
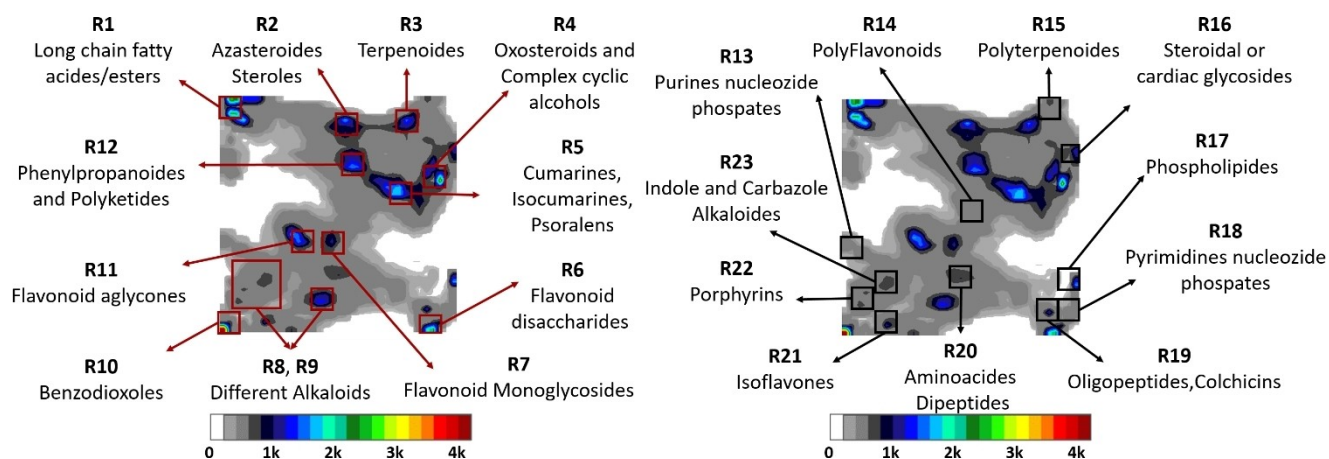
Figure 1. Density landscape of NPs from COCONUT. On the left – chemotypes for the highly populated regions, on the right – for the low populated ones. Multicolored areas correspond to the highly populated regions, while gray color defines moderately occupied areas. White zones are empty.

previously constructed universal maps.[29] As one can see, NPs (blue regions) agglutinate in specific zones. This forces a lot of different NP chemotypes to "collide" in the same nodes, preventing their meaningful separation and clustering. Therefore, fitting of a NP-dedicated Universal map (NP-Umap) is proven mandatory.

By definition, a Pareto-front driven optimization does not produce a single best solution unless all objectives are correlated and a configuration simultaneously maximizing all of them exists. This is not expected to be the case here. Thousands of map configurations were retrieved, each having locally some competitive edge over others, in terms of specific objectives. Note that perfect separation of the members of considered classes is neither necessary nor expected (actually, some compounds are "promiscuous" and included in several classes – ion channels and GPCRs, for example, are notoriously sharing many actives). In these cases, the same molecule is present twice in the cross-validation set – labeled both as "C_i" and "C_j", making overlap unavoidable. The goal is to maximize separation as far as this is possible, not to aim for perfect separation.

Eventually, one map was hand-picked, amongst those with worst balanced accuracy exceeding some minimal threshold (here, 0.59), all while being based on the technically most convenient descriptors amongst the ones allowing such level of performance. The selected "best" map consists of 1,225 nodes (35×35) coupled with 324 RBFs (18×18). The descriptors used to define NPs chemical space are ISIDA symmetrical atom-centered fragments with topological distance from 1 to 2 including both atoms and bonds information. These are easier to calculate than the topological pharmacophore fragments very often encountered in good maps (the latter require an additional pharmacophore typing step, which may be expensive as it involves an explicit protonation state prediction). The average BA in class separation is 0.67 (Table S1). This map is

"NP-universal" in the sense that chosen set of parameters, including descriptor type, embodies a simultaneous capacity to satisfactorily separate NPs, active associated to various (here, nine) biologically unrelated target classes. This broadens NP-Umap application for chemical space analysis in a medicinal chemistry context.

3.2 Chemical Space of Natural Products – Chemotype Distribution

The entire NP dataset has been projected onto the newly constructed NP-Umap. Figure 1 shows the obtained density landscape, colored according to the cumulative sum of responsibilities of compounds residing in each node. According to the color scale, colored areas correspond to the highly populated regions, while moderately occupied areas are gray. White zones are empty. As shown in Figure 1, the densest areas are unsurprisingly populated by the most common NP families e.g. lipids, alkaloids, sugars, flavonoids etc.

In general, the northern part of the map corresponds to the NPs with a high proportion of carbon atoms – long-chain fatty acids and corresponding lipids (R1), steroid-like compounds (R2), terpenoids (R3) etc. While heading southeast, the number of oxygen atoms increases resulting in dense regions of polyketides (R12), oxosteroids (R4) coumarins and psoralenes (R5). Close to the oxosteroids, a small island of steroidal or cardiac glycosides (R16) can be found – compounds that contain both carbocyclic steroid moiety and oxygen-enriched sugar fragments. In the central part, flavone-containing compounds can be found – polyflavonoids (R14), flavonoid aglycones (R11) and mono-glycosides (R7). However, flavonoid disaccharides are residing on the far south-east of the map (R6), next to the colchicines and oligopeptides (R19). At the same time,
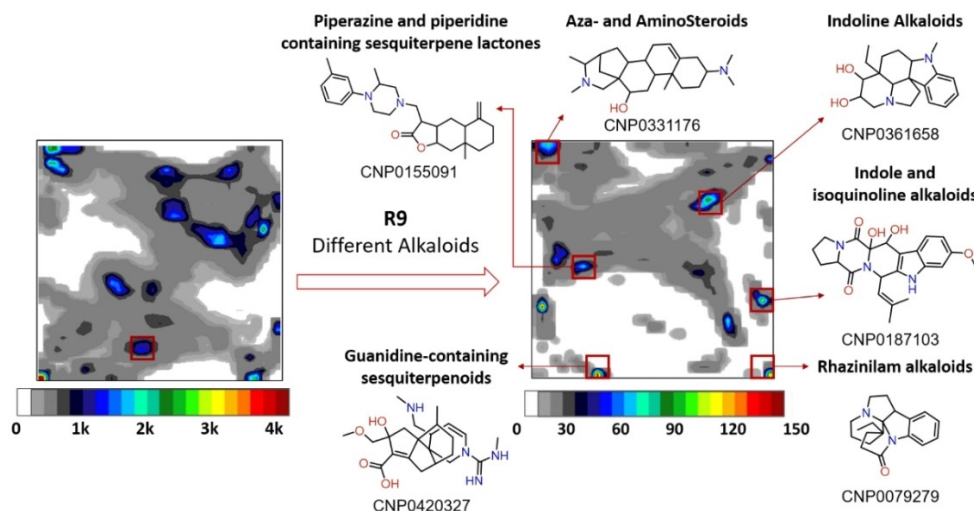
Figure 2. Zoomed density landscape for the region R9 that contains different type on alkaloids. On the finer scale of the zoomed map one can observe better chemotypes separation. Multicolored areas correspond to the highly populated regions, while gray color defines moderately occupied areas. White zones are empty

aminoacids and dipeptides (R20) are neighboring flavonoid monoglycosides from one side and large area of N-hetero-cycles – different types of alkaloids (R8, R9) – from another. Extreme south-west of the map is populated by numerous benzodioxol-containing compounds and their analogs.

Interestingly, nucleotides are not situated in the same regions – pyrimidine nucleoside phosphates (R18) reside close to phospholipids (R17) on the south-eastern part of the map, while purine nucleoside phosphates (R13) are found in the far west – neighboring the alkaloids area. Such distancing of (by human perception) similar compound subfamilies illustrates the competitive contribution of several underlying chemotypes to the compound's position in the chemical space. Pyrimidine nucleotides with their relatively smaller N-heterocycle moiety tend to be closer to phospholipids. In purines, N-heterocycles are dominant placing those compounds near the alkaloids area.

The NP-Umap supports a significant separation of the most common NP compound families, which makes it an efficient tool for NPs chemical space navigation. However, for more detailed structural analysis hierarchical zooming needs to be applied. In Figure 2, zooming of the alkaloid-containing region (R9) is shown as an example. With a better resolution, we can distinguish several density picks, corresponding to the different alkaloid subfamilies – piperazine and piperidine containing sesquiterpene lactones, guanidine-containing alkaloids, indoline, indole, isoquinoline and rhazinilam alkaloids. While all are members of one of the largest NP classes and thus to some extend similar, they nevertheless possess unique structural features that could be captured only with a help of HGTM.

3.3 Commercial Availability of Natural Products and Amount of Associated Biological Testing Data, as Functions of Drug-likeness

As already mentioned, multiple different landscapes can be created for a same map. They can be used separately or combined allowing to analyze projected compound libraries from different perspectives – comparing, for example, the availability of bioactivity test results versus commercial availability of NPs. COCONUT was intersected with ChEMBL and NP-like ZINC datasets, resulting in almost 45 K of biologically tested compounds and 11 K commercially available NPs, respectively. Their distribution within the entire COCONUT NP dataset is shown in Figure 3. The left-hand map is a fuzzy class landscape contrasting biologically untested NPs (COCONUT - ChEMBL) in black, versus experimentally tested NPs (COCONUT∩ChEMBL) in red, mixed regions in intermediate colors. On the right-hand map, commercially unavailable NPs (COCNUT – ZINC) – black regions – cover largely the same map zones as untested NPs (COCONUT-ChEMBL). It is no surprise that compounds that are difficult to access are not amongst the most tested ones. The middle map shows the COCONUT drug-likeness landscape, based on the drug-likeness (QED) score.[32] It varies from zero to one – the bigger the score the more drug-like properties the compound possesses. It appears that both biologically tested and commercially available NPs-enriched regions coincide fairly well with areas of the high QED values, showing that one of the driving forces of the NPs exploration in bioactivity and purchasability context is their physicochemical properties and thus their potential to be used as drugs. This is just one of many possible examples of how integrated analysis of multiple property landscapes can shed the light onto
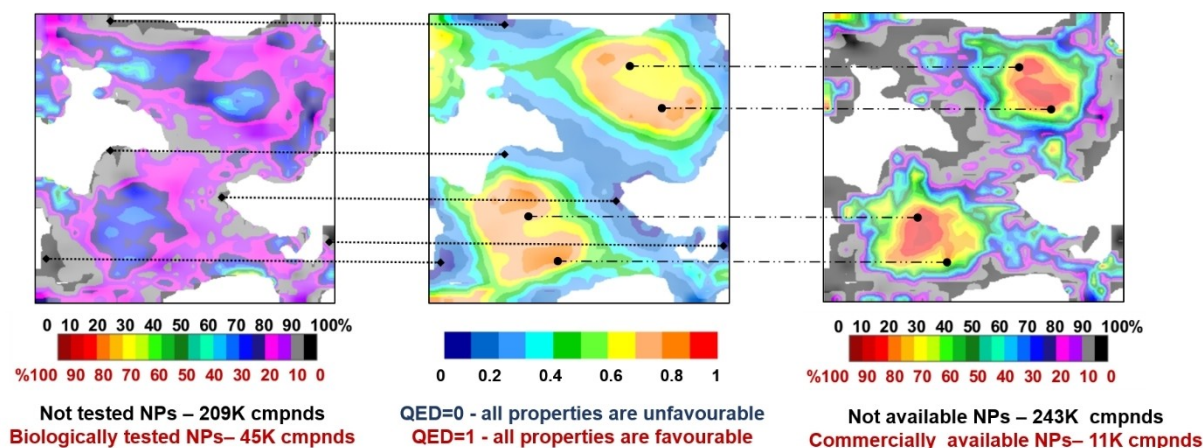
Figure 3. Amount of existing (ChEMBL-reported) NP bioactivity data and NP commercial availability relate to the drug-likeness of compounds. Map on the left - class landscape comparing biologically tested (red) and not tested (black) NPs. Map in the middle – property landscape showing distribution of quantitative estimate of drug-likeness (QED) of NPs. Blue regions correspond to the compounds with all physicochemical parameters being unfavorable for oral drugs, red ones – with all properties being favorable. Map on the right – normalized class landscape comparing commercially available (red) and not available (black) NPs.
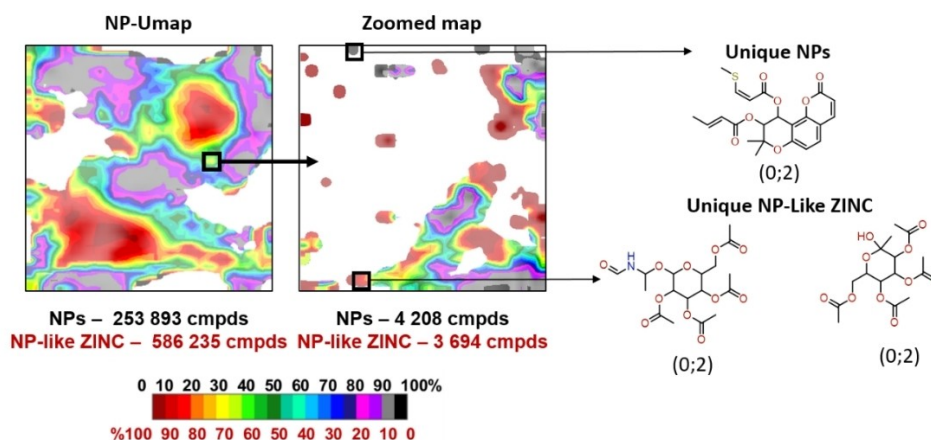
Figure 4. Examples of the zooming (HGTM) procedure in a search for NP-specific and ZINC-specific MCSs. First number in parenthesis gives number of hits in COCONUT, second one – in NP-like ZINC.

different aspects of the NPs chemical space providing generalized understanding of its global features.

3.4 Natural Products vs NP-like ZINC Compounds

The newly constructed NP-Umap is not limited only to NPs – any compounds populating the regions of the chemical space, covered by the map can be projected. Considering the neighborhood behavior principle,[33] those compounds should be structurally similar to the natural compounds used for GTM construction – NP-like compounds - and thus possess similar properties. Mapping the external dataset of the NP-like commercially available compounds and their structural comparison with NPs can provide valuable insight into similarities and differences between artificially synthe-

sized and naturally produced molecules. Reversely, pseudo-NP (synthetic analogs of natural compounds) detection of NP-zone residents stemming from synthetic sources can be easily performed.

Thus, 254k NPs and 586k NP-like ZINC compounds were projected onto NP-Umap. In Figure 4 the first map is a fuzzy class landscape where black regions correspond to the NPs and red – to the NP-like ZINC compounds. Even on the global "bird's-eye" scale of NP-Umap, regions significantly dominated by members of each library can be spotted. However, there are plenty of mixed zones, containing both NPs and commercially available NP-like compounds. In Figure 4, one example of the more detailed HTGM-based analysis is pursued. A mixed green zone (square of 3*3 nodes), containing 7 902 compounds with almost 50:50 ratio of members of each library, has been zoomed

resulting in a new map of finer scale with a better class separation – multiple regions occupied by compounds from only one library can be found. For further structural analysis of those regions, maximum common substructures (MCS) were used as a way to generalize structural features of compounds populating them. MCS was preferred over the popular scaffold concept due to its flexibility and adaptability. MCS can either contain only rings and linkers, in such a way coinciding with the corresponding scaffold or be more specific by including side-chain substituents if that is beneficial for capturing distinctive structural features of the analyzed libraries. Here we aimed to identify unique MCSs, found either only in COCONUT, or in NP-like ZINC respectively. As a result of the iterative HGTM application, 241 HGTMs have been built with up to two levels of zooming. With the help of those maps 15,891 locally NP-like ZINC-specific MCSs and 9,357 locally COCONUT-specific MCS have been found. "Locally specific" means that in the analyzed region this MCS occurs only in one library. However, as observed with the nucleotides, compounds sharing similar structural patterns can be situated far from each other on the map due to the contribution of other underlying chemotypes to the molecule position. As a result, locally specific MCSs may still be present in the other library, but outside the analyzed area. Therefore, an additional substructure search is needed to ensure (absolute) specificity of locally identified MCSs. NP-Like ZINC-specific MCSs have been checked against COCONUT NPs leaving only 12,981 ZINC-specific MCSs (10,545 of which are absent also in the uncleaned COCONUT dataset). Local NP-specific MCSs in their turn have been substructure-queried against the NP-like In-Stock ZINC library, with 8,282 MCSs returning no matches. However, 1,337 of these NP-specific MCSs have been found in the NP-like Tangible ZINC dataset, making compounds incarnating them purchasable in principle (acquisition success rate for tangible compounds is around 70%). The complete list of detected NP- and NP-LIke ZINC-specific chemotypes is available upon quick registration by the link https://forms.gle/LHQPvqitKEJv7e4K8.

Figure 5 displays the most often encountered NP-like ZINC-specific and COCONUT-specific MCSs. The first number in parenthesis represents occurrences in COCONUT, the second in NP-like ZINC. Among the ZINC-specific MCSs there are some purely synthetic chemotypes like bicyclo (1.1.1)pentane derivatives (R4) or dioxaborolanes (R3). However, some contain typical rings often seen in NPs e.g. furane (R5) or pyrrole (R17). Here, the ring substitution patterns typically produced by chemical synthesis are conferring ZINC-specificity to these MCS. There are also ZINC-specific MCSs representing synthetic peptidomimetics (R10) and synthetically modified natural compounds (R6). In any case, 90% of them contain nitrogens as key heteroatoms. In contrast, the majority of COCONUT-specific MCSs corresponds to the complex carbo- or oxoheterocycles with oxygen-containing sidechains. Thus, nitrogen-containing compounds and alkaloids, in particular, are better explored by synthetic chemistry than complex oxygen-containing NPs.

3.5 Biological Activity of Natural Products

As mentioned before, ChEMBL bioactivity data are available for about 45k NPs. Those compounds are almost evenly distributed around the map, typically within high QED regions (Figure 3). By contrast, the most common chemotypes for untested NPs (Figure 6) contain either complex ring systems or long hydrocarbon chains, shifting them outside of the drug-likeness domain.

NP-Umap can be also used for the target-based bioactivity analysis. Figure 7 and Figure 8 display fuzzy classification landscapes contrasting NP ligands of each of the target classes (C) used for NP-Umap optimization – black regions - against NPs active against all other targets reunited into one non-C class – red zones. Note – non-C pool does not include any of COCONUT compounds that were not labeled by activity class. Landscapes have been normalized due to the high dataset imbalance (mid-range color green corresponds to zones populated by classes C and non-C at local cumulated responsibility ratio equaling the default ratio of those set sizes). Target class-specific MCSs are shown below, except for the 70 enzyme-specific MCSs out of which only 5 most populated are shown.

3.6 NP Navigator

The hierarchical ensemble of maps was used as the basis for NP Navigator – a multifunctional tool for the analysis of the chemical space of NPs and NP-like molecules. It is openly accessible via web-interface by the https://infochm.-chimie.unistra.fr/npnav/chematlas_userspace. NP Navigator provides access to the library of multiple pregenerated property landscapes – density, various physico-chemical parameters, QED, ZINC vs NPs and ChEMBL vs NPs comparative landscapes, biological activity landscapes, etc. Each predefined zone (square of 3*3 nodes) of these maps is assigned to the NPs, NP-like ZINC and ChEMBL compounds populating it. Those compounds as well as MCSs characterizing them can be displayed and/or downloaded. If the zone was zoomed, the HGTM landscape will be shown prior to the associated compounds list. In such a way users can by themselves navigate through the chemical space of NPs and explore its different aspects. NP Navigator can be used for different purposes – chemical space analysis, NP-like libraries comparison (Figure 4 and Figure 5), searching for the NP-analogs of the compound of interest (Figure 9), analysis of the biological activity of NPs (Figure 7 - Figure 9).

The detailed description of NP Navigator can be found in the Supporting Information.
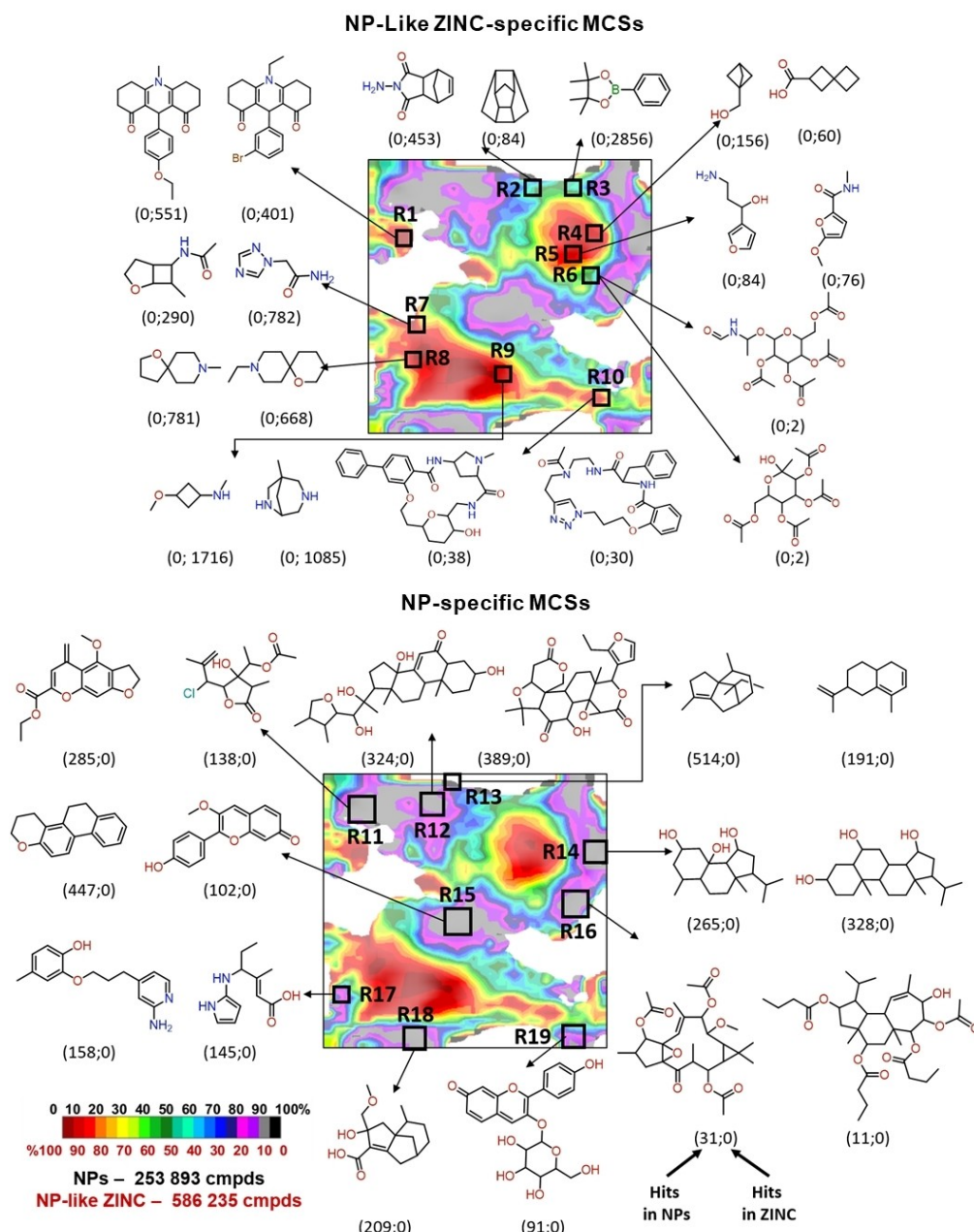
Figure 5. Class landscape comparing COCONUT natural products(black) with NP-like ZINC compounds(red). Upper scheme provides examples of ZINC-specific MCSs, while lower one demonstrates NP-specific MCSs. First number in parenthesis gives number of hits in c-COCONUT, second one – in NP-like ZINC

4 Conclusions

In this work, hierarchical GTM has been used to perform a thorough analysis of the chemical space represented by natural products. More than 200 HGTMs based on the universal map of natural products (NP-Umap) have been constructed. It has been shown that the ensemble of those maps – accessible via web-interface NP Navigator – provides a meaningful chemotypes separation, which can

be used for structural analysis of NPs and in a search of natural or synthetic analogs of the molecule of interest.

Comparison of COCONUT NPs and NP-like ZINC subsets resulted in almost 20 thousand unique MCSs, specific to only one library (https://forms.gle/LHQPvqitKEJv7e4K8). 90% of ZINC-specific MCSs contain a nitrogen atom. Concerning NPs-specific MCSs, the majority of them correspond to the complex carbo- or oxoheterocycles with oxygen-containing sidechains. This illustrates the well-
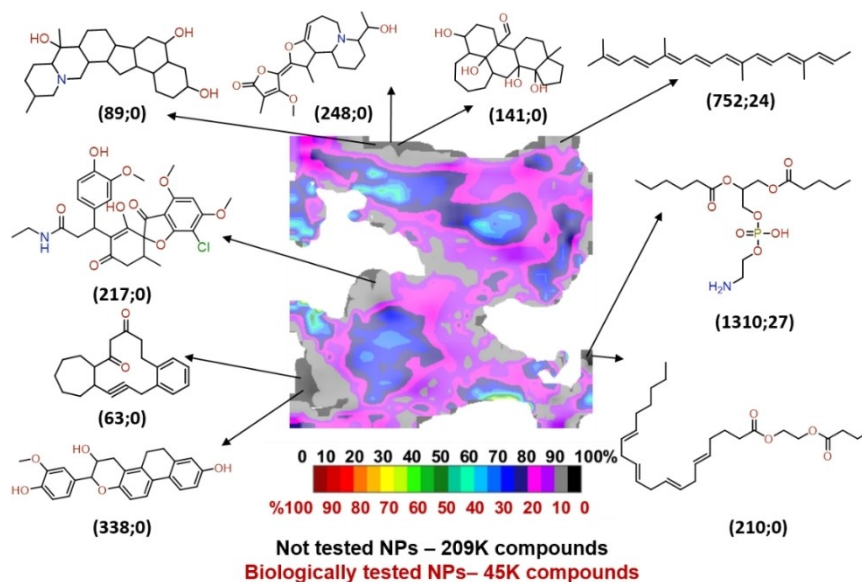
Figure 6. Class landscape comparing biologically tested (red) and not tested (black) NPs. Given substructures correspond to the MCSs, specific to the not tested subset. First number in parenthesis gives number of hits in not tested subset, second one – in tested.
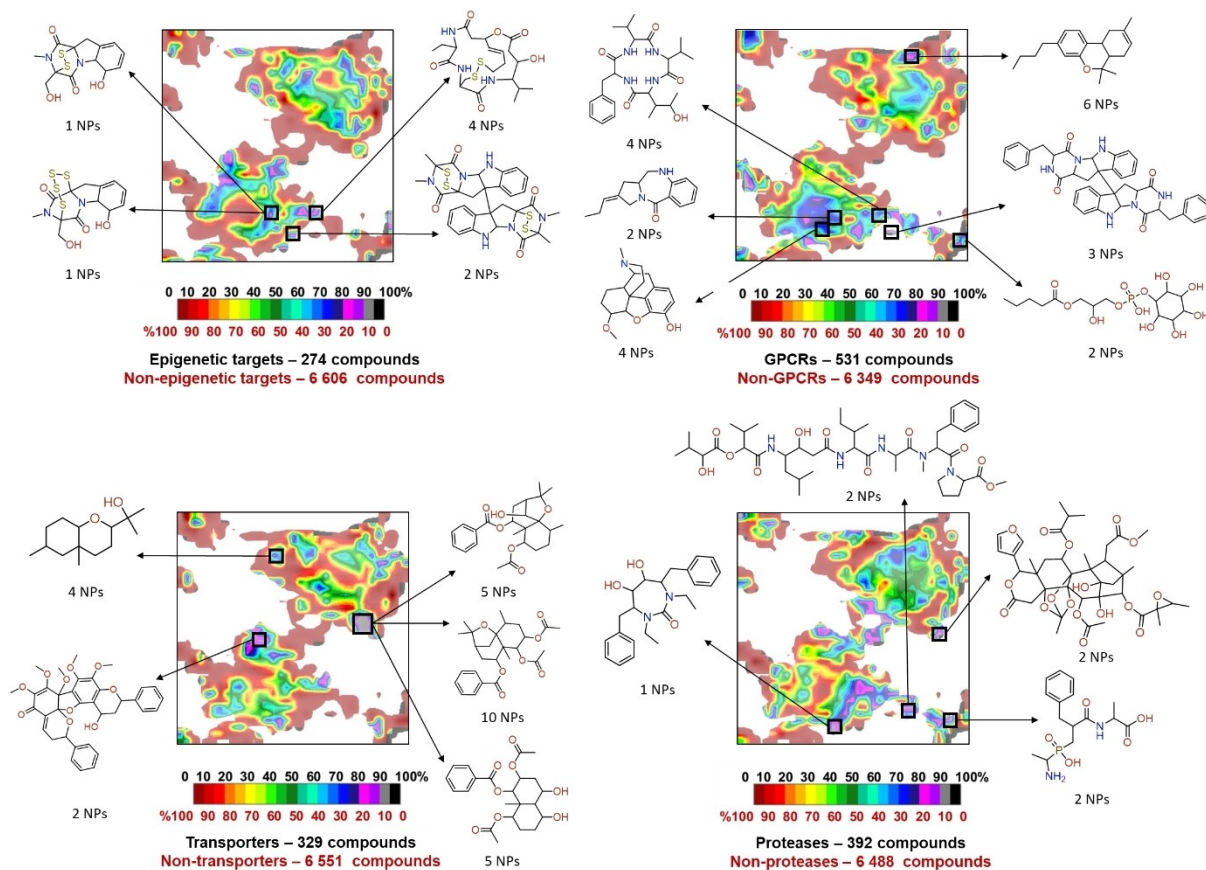
Figure 7. Target-specific NP chemotypes and corresponding regions of chemical space: epigenetic targets, GPCRs, transporters and proteases.
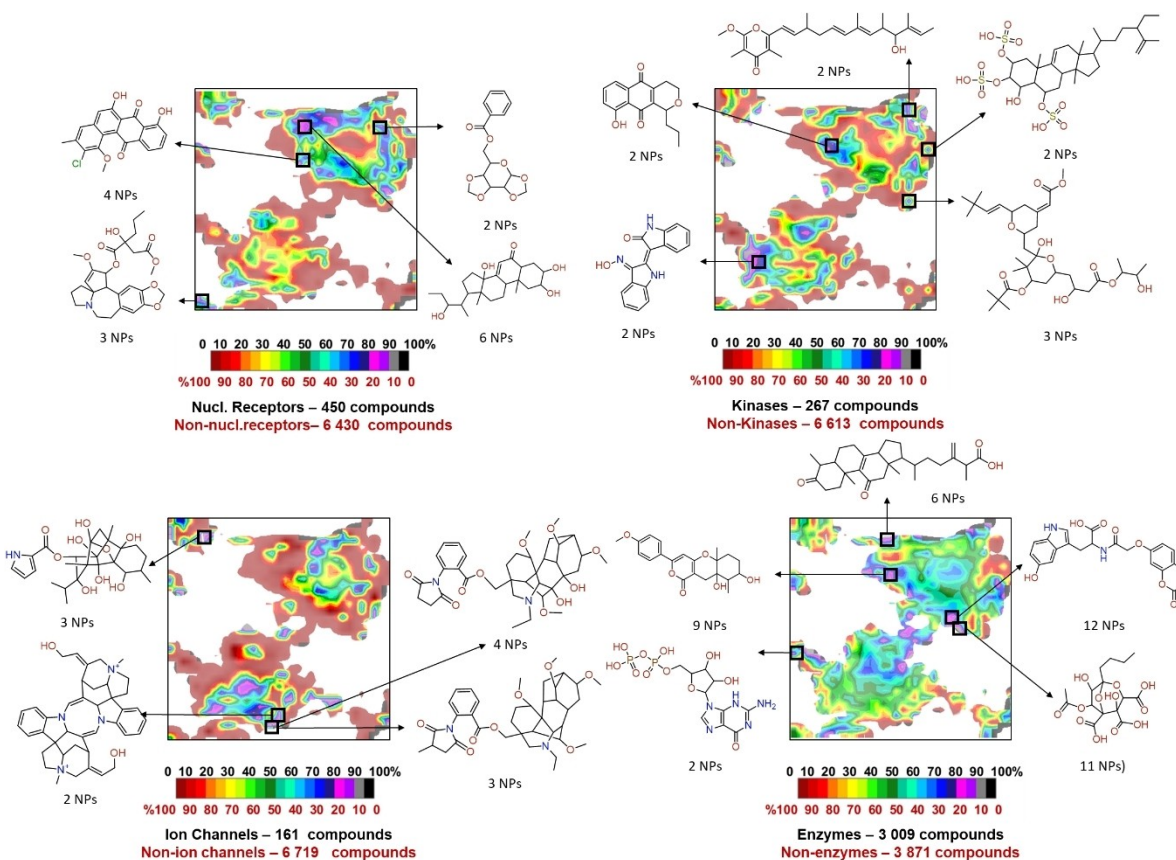
Figure 8. Target-specific NP chemotypes and corresponding regions of chemical space: nuclear receptors, kinases, ion channels and other enzymes
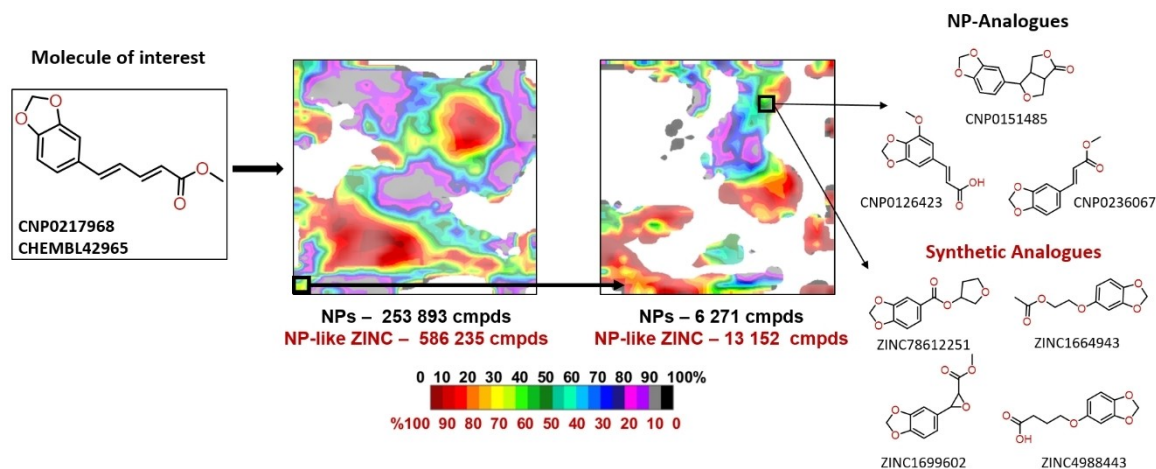
Figure 9. Search of the NPs and synthetic analogs of a compounds of interest using NP Navigator (241 GTM in total). After being projected onto the NP-Umap, compound is followed down to the last level of zoom. Neighboring compounds on the last zoomed map can be considered as a close NP-analogs and synthetic analogues of the initial compound of interest

known fact that nitrogen-containing compounds in general and alkaloids, in particular, are better explored by synthetic chemistry than complex oxygen-containing NPs. ZINC-specific MCSs, being the chemotypes found in NP-like ZINC but never occurring in NPs, can be used as a filtering set applicable together with NP-likeness score in order to improve NP-likeness of the designed library.

Biological activity of NPs has been also investigated. It was shown that one of the driving forces of NP-focused investigation for biomedical applications is their physico-chemical profile and thus their potential to be used as drugs – NPs with a higher QED score tend to appear more often in ChEMBL and ZINC than other compounds.

NPs active against popular target families (kinases, proteases, other enzymes, ion channels, nuclear receptors, GPCRs, epigenetic targets, transporters), have been analyzed in order to find characteristic structural features unique for each of the ligand series. However, it appears, that NP active against different target classes may significantly overlap in the chemical space if those targets are naturally "promiscuous" with respect to each other's ligands. Thus only a few specific MCSs have been found for each target-based subset.

Author Contribution

The manuscript was written through contributions of all authors, and all authors have given approval to the final version of the manuscript.

Data Availability Statement

The initial data used in this work were derived from the following resources available in the public domain: ChEMBL[25] (version 26) – https://www.ebi.ac.uk/chembl/, COCONUT[5b,17] (v.2020.4) – https://coconut.naturalproducts.-net/, ZINC20[18] – http://zinc20.docking.org/

The results of this study are openly available: NP Navigator – tool for chemical space analysis and exploration is available via a web interface in https://infochm.chimie.u-nistra.fr/npnav/chematlas_userspace, the list of unique COCONUT- and NP-like ZINC-specific Maximum Common Substructure (MCS) – in https://forms.gle/LHQPvqitKEJ-v7e4K8

References

[1] N. Dixon, L. S. Wong, T. H. Geerlings, J. Micklefield, *Nat. Prod. Rep.* **2007**, *24*, 1288–1310.

[2] a) R. Liu, X. Li, K. S. Lam, *Curr. Opin. Chem. Biol.* **2017**, *38*, 117–126; b) O. Ramström, J.-M. Lehn, *Nat. Rev. Drug Discovery* **2002**, *1*, 26–36.

[3] a) J. Inglese, D. S. Auld, *Wiley Encyclopedia of Chemical Biology* **2008**, 1–15; b) R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, G. S. Sittampalam, *Nat. Rev. Drug Discovery* **2011**, *10*, 188–195.

[4] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629–661.

[5] a) Y. Chen, C. de Bruyn Kops, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111; b) M. Sorokina, C. Steinbeck, *J. Cheminf.* **2020**, *12*, 20.

[6] a) S. Wetzel, A. Schuffenhauer, S. Roggo, P. Ertl, H. Waldmann, *Chimia* **2007**, *61*, 355–360; b) C. F. Stratton, D. J. Newman, D. S. Tan, *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4802–4807; c) T. A. Wenderski, C. F. Stratton, R. A. Bauer, F. Kopp, D. S. Tan, *Methods Mol. Biol.* **2015**, *1263*, 225–242; d) N. Singh, R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten, J. L. Medina-Franco, *J. Chem. Inf. Model.* **2009**, *49*, 1010–1024; e) P. Ertl, A. Schuffenhauer, in *Natural Compounds as Drugs: Volume II* (Eds.: F. Petersen, R. Amstutz), Birkhäuser Basel, Basel, **2008**, pp. 217–235; f) J. Rosén, J. Gottfries, S. Muresan, A. Backlund, T. I. Oprea, *J. Med. Chem.* **2009**, *52*, 1953–1962; g) X. Lucas, B. A. Grüning, S. Bleher, S. Günther, *J. Chem. Inf. Model.* **2015**, *55*, 915–924; h) M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227; i) T. Henkel, R. M. Brunne, H. Müller, F. Reichel, *Angew. Chem. Int. Ed.* **1999**, *38*, 643–647; *Angew. Chem.* **1999**, *111*, 688–691; j) M. L. Lee, G. Schneider, *J. Comb. Chem.* **2001**, *3*, 284–289; k) A. B. Yongye, J. Waddell, J. L. Medina-Franco, *Chem. Biol. Drug Des.* **2012**, *80*, 717–724; l) Y. Chen, M. Garcia de Lomana, N.-O. Friedrich, J. Kirchmair, *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532.

[7] a) F. L. Stahura, L. Godden, J. Fau-Xue, J. Xue, L. Fau-Bajorath, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2000** *40*, 1245–1252; b) P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68–74; c) K. Vanii Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, C. Steinbeck, *BMC Bioinf.* **2012**, *13*, 106.

[8] H. Lachance, K. Wetzel, S. Fau-Kumar, H. Kumar, K. Fau-Waldmann, H. Waldmann, *J. Med. Chem.* **2012**, *55*, 5989–6001.

[9] I. S.-G. Fernanda, B. A. Pilón-Jiménez, L. M.-F. José, *Phys. Sci. Rev.* **2018**, *4*, 20180103.

[10] F. I. Saldívar-González, E. Lenci, A. Trabocchi, J. L. Medina-Franco, *RSC Adv.* **2019**, *9*, 27105–27116.

[11] P. Ertl, T. Schuhmann, *Mol. Inf.* **2020**, *39*, 2000017.

[12] K. Grabowski, G. Baringhaus, K. Fau-Schneider, G. Schneider, *Nat. Prod. Rep.* **2008**, *25*, 892–904.

[13] T. Miyao, D. Reker, P. Schneider, K. Funatsu, G. Schneider, *Planta Med.* **2015**, *81*, 429–435.

[14] a) D. Probst, J.-L. Reymond, *J. Cheminf.* **2020**, *12*, 12; b) A. Capecchi, J.-L. Reymond, *Biomol. Eng.* **2020**, *10*; c) A. L. Chávez-Hernández, N. Sánchez-Cruz, J. L. Medina-Franco, *Biomol. Eng.* **2020**, *10*.

[15] J. A. van Santen, G. Jacob, A. L. Singh, V. Aniebok, M. J. Balunas, D. Bunsko, F. C. Neto, L. Castaño-Espriu, C. Chang, T. N. Clark, J. L. Cleary Little, D. A. Delgadillo, P. C. Dorrestein, K. R. Duncan, J. M. Egan, M. M. Galey, F. P. J. Haeckl, A. Hua, A. H. Hughes, D. Iskakova, A. Khadilkar, J.-H. Lee, S. Lee, N. LeGrow, D. Y. Liu, J. M. Macho, C. S. McCaughey, M. H. Medema, R. P. Neupane, T. J. O'Donnell, J. S. Paula, L. M. Sanchez, A. F. Shaikh, S. Soldatou, B. R. Terlouw, T. A. Tran, M. Valentine, J. J. J. van der Hooft, D. A. Vo, M. Wang, D. Wilson, K. E. Zink, R. G. Linington, *ACS Cent. Sci.* **2019**, *5*, 1824–1833.

[16] B. I. Díaz-Eufracio, O. Palomino-Hernández, A. Arredondo-Sánchez, J. L. Medina-Franco, *Mol. Inf.* **2020**, *39*, 2000035.

[17] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, C. Steinbeck, *Preprint* **2019**, *10.21203/rs.3.rs-75600/v1*.

[18] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, R. A. Sayle, *J. Chem. Inf. Model.* **2020**.

[19] C. M. Bishop, M. Svensén, C. K. I. Williams, *Neural Comput.* **1998**, *10*, 215–234.

[20] a) A. Lin, D. Horvath, V. Afonina, G. Marcou, J.-L. Reymond, A. Varnek, *ChemMedChem* **2018**, *13*, 540–554; b) A. Lin, B. Beck, D. Horvath, G. Marcou, A. Varnek, *J. Comput.-Aided Mol. Des.* **2019**; c) Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Voloch-nyuk, A. Varnek, *J. Chem. Inf. Model.* **2020**.

[21] M. Grigalunas, A. Burhop, A. Christoforow, H. Waldmann, *Curr. Opin. Chem. Biol.* **2020**, *56*, 111–118.

[22] *ChemAxon. JChem, Version 20.8.3, ChemAxon, Ltd: Budapest, Hungary* **2020**.

[23] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855–868.

[24] J. Schaub, A. Zielesny, C. Steinbeck, M. Sorokina, *J. Cheminf.* **2020**, *12*, 67.

[25] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic Acids Res.* **2019**, *47*, D930–D940.

[26] T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59–69.

[27] a) N. Kireeva, I. I. Baskin, H. A. Gaspar, D. Horvath, G. Marcou, A. Varnek, *Mol. Inf.* **2012**, *31*, 301–312; b) H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *Mol. Inf.* **2015**, *34*, 348–356; c) A. Lin, D. Horvath, G. Marcou, B. Beck, A. Varnek, *J. Comput.-Aided Mol. Des.* **2019**, *33*, 331–343; d) D. Horvath, G. Marcou, A. Varnek, *Drug Discovery Today Technol.* **2020**.

[28] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087–1108.

[29] I. Casciuc, Y. Zabolotna, D. Horvath, G. Marcou, J. Bajorath, A. Varnek, *J. Chem. Inf. Model.* **2019**, *59*, 564–572.

[30] D. Horvath, J. B. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.

[31] P. Tino, I. Nabney, *IEEE PAMI* **2002**, *24*, 639–656.

[32] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, A. L. Hopkins, *Nat. Chem.* **2012**, *4*, 90–98.

[33] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, L. E. Weinberger, *J. Med. Chem.* **1996**, *39*, 3049–3059.

Summary

The chemical space of NPs has been analyzed here, featuring the largest publicly available database of compounds with natural origin COCONUT. Due to the limited number of NPs in ChEMBL, the applicability of the previously constructed uMaps to the NP chemical space analysis is not appropriate. Indeed, as one can see in **Figure 20**, NPs (blue regions) agglutinate in specific zones. This forces a lot of different NP chemotypes to "collide" in the same nodes, preventing their meaningful separation and clustering. Therefore, optimization of the NP-dedicated Universal map (NP-uMap) is proven mandatory.
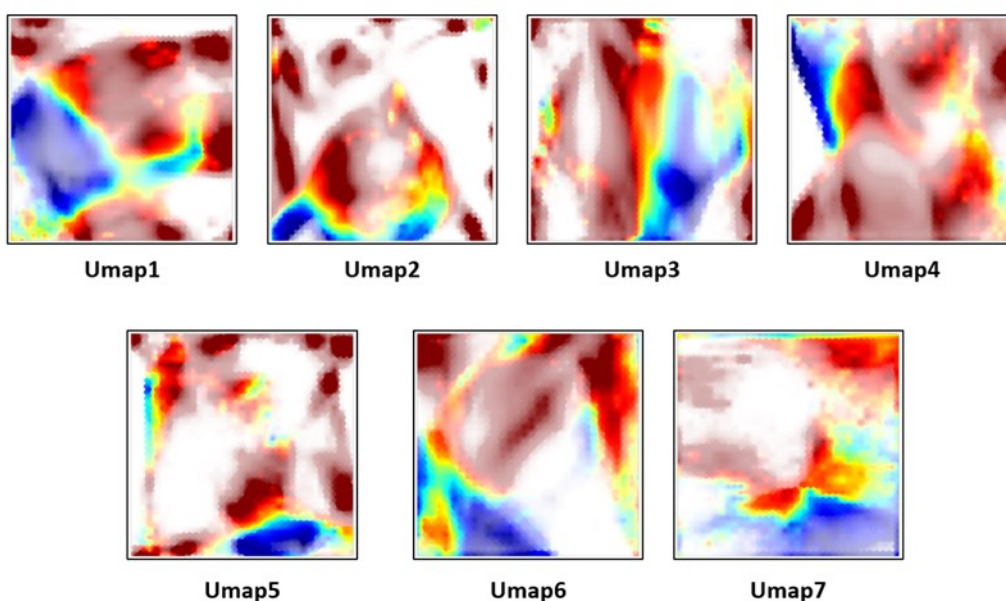
Figure 20. The comparative landscape of ChEMBL26 compounds (red regions) and NPs from COCONUT (blue regions) on the seven previously constructed universal maps of the ChEMBL chemical space.

New NP-uMap has been optimized and analyzed, demonstrating meaningful clusterization and NP classes separation. It consists of 1 225 nodes (35x35) coupled with 324 RBFs (18x18). The descriptors used to define NPs chemical space are ISIDA symmetrical atom-centered fragments with topological distance from 1 to 2 including both atoms and bonds information. The average BA in class separation is 0.67 (**Table 8**). This map is "NP-universal" in the sense that chosen set of parameters, including descriptor type, embodies a simultaneous capacity to satisfactorily separate NPs, active against nine

biologically unrelated target classes. This broadens NP-Umap application for chemical space analysis in a medicinal chemistry context.

Table 8. Genetic algorithm optimization of NP-uMap: pairwise class separation BA for the "best" manifold.

Pair of target classes	Class separation BA	Pair of target classes	Class separation BA
enzyme-epg	0.62	gpcr-other	0.66
enzyme-gpcr	0.67	gpcr-protease	0.67
enzyme-ic	0.71	gpcr-transporter	0.72
enzyme-kinase	0.62	ic-kinase	0.69
enzyme-nr	0.64	ic-nr	0.72
enzyme-other	0.60	ic-other	0.68
enzyme-protease	0.61	ic-protease	0.71
enzyme-transporter	0.67	ic-transporter	0.73
epg-gpcr	0.68	kinase-nr	0.65
epg-ic	0.71	kinase-other	0.60
epg-kinase	0.61	kinase-protease	0.62
epg-nr	0.67	kinase-transporter	0.71
epg-other	0.59	nr-other	0.65
epg-protease	0.64	nr-protease	0.68
epg-transporter	0.67	nr-transporter	0.68
gpcr-ic	0.66	protease-other	0.63
gpcr-kinase	0.71	protease-transporter	0.70
gpcr-nr	0.73	transporter-other	0.70

Hierarchical zooming has been applied to the zones with the highest density in order to increase the map's detalization ability. The repetitive hGTM application produced 241 hGTMs with up to two levels of zooming. The resulting maps were used to analyze and compare COCONUT to NP-like ChEMBL and ZINC subsets, revealing various structural features inherent to only one of the libraries.

The NPs, active against popular target families (kinases, proteases, other enzymes, ion channels, nuclear receptors, GPCRs, epigenetic targets, transporters), have been analyzed in order to find characteristic structural features unique for each of the ligand series. However, it appears that NPs active against different target classes are significantly overlapping in the chemical space. Thus only a few specific MCSs have been found for each target-based subset.

It has been shown that the ensemble of herein constructed maps provides a meaningful chemotypes separation, which can be used for both structural analysis of NPs and a search of natural or synthetic analogs of the molecule of interest. The resulting hierarchy of GTMs was used as a framework of the NP-Navigator - a part of ChemSpace Atlas web tool dedicated to analyzing the NP chemical space.

5 Chemspace Atlas – a polyvalent tool for the efficient exploration of chemical space

ChemSpace Atlas is an intuitive polyvalent tool for the efficient exploration of the ultra-large chemical space and its analysis with respect to medicinal chemistry problems. It is based on the hierarchical ensemble of tens of thousands GTMs, featuring biologically relevant chemical space. This hierarchy enables convenient navigation through the hundreds of millions of compounds from a global bird's eye view to structural pattern detection. One of the main advantages of such an approach is the ability to capture specific features of the chemical space, compare several libraries on different scales and perform structural analysis.

5.1 Featured chemical space and underlying ensemble of GTMs

As soon as drug design encompasses various strategies that significantly diverge in terms of relevant chemical space, ChemSpace Atlas was designed as a container for several subspace navigators: fragment-like, lead-like, drug-like, PPI-like natural products, and NP-like compounds, DNA-encoded libraries (DEL), and synthons navigators (**Figure 21**). The last two are under development and are not yet available online. Their interface and functionality will differ from those already implemented and will be discussed in the next chapter as a perspective for further ChemSpace Atlas development. In addition, there is a ChEMBL activity space Navigator and activity Profiler staring compounds with reported biological activity against 749 biological targets and enabling pharmacological profiling using consensus activity class prediction on seven universal maps, described in Chapter 3.
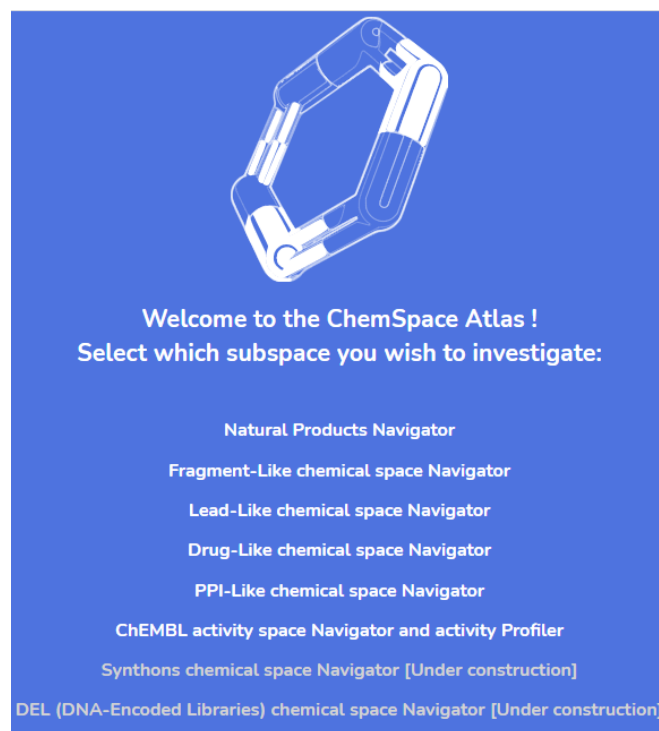
Figure 21. Starting page of ChemSpace Atlas.

Each of the navigators listed in **Figure 21** is focused on specific subspaces of the biologically relevant chemical space that differ in size (**Table 9**): from 10^5 in the case of synthons to 10^9 for DELs. Each of the eight navigators is based on the separate hierarchy of maps reported in previous chapters (**Table 10**). The uGTMs were evolved with the help of GA, which allowed optimal descriptor space and GTM parameters selection. Zoomed maps were then constructed using the parameters of the main map and frameset composed of compounds localized in a specific zone. The descriptors defining chemical spaces are different variations of ISIDA fragment descriptors from simple atom sequencing to complex variations labeled by force-field constants, formal charges, pharmacophoric features, and even position of reactive centers in the case of synthons. Apart from the libraries that have been already projected onto the hGTMs, new collections can be placed on these maps leaving numerous possibilities for further ChemSpace Atlas extensions.

5.2 Interface and functionality

From the main page of ChemSpace Atlas (https://chematlas.chimie.unistra.fr/), one can select the section of the chemical space to explore(**Figure 21**). The functionality of each of so far implemented navigators is the same:

- Physicochemical properties visualization (18 calculated properties)

- Activity visualization (749 ChEMBL activities)

- Activity prediction (749 ChEMBL activities)

- Tracking specific areas of the chemical space based on structural features

- Analogs search

- Structural analysis of selected regions of the chemical space with the help of MCSs

- Precomputed libraries comparison

Almost twenty various physicochemical properties and more than 700 activity landscapes allow users to analyze libraries from different perspectives. In order to facilitate navigation, a small set of "tracking" compounds can be provided by the user. These molecules will be projected onto the GTMs, appearing as dots on the selected landscapes. These dots will help to choose the zones of chemical space worth exploring in the context of users' needs. Apart from simple navigation, ChemSpace Atlas can be used for efficient analysis of underlying libraries - chemotype distribution, physicochemical properties, (reported and/or predicted) biological activity, and commercial availability. Moreover, activity prediction based on the consensus model of seven universal maps is also available.

Here the interface of ChemSpace Atlas is demonstrated on the example of NP Navigator. From the main page of NP Navigator, one can access the input page of ChemSpace tracker (**Figure 22**). Here, the user can provide a list of SMILES (**Figure 22 (1)**) or draw a molecular structure in the sketcher window (**Figure 22 (2)**). These molecules will play the role of chemical "trackers" that allow pinpointing the regions of the chemical space that the user wants to explore. On the right part of the page, the drop-down menus enable the choice of the type of map coloration, e.i. type of landscape (**Figure 22 (3)**).

Table 9. Description of eight navigators composing ChemSpace Atlas: featured libraries, their size, underlying uMap and the size of the hierarchy in case if hierarchical zooming was performed.

Navigator name	Featured libraries	Size of the analyzed chemical space	Main uMap	Number of hGTMs in hierarchy
Natural Products Navigator	COCONUT	253K	NP-uMap	241 hGTMs
	NP-Like ChEMBL	474K		
	NP-Like ZINC20	586K		
Fragment-Like chemical space Navigator	FL ChEMBL	15K	1st uMap of ChEMBL	880 hGTMs
	FL ZINC15 (stock)	103K		
	FL ZINC15 (tangible)	2.7M		
Lead-Like chemical space Navigator	LL ChEMBL	363K	1st uMap of ChEMBL	11 150 hGTMs
	LL ZINC15 (stock)	3.2M		
	LL ZINC15 (tangible)	329M		

Drug-Like chemical space Navigator	DL ChEMBL	668K	1st uMap of ChEMBL	22 325 hGTMs
	DL ZINC15 (stock)	5.1M		
	DL ZINC15 (tangible)	516M		
PPI-Like chemical space Navigator	PPIL ChEMBL	229K	1st uMap of ChEMBL	3 294 hGTMs
	PPIL ZINC15 (stock)	1.2K		
	PPIL ZINC15 (tangible)	603K		
ChEMBL activity space Navigator and activity Profiler	Visualization: ChEMBL (v26)	1.7M	1st uMap of ChEMBL	241 hGTMs
	Profiler: ChEMBL(v24)	1.6M		
Synthons chemical space Navigator [Under construction]	PBB synthons	799K	Synthons-uMap	-
	ChEMBL-derived synthons	372K		
DEL chemical space Navigator [Under construction]	2,5K generated DELs	2.5B	1st uMap of ChEMBL	-

Table 10. Description of nine universal maps behind ChemSpace Atlas.

uMap	Type of ISIDA descriptors	GTM parameters	Role in ChemSpace Atlas
1st uMap of ChEMBL	Sequences of atoms with a length of 2−3 atoms labeled by force field types and formal charge status, using all paths	Nodes: 41x41 RBFs:23x23	Fragment-like, Lead-like, Drug-like, PPI-like and DEL chemical space Navigatorsn; Activity profiler
2nd uMap of ChEMBL	Symmetrical atom-centered fragments of atom and bonds of 1−2 atoms labeled by force field types	Nodes:47x47 RBFs:29x29	Activity profiler
3rd uMap of ChEMBL	Sequences of atoms and bonds of a length 2−4 atoms labeled by pharmacophoric atom types and formal charges using all paths	Nodes:37x37 RBFs:19x19	Activity profiler
4th uMap of ChEMBL	Sequences of 2−7 atoms	Nodes:38x38 RBFs:19x19	Activity profiler

5th uMap of ChEMBL	Sequences of atoms and bonds of 2−4 atoms labeled by formal charge, using all paths	Nodes:37x37 RBFs:17x17	Activity profiler
6th uMap of ChEMBL	Sequences of atom pairs with a length of 2−6 intercalated bonds, labeled by Force Field type	Nodes:32x32 RBFs:30x30	Activity profiler
7th uMap of ChEMBL	Atom triplets labeled by pharmacophoric atom types with topological distance from 3 to 6 bonds	Nodes:36x36 RBFs:25x25	Activity profiler
NP-uMap	Symmetrical atom-centered fragments of atom and bonds of 1−2 atoms	Nodes: 35x35 RBFs: 18x18	NP and NP-like chemical space Navigator
Synthons-uMap	Symmetrical atom-centered fragments of atom and bonds of 1−2 atoms reactive centers positions information	Nodes: 29*29 RBFs: 25*25	Synthons chemical space Navigator

Figure 22. Input page on Chemspace Tracker. 1) zone of text input (SMILES); 2) Structure sketcher; 3) selection of up to 5 landscape types.

Upon compound submission, they will be standardized, filtered according to the NP-likeness score, fragmented to calculate respective descriptor vectors, and projected into the universal map. Selected in the previous step landscapes will be generated. The progress of the whole preparation procedure will be displayed on the Progress page. In case if provided compounds are out of AD of NP-navigator, the error message will be displayed here. It can happen in two cases – either the compound is not NP-like (NP-likeness score filtering with a lower limit of -0.5) or situated too far from the manifold and thus cannot be analyzed with its help.

After the projection, the user will be redirected to the main result page containing one of the selected landscapes (**Figure 23**). The colored background of the map corresponds to the library (libraries) that were selected as a basis for the landscape (in provided example - ZINC (red regions) and ChEMBL (black regions); all colors in between correspond to the areas occupied by both libraries). User-defined compounds are displayed as black dots (**Figure 23 (5)**).
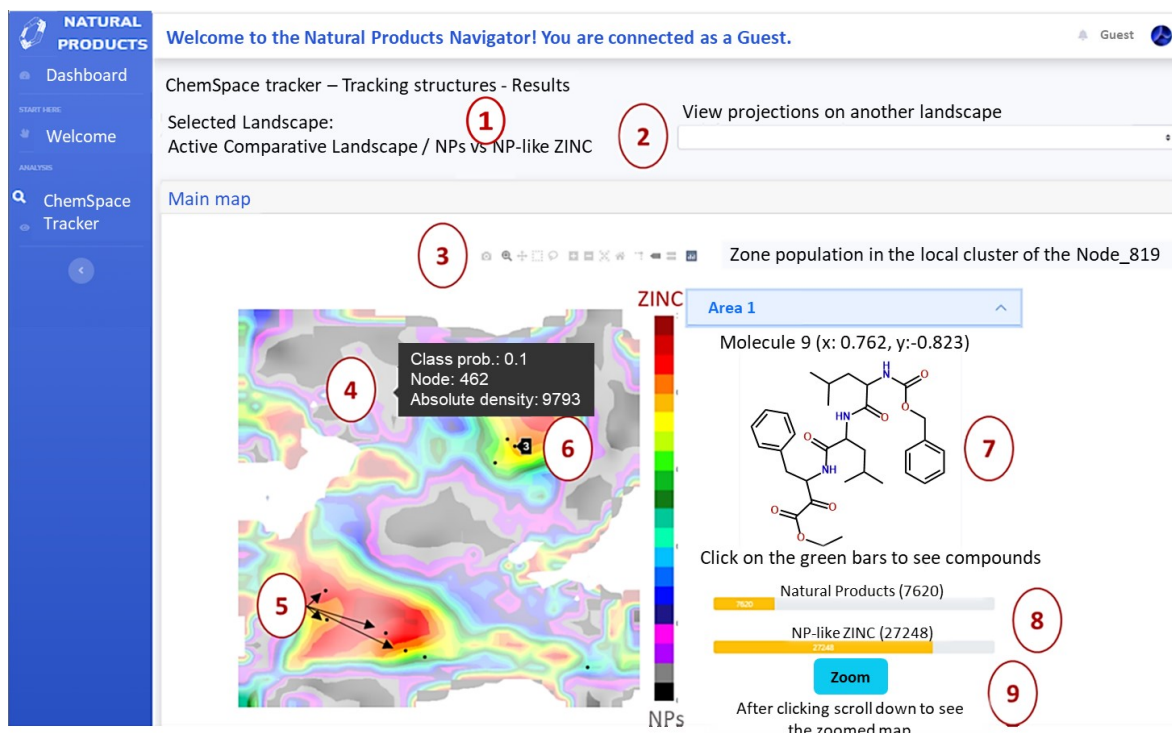
Figure 23. Main level landscape visualization: 1) type of the displayed landscape; 2) drop-down menu allowing to change displayed landscape; 3) Plotly toolbar allowing different types of navigation through the plot; 4) hoover-activated information about the node composition (Absolute density correspond approximately to the number of compounds residing in the node, and class probability indicates the proportion of NP(0) and ZINC(1) compounds); 5) black dots represent user-defined molecules - ChemSpace trackers; 6) hoover-activated ChemSpace tracker information (index number of compound in the provided list); 7) selected tracking compound; 8) the number of closest analogs of the selected compound on this level of hGTM (if green, bars become clickable and corresponding compounds can be displayed); 9) zoom button enabling display of the next level of navigation focusing on the selected zone of the chemical space.

After clicking on one of the dots, the respective compound will be shown on the right side of the map (**Figure 23 (7)**). Below the chemical structure, two bars illustrate the proportion of NP and NP-like ZINC compounds found in the closest surrounding of the selected "tracker" (**Figure 23 (8)**). As soon as bars are yellow, corresponding compounds cannot be displayed, as there are too many of them. In such a case, the "Zoom" button (**Figure 23 (9)**) should be present, allowing to visualize zoomed map – the next level of navigation (**Figure 24**).
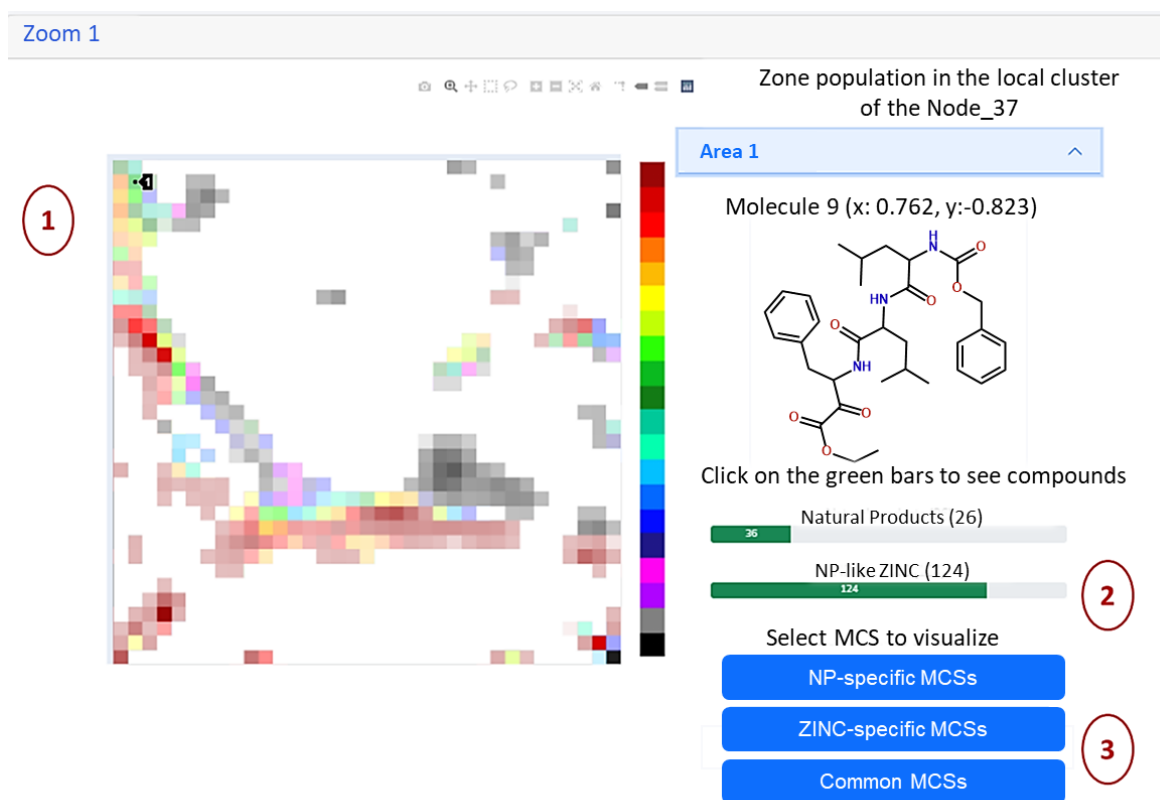
Figure 24. Zoomed level landscape visualization: 1) zoomed map with 1 "tracking" compound projection; 2) the number of closest analogs of the selected compound on this level of HGTM (if green, bars become clickable and corresponding compounds can be displayed); 3) buttons to perform structural analysis of the zone – the list of common and library-specific MCSs will be displayed.

Once the bars become green (**Figure 24 (2)**), the closest neighbors of the selected tracking compound can be displayed (**Figure 25**). The source identifiers provided for each molecule are hyperlinked to the corresponding library's web interface allowing direct access to the compound's information. One compound can have multiple identifiers if in the source library there were several stereoisomers. For simplicity reasons, stereochemistry was omitted in the analysis of ultra-large libraries. Therefore, all stereoisomer IDs were assigned to only one stereochemistry-depleted chemical structure. At the last level of zooming, MCSs analysis is available (**Figure 24 (3)**). Users can retrieve library-specific and common MCSs characterizing selected zone (**Figure 26**).
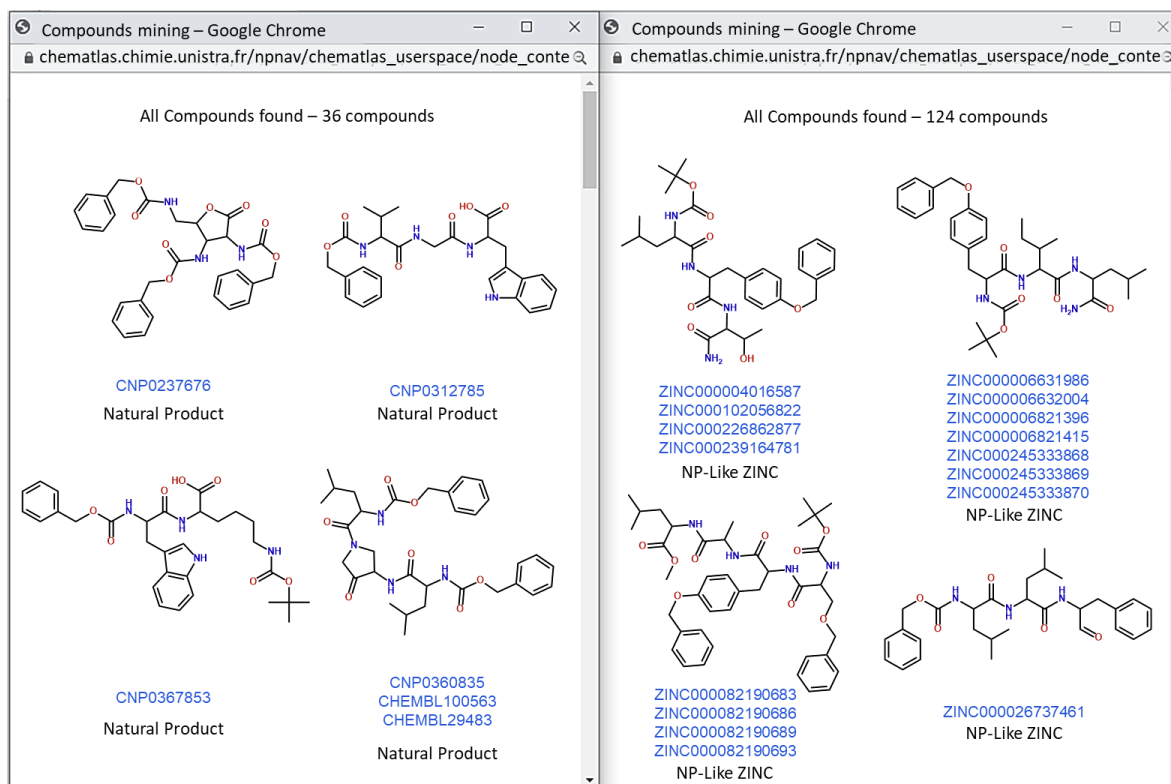
Figure 25. List of the closest analogs of the selected compound, their source, and external ID links. NPs (window on the left), that have been tested biologically or are available commercially will have not only COCONUT ids (CNPxxxxxx) but also ChEMBL or ZINC identifiers.

5.3 Technical details on web implementation

ChemSpace Atlas runs on an server version of Ubuntu 18.04 [131] with Apache 2.4[132] as an open-source HTTP webserver. An Anaconda[133] installation with Python 3.6 is linked to the Apache server. All physicochemical properties, respective landscapes, and MCSs are precomputed, hierarchically organized, and stored on a dedicated server. The ChemSpace Atlas front-end is developed with jQuery[134], a fast, lightweight, cross-browser, and feature-rich JavaScript library. The Bootstrap toolkit[135] is used to design the responsive interface. Chemical structures handling is done using two libraries: Epam sketcher[136] as a web-based chemical structure editor and OpenChemLib-js[137] for compounds visualization in 2D in the results pages.
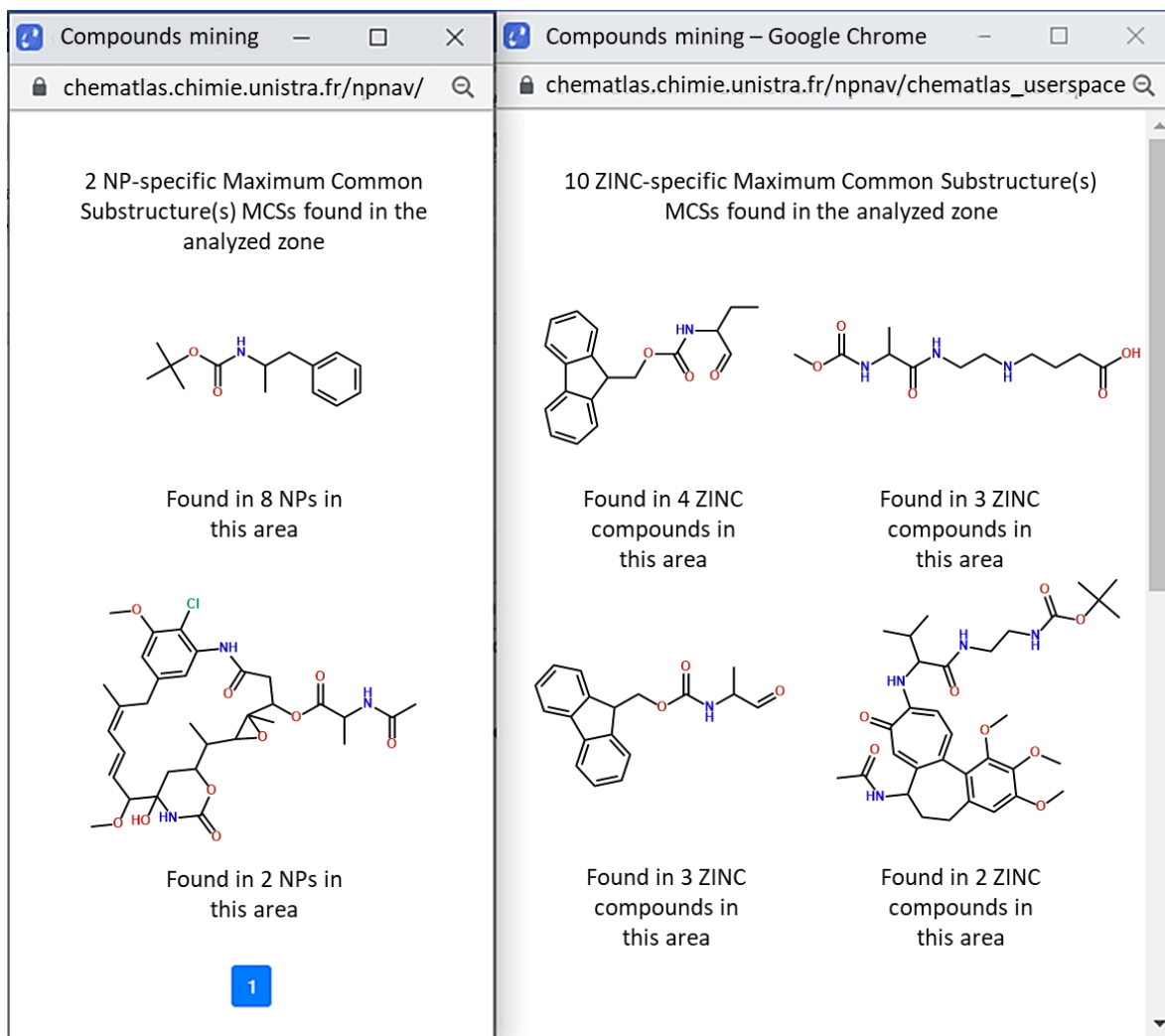
Figure 26. List of the library-specific MCSs characterizing selected zone.

The ChemSpace Atlas back-end is developed using custom PHP and Python CGIs that process the data entered by the user (either list of SMILES or single compound drown in sketcher). Standardization is performed using ChemAxon[117] Standardizer and pKa calculations plugins. Compounds projection followed by landscapes visualization is performed dynamically with custom Python scripts in the context of the ChemSpace tracker, using the Plotly library[138] version 4.8.

6   Conclusions and Perspectives

A close-up look at the chemical space for medicinal chemistry

The Big Data era in medicinal chemistry is marked by a boom of novel chemical and biological information reported on a daily basis. Even though currently available chemical libraries of synthesized and feasible compounds account for only a tiny portion of the anticipated number of all possible drug-like molecules (at least 10^{33} structures), they are far from being thoroughly studied and apprehended by medicinal chemists. Existing studies of the chemical space of purchasable screening compounds - one of the main sources of the hits in drug design campaigns - covered only a few percentages of the available chemical data. Moreover, there was a lack of analysis of their structural features and biological relevance. There was even less understanding of the chemical space of building blocks (BBs) used to synthesize the abovementioned screening compounds and DNA-encoded libraries (DELs).

Therefore, one of the main objectives of this thesis was to perform a detailed analysis of the compound libraries that medicinal chemists most frequently use in various stages of drug design: fragment-, lead-, drug-, PPI-like compounds, natural products (NPs) and DELs. As soon as Generative Topographic Mapping (GTM) has proven to be a highly efficient method of chemical space visualization and analysis, it was used as the main method for producing a 2D representation of the corresponding libraries. Ensemble of universal maps (uGTM) has been prepared in this work with the help of a genetic algorithm in order to separate biological activity (for ChEMBL molecules) or chemical reactivity classes (for synthons) in different groups of compounds. Each uGTM is able to simultaneously accommodate numerous predictive landscapes manifesting satisfactory performance in different classification/regression tasks. Moreover, universal maps have also proven to be efficient frameworks for the analysis and comparison of large chemical libraries. The hierarchical zooming (hGTM) applied to these maps allowed processing of the

unprecedented amount of data, increasing the limit for the size of analyzed libraries from previously reported 10^7 to studied herein 10^9 compounds.

As a result of detailed structural analysis of fragment-like, lead-like, drug-like, and PPI-like chemical spaces, several thousands of ChEMBL- and ZINC-specific maximum common substructures (MCSs) have been identified and made publicly available. ChEMBL-specific MCSs can be used as an inspiration for the stock-enhancement campaigns, while the ZINC-specific ones represent potential novel paths for the biological exploration of chemical space. It was also shown that the chemical space is unbalanced with a shift towards easily synthesizable sulfonamides, amides, etc. A similar imbalance, observed in "younger" tangible libraries, questions the efficiency of current stock enhancement techniques. Closer analysis of the most frequently used reactions and BBs may provide insight into the ways of improving these techniques.

Therefore, the first broad analysis of the purchasable BBs (PBBs) in a medicinal chemistry context has been performed. It required developing a new toolkit, SynthI, that employs synthons-based representation for the analysis of 150 different types of BBs without considering the leaving and protective groups. Moreover, SynthI also allows synthons generation as a result of pseudo-retrosynthetic fragmentation of reference compounds according to the 38 bond disconnection reaction rules. With the help of SynthI, GTM, and ISIDA descriptors, sensitive to the position of the reactive center in a synthon, it was shown that there is a lack of C- and S-nucleophiles and nucleophilic radicals, while O- and N-nucleophiles and electrophilic reagents are overrepresented in PBBs libraries. New synthons-uGTM has shown that only in the case of reagents for metathesis, acylation agents, O- and N-nucleophiles PBBs cover largely ChEMBL-derived synthons chemical space. For other groups of BBs, there are plenty of ChEMBL-specific areas of chemical space without any PBBs counterparts. Most of these areas correspond to the underrepresented on the market polyfunctional BBs. ChEMBL-derived synthons can serve as potential sources of inspiration for BBs libraries enhancement. In addition, the detailed GTM-based analysis of tangible BBs libraries and virtually generated ones, like GDB13, can also provide high-quality BBs structures, absent for now from in-stock libraries. Such investigation can significantly improve existing BBs libraries.

In another project, PBBs were used to enumerate the largest reported chemical space of DELs. Almost 2500 DELs have been designed with the help of eDesigner tool based on DNA-compatible reactions. For each library, 1M representative dataset has been generated from prefiltered PBBs and analyzed with GTM. With the help of the universal GTM, it was

shown that all 2.5B compounds that can be produced employing DEL technology cover largely the chemical space of biologically relevant compounds from ChEMBL. However, some small ChEMBL-specific areas are populated by complex natural products expectedly unreachable by DEL. GTM-based analysis of the regions of the chemical space populated by both ChEMBL and each separate DEL allowed us to identify the optimal DEL, covering the chemical space of ChEMBL to the highest extent and thus containing the maximum possible percentage of biologically relevant chemotypes.

Being the first chemoinformatics analysis of DELs of such scale, this project opens plenty of possibilities for future investigations. Indeed, a more detailed analysis of chemical structures, composing DELs, and their comparison to ChEMBL and commercially available compounds from ZINC will improve our understanding of the chemical space accessible via this technology with PBBs. Further GTM analysis and comparison of generated DELs can be helpful for the enhancement of available BBs libraries and prioritizing some promising synthetic procedures in order to improve the biological relevance of DEL chemical space. Another direction of DEL chemoinformatics research, which can also be handled with the help of GTM, is the development of the efficient methodology for BBs and reaction selection for the design of focused DELs - libraries structurally biased towards a particular class of biological targets.

ChemSpace Atlas as an efficient tool for chemical space navigation

Thousands of hierarchically related GTMs generated in our projects can be efficiently applied for highly informative analysis of featured and external libraries. However, the usage of GTM tools requires specific skills that are not necessarily components of medicinal chemists' training. Therefore, the second objective of this thesis was to develop the web interface incorporating all hGTMs created herein in order to enable easy and efficient usage of GTM for ultra-large chemical space navigation and analysis.

As a result, a highly polyfunctional web tool that allows navigating through the chemical space of unprecedently large libraries was created and made freely available. More than 40 thousand hierarchically related GTMs enable intuitive navigation through the hundreds of millions of compounds. The distinctive feature of the ChemSpace Atlas comparing to other online tools is that it allows users to analyze ultra-large libraries on different scales: from a global bird's eye view of the whole dataset to structural pattern detection in small clusters. A user-defined compound set can be used to "track" the chemical space regions containing molecules with specific structural features. It also can be used for

analogs search. Almost twenty precomputed physicochemical properties and thousands of MCSs characterizing each zone enable a detailed analysis of featured libraries in a different context. More than 700 biological activities from ChEMBL can also be visualized and pharmacological profiling using consensus of seven universal maps is available.

ChemSpace Atlas was designed as a container for several subspace navigators: ChEMBL, fragment-like, lead-like, drug-like, PPI-like, natural products, and NP-like compounds, DELs, and synthons navigators. Considering the scale of analyzed data, incorporating the results of the performed GTM analysis is a long-lasting process. Therefore, the ChemSpace atlas content is constantly updating. The functionality of the first five navigators is virtually equivalent and is described in this thesis. Implementation of the DEL and synthons navigators, based on the work reported here, would still demand additional efforts. Indeed, the detailed GTM-based analysis and structural comparison of all DELs to ZINC and ChEMBL that would allow creating a hierarchical navigator were not yet performed. Moreover, the DEL navigator will have extended functionality, allowing to compare all 2.5K libraries to the reference one (e.g., actives of a selected biological target). It will allow the selection of the best-suited DELs for each particular task. On the other hand, synthons navigator functionality might be coupled with SynthI, allowing users to analyze synthons generated from the user-provided list of BBs or reference libraries.

In the future, ChemSpace Atlas should not be limited to the navigators and libraries featured in this thesis. They are simply a starting core that can easily be updated in order to increase functionality, the scope of analyzed chemical space, or even the domain of its application. One of the possible directions of improvement can be the analysis and prediction of ADMETox properties, which was not considered herein.

Another significant functionality to include in any tool used in drug design is de novo compound generation. It allows the generation of novel compounds with desired pharmacological properties[139]. The autoencoder sequence-to-sequence neural network has already been combined with GTM in recent work by Sattarov et al.[140] The incorporation of such methodology in ChemSpace Atlas will complement its usage by introducing the guided rational exploration of the novel regions of the chemical space.

7 List of abbreviations

AD	Applicability Domain
AT	Activity Threshold
BA	Balanced Accuracy
BB	Building Block
BHA	Buchwald-Hartwig Amination
COCONUT	COlleCtion of Open Natural prodUcTs
CSN	Chemical Space Networks
DEL	DNA-Encoded Library
DUD	Directory of Useful Decoys
EC50	Half maximal Effective Concentration
FSc	Fitness Score
Fsp3	Fraction of saturated carbons
GA	Genetic Algorythm
GTM	Generative Topographic Mapping

hGTM	Hierarchical Generative Topographic Mapping
HTS	High Throughput Screening
IC50	Half maximal Inhibitory Concentration
iGTM	Incremental Generative Topographic Mapping
Ki	Inhibitory constant
LSH	Locality Sensitive Hashing
MCS	Maximum Common Substructure
MW	Molecular Weight
NB	Neighborhood Behavior
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NN	Nearest Neighbor
NP	Natural Products
PBB	Purchasable Building Blocks
PC	Principal Components
PCA	Principal Component Analysis
PPI	Protein-protein Interactions
QSAR	Quantitative Structure−Activity Relationship
QSPR	Quantitative Structure−Property Relationship

RBF   Radial Basis Functions

RP   Responsibility Patterns

SOM   Self-Organazing Maps

SVM   Support Vectors Machines

TMAP   Tree Map

t-SNE   t-distributed Stochastic Neighbor Embedding

uGTM   Universal Generative Topographic Mapping

VS   Virtual Screeening

8 References

1.	Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* **2014,** 19, 859-68.

2.	Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998,** 10, 215-234.

3.	Tino, P.; Nabney, I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE PAMI* **2002,** 24, 639-656.

4.	Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. Diversifying chemical libraries with generative topographic mapping. *J. Comput. Aided Mol. Des.* **2019**.

5.	Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J Chem Inf Model* **2019,** 59, 564-572.

6.	Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **2003,** 8, 876-877.

7.	Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty years on: the impact of fragments on drug discovery. *Nat Rev Drug Discov* **2016,** 15, 605-619.

8.	Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008,** 51, 817-34.

9.	Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004,** 8, 255-263.

10.	Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000,** 44, 235-49.

11.	Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004,** 1, 337-341.

12.	Morelli, X.; Bourgeas, R.; Roche, P. Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.* **2011,** 15, 475-81.

13.	Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J. L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018,** 13, 540-554.

14.	Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of commercially available compounds for HTS. *Drug Discov. Today* **2019,** 24, 390-402.

15.	Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S. DNA Encoded Libraries: A Visitor's Guide. *Isr. J. Chem.* **2020,** 60, 268-280.

16.	Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015,** 20, 11-7.

17.   Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: a New Look at the Natural Product Chemical Space. *Mol Inform, doi:10.1002/minf.202100068* **2021**.

18.   Erhardt, P. W.; Proudfoot, J. R. Drug Discovery: Historical Perspective, Current Status, and Outlook. In *Comprehensive Medicinal Chemistry II*, Taylor, J. B.; Triggle, D. J., Eds. Elsevier: Oxford, 2007; pp 29-96.

19.   Ciancetta, A.; Jacobson, K. A. Breakthrough in GPCR Crystallography and Its Impact on Computer-Aided Drug Design. *Methods Mol Biol* **2018,** 1705, 45-72.

20.   Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958,** 181, 662-6.

21.   Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T. L.; Homans, S. W.; Kessler, H.; Luchinat, C.; Meyer, B.; Oschkinat, H.; Peng, J.; Schwalbe, H.; Siegal, G. Perspectives on NMR in drug discovery: a technique comes of age. *Nat Rev Drug Discov* **2008,** 7, 738-45.

22.   Rabi, I. I.; Zacharias, J. R.; Millman, S.; Kusch, P. A New Method of Measuring Nuclear Magnetic Moment. *Physical Review* **1938,** 53, 318-318.

23.   Sugiki, T.; Furuita, K.; Fujiwara, T.; Kojima, C. Current NMR Techniques for Structure-Based Drug Discovery. *Molecules* **2018,** 23, 148.

24.   De Carlo, S.; Rémigy, H.-W. Cryo-electron Microscopy as a Tool for Drug Discovery in the Context of Integrative Structural Biology. *Structural Biology in Drug Discovery* **2020**, 613-632.

25.   Trobe, M.; Burke, M. D. The Molecular Industrial Revolution: Automated Synthesis of Small Molecules. *Angew. Chem. Int. Ed. Engl.* **2018,** 57, 4192-4214.

26.   Liu, R.; Li, X.; Lam, K. S. Combinatorial chemistry in drug discovery. *Curr. Opin. Chem. Biol.* **2017,** 38, 117-126.

27.   Ramstrom, O.; Lehn, J. M. Drug discovery by dynamic combinatorial libraries. *Nat Rev Drug Discov* **2002,** 1, 26-36.

28.   Inglese, J.; Auld, D. S. High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. *Wiley Encyclopedia of Chemical Biology* **2008**, 1-15.

29.   Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011,** 10, 188-95.

30.   Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007,** 25, 71-5.

31.   Collins, F. S.; McKusick, V. A. Implications of the Human Genome Project for medical science. *JAMA* **2001,** 285, 540-4.

32.   Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003,** 43, 218-27.

33.   Kodadek, T. The rise, fall and reinvention of combinatorial chemistry. *Chem Commun (Camb)* **2011,** 47, 9757-63.

34.   van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J Chem Inf Model* **2019,** 59, 644-651.

35.   Reymond, J. L. The chemical space project. *Acc. Chem. Res.* **2015,** 48, 722-30.

36.   Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J Chem Inf Model* **2015,** 55, 1824-35.

37.   Patel, H.; Ihlenfeldt, W. D.; Judson, P. N.; Moroz, Y. S.; Pevzner, Y.; Peach, M. L.; Delannee, V.; Tarasova, N. I.; Nicklaus, M. C. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci Data* **2020,** 7, 384.

38.   Humbeck, L.; Weigang, S.; Schafer, T.; Mutzel, P.; Koch, O. CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein-Protein Interaction Modulators. *ChemMedChem* **2018,** 13, 532-539.

39.   Shivanyuk, A.; Ryabukhin, S. V.; Bogolubsky, A. V.; Tolmachev, A. Enamine REAL database: making chemical diversity real. *Chimica Oggi-Chemistry Today* **2007**, 58-59.

40.   Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020,** 23, 101681.

41.   Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019,** 47, D1102-D1109.

42.   Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J Chem Inf Model* **2020,** 60, 6065-6073.

43.   Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J Chem Inf Model* **2016,** 56, 1253-66.

44.   Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with feature trees. *J Chem Inf Model* **2009,** 49, 270-9.

45.   Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A. Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb Sci* **2012,** 14, 579-89.

46.   Probst, D.; Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* **2020,** 12, 12.

47.   Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982,** 43, 59-69.

48.   Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **2015,** 34, 348-56.

49.   Kireeva, N.; Baskin, II; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012,** 31, 301-12.

50.   Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-task generative topographic mapping in virtual screening. *J. Comput. Aided Mol. Des.* **2019,** 33, 331-343.

51.   Nicola, G.; Liu, T.; Gilson, M. K. Public domain databases for medicinal chemistry. *J. Med. Chem.* **2012,** 55, 6987-7002.

52.   Kim, S. Public Chemical Databases. In *Encyclopedia of Bioinformatics and Computational Biology*, Ranganathan, S.; Gribskov, M.; Nakai, K.; Schönbach, C., Eds. Academic Press: Oxford, 2019; pp 628-639.

53.   Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003,** 10, 980.

54.   Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** 28, 235-42.

55.   Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2013,** 41, D36-42.

56.   UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021,** 49, D480-D489.

57.   Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007,** 35, D198-201.

58.   Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019,** 47, D930-D940.

59.   Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** 40, D1100-7.

60.   Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006,** 34, D668-72.

61.   Sterling, T.; Irwin, J. J. ZINC 15--Ligand Discovery for Everyone. *J Chem Inf Model* **2015,** 55, 2324-37.

62.   *eMolecules, Inc. https://www.emolecules.com/.*

63.   Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Brief Bioinform* **2021,** 22, 1790-1818.

64.   CHEMBL23. In May 2017 ed.; 10.6019/CHEMBL.database.23.

65.   CHEMBL24. In May 2018 ed.; 10.6019/CHEMBL.database.24.

66.   CHEMBL25. In March 2019 ed.; 10.6019/CHEMBL.database.25.

67.   CHEMBL26. In March 2020 ed.; 10.6019/CHEMBL.database.26.

68.   CHEMBL28. In Feb 2021 ed.; 10.6019/CHEMBL.database.28.

69.   Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006,** 49, 6789-801.

70.   Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012,** 52, 1757-68.

71.   Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **2005,** 45, 177-82.

72.   Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J Cheminform* **2021,** 13, 2.

73.   Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *J Cheminform* **2020,** 12, 20.

74.   Chen, C. Y. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* **2011,** 6, e15939.

75.   Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M. T.; Piperno, A.; Rescifina, A. Putative Inhibitors of SARS-CoV-2 Main Protease from A Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Mar Drugs* **2020,** 18.

76.   Zeng, X.; Zhang, P.; Wang, Y.; Qin, C.; Chen, S.; He, W.; Tao, L.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res.* **2019,** 47, D1118-D1127.

77.   Banerjee, P.; Erehman, J.; Gohlke, B. O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II--a database of natural products. *Nucleic Acids Res.* **2015,** 43, D935-9.

78.   Carbó-Dorca, R. About the concept of Chemical Space: a concerned reflection on some trends of modern scientific thought within theoretical chemical lore. *J. Math. Chem.* **2012,** 51, 413-419.

79.   Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality reduction: a comparative review. *J Mach Learn Res* **2009,** 10, 66-71.

80.   Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* **2007,** 47, 47-58.

81.   Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996,** 39, 2887-93.

82.   Maggiora, G. M.; Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput. Aided Mol. Des.* **2014,** 28, 795-802.

83.   Lu, J.; Carlson, H. A. ChemTreeMap: an interactive map of biochemical similarity in molecular datasets. *Bioinformatics* **2016,** 32, 3584-3592.

84.   Bawa, M.; Condie, T.; Ganesan, P. LSH forest. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*, Association for Computing Machinery: Chiba, Japan, 2005; pp 651–660.

85.   Dobson, C. M. Chemical space and biology. *Nature* **2004,** 432, 824-8.

86.   Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010,** 1, 30-38.

87.   Wenderski, T. A.; Stratton, C. F.; Bauer, R. A.; Kopp, F.; Tan, D. S. Principal component analysis as a tool for library design: a case study investigating natural products, brand-name drugs, natural product-like libraries, and drug-like libraries. *Methods Mol Biol* **2015,** 1263, 225-42.

88.   van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008,** 9, 2579--2605.

89.   Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010,** 29, 855-68.

90.   Villoutreix, B. O.; Lagorce, D.; Labbé, C. M.; Sperandio, O.; Miteva, M. A. One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discov. Today* **2013,** 18, 1081-1089.

91.   Awale, M.; van Deursen, R.; Reymond, J.-L. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013,** 53, 509-518.

92.   Gütlein, M.; Karwath, A.; Kramer, S. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. *J. Cheminformatics* **2012,** 4, 7.

93.   Janssen, A. P. A.; Grimm, S. H.; Wijdeven, R. H. M.; Lenselink, E. B.; Neefjes, J.; van Boeckel, C. A. A.; van Westen, G. J. P.; van der Stelt, M. Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome–Inhibitor Interaction Landscapes. *J. Chem. Inf. Model.* **2019,** 59, 1221-1229.

94.   González-Medina, M.; Medina-Franco, J. L. Platform for Unified Molecular Analysis: PUMA. *J. Chem. Inf. Model.* **2017,** 57, 1735-1740.

95.   Borrel, A.; Kleinstreuer, N. C.; Fourches, D. Exploring drug space with ChemMaps.com. *Bioinformatics* **2018,** 34, 3773-3775.

96.   Karlov, D. S.; Sosnin, S.; Tetko, I. V.; Fedorov, M. V. Chemical space exploration guided by deep neural networks. *RSC Adv.* **2019,** 9, 5151-5157.

97.   Donmez, A.; Rifaioglu, A. S.; Acar, A.; Doğan, T.; Cetin-Atalay, R.; Atalay, V. iBioProVis: interactive visualization and analysis of compound bioactivity space. *Bioinformatics* **2020,** 36, 4227-4230.

98.   Cortés-Cabrera, Á.; Morreale, A.; Gago, F.; Abad-Zapatero, C. AtlasCBS: a web server to map and explore chemico-biological space. *J. Comput. Aided Mol. Des.* **2012,** 26, 995-1003.

99.   Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminformatics* **2020,** 12, 12.

100.   Awale, M.; Reymond, J.-L. Web-based 3D-visualization of the DrugBank chemical space. *J. Cheminformatics* **2016,** 8, 25.

101.   Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space. *J. Nat. Prod.* **2007,** 70, 789-794.

102.   Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2018,** 34, 1433-1435.

103.   Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges* **2014,** 5, 450-472.

104.   Sidorov, P.; Viira, B.; Davioud-Charvet, E.; Maran, U.; Marcou, G.; Horvath, D.; Varnek, A. QSAR modeling and chemical space analysis of antimalarial compounds. *J. Comput. Aided Mol. Des.* **2017,** 31, 441-451.

105.   Kayastha, S.; Horvath, D.; Gilberg, E.; Gutschow, M.; Bajorath, J.; Varnek, A. Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. *J Chem Inf Model* **2017,** 57, 1218-1232.

106.   Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. *J Chem Inf Model* **2016,** 56, 1438-54.

107.   Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping in drug design. *Drug Discov Today Technol* **2019,** 32-33, 99-107.

108.   Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J Chem Inf Model* **2015,** 55, 84-94.

109.   Lin, A.; Baskin, II; Marcou, G.; Horvath, D.; Beck, B.; Varnek, A. Parallel Generative Topographic Mapping: An Efficient Approach for Big Data Handling. *Mol. Inform.* **2020,** 39, e2000009.

110.   Williams, C. B. a. C. K. I. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998,** 21, 203-224.

111.   Orlov, A. A.; Khvatov, E. V.; Koruchekov, A. A.; Nikitina, A. A.; Zolotareva, A. D.; Eletskaya, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Horvath, D.; Osolodkin, D. I.; Varnek, A. Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol. Inform.* **2019,** 38, e1800166.

112.   Visini, R.; Awale, M.; Reymond, J. L. Fragment Database FDB-17. *J Chem Inf Model* **2017,** 57, 700-709.

113.   Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J Chem Inf Model* **2012,** 52, 2864-75.

114.   Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided Mol. Des.* **2015,** 29, 1087-108.

115.   Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*, Liu, L.; ÖZsu, M. T., Eds. Springer US: Boston, MA, 2009; pp 532-538.

116.   Attene-Ramos, M. S.; Austin, C. P.; Xia, M. High Throughput Screening. In *Encyclopedia of Toxicology*, Wexler, P., Ed. Academic Press: Oxford, 2014; pp 916-917.

117.   *ChemAxon. JChem, Version 20.8.3, ChemAxon, Ltd: Budapest, Hungary* **2020**.

118.   Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **2008,** 3, 435-44.

119.   Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **2012,** 55, 9763-72.

120.   Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010,** 53, 2719-40.

121.   Landrum, G. *RDKit: Open-Source Cheminformatics Software.* *http://www.rdkit.org*.

122.   Furka, Á.; SebestyÉN, F.; Asgedom, M.; DibÓ, G. General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Pept. Protein Res.* **1991,** 37, 487-493.

123.   Behjati, S.; Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **2013,** 98, 236-238.

124.   Szmant, H. H. *Organic Building Blocks of the Chemical Industry*. New York: John Wiley & Sons.: 1989; p 736.

125.   Zhou, J. Z. Chemoinformatics and Library Design. In *Chemical Library Design*, Zhou, J. Z., Ed. Humana Press: Totowa, NJ, 2011; pp 27-52.

126.   Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008,** 3, 1503-7.

127.   Schneider, G.; Hartenfeller M Fau - Reutlinger, M.; Reutlinger M Fau - Tanrikulu, Y.; Tanrikulu Y Fau - Proschak, E.; Proschak E Fau - Schneider, P.; Schneider, P. Voyages to the (un)known: adaptive design of bioactive compounds.

128.   Lachance, H.; Wetzel, S.; Kumar, K.; Waldmann, H. Charting, navigating, and populating natural product chemical space for drug discovery. *J. Med. Chem.* **2012,** 55, 5989-6001.

129.   Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* **2008,** 48, 68-74.

130.   Jayaseelan, K. V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* **2012,** 13, 106.

131.   Sobell, M. G. *A practical guide to Ubuntu Linux*. Pearson Education: 2015.

132.   *Apache2 Software Distribution (2021) Apache2 Documentation. License from https://www.apache.org/licenses/LICENSE-2.0*.

133.   *Anaconda Software Distribution. (2021). Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/*.

134.   *jQuery Software Distribution (2021). jQuery Documentation. Retrieved from https://jquery.com*.

135.   *Bootstrap Software Distribution (2021). Bootstrap Documentation. Retrieved from https://getbootstrap.com*.

136.   *LifeSciences unit of EPAM Systems. EPAM Documentation. Retrieved from https://lifescience.opensource.epam.com/ketcher/*.

137.   *JavaScript port of the OpenChemLib Java library. OpenChemLib-js . Retrieved from https://github.com/cheminfo/openchemlib-js*.

138.   *Inc., P. T. (2015). Collaborative data science. Montreal, QC: Plotly Technologies Inc. Retrieved from https://plot.ly/*.

139.   Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* **2005,** 4, 649-63.

140.   Sattarov, B.; Baskin, II; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J Chem Inf Model* **2019,** 59, 1182-1196.

Yuliana ZABOLOTNA

Exploration par chémographie des espaces chimiques ultra-larges pour la chimie médicinale

Résumé

Cette thèse est dédiée à l'analyse détaillée de l'espace chimique des chimiothèques ultra-larges à l'aide de l'approche GTM et au développement de ChemSpace Atlas – un outil en ligne conçu pour la navigation à travers des milliards de composés. L'efficacité et la polyfonctionnalité de la GTM ont permis de produire une image détaillée de l'espace chimique actuellement disponible pour les chimistes médicinaux. Plusieurs groupes de composés (fragment-, lead-, drug-, PPI- and NP-like, produits naturels, building blocks, et les bibliothèques codées par l'ADN) ont été systématiquement analysés à l'aide de la GTM hiérarchique. Les dizaines de milliers de cartes ainsi obtenues ont été utilisées comme base principale de l'atlas ChemSpace. Cet outil permet une exploration efficace de l'espace chimique ultra-large sous des angles différents : chimiotypes, diverses propriétés physicochimiques, activités biologiques, etc. En outre, la hiérarchie des cartes offre de multiples niveaux de détail : d'une vue globale de l'ensemble des données sur la carte universelle à la détection de motifs structurels dans des zones distinctes sur les cartes zoomées dédiées aux régions spécifiques.

Résumé en anglais

This thesis is dedicated to the detailed GTM-based analysis of the chemical space of ultra-large libraries and development of the online tool for navigation through up to billions of compounds, called ChemSpace Atlas. The efficiency and polyfunctionality of GTM allowed producing a detailed picture of the chemical space currently available to medicinal chemists. Fragment-, lead-, drug-, PPI- and NP-like compounds, genuine NPs, purchasable building blocks, and DNA-encoded libraries were systematically analyzed using hierarchical GTM. The resulting tens of thousands of maps were employed as the main basis of the ChemSpace Atlas. This tool enables efficient exploration of the ultra-large chemical space from different perspectives: chemotypes, various physicochemical properties, biological activities, etc. Moreover, the hierarchy of maps provides multiple levels of detalization: from a global bird's eye view of the whole dataset on the universal map to the structural pattern detection in separate areas of the region-dedicated zoomed maps.