Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

EDSC
École Doctorale des
Sciences Chimiques

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :

Arkadii LIN

soutenue le : **16 septembre 2019**

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

Discipline/ Spécialité : **Chimie / Chémoinformatique**

Cartographie Topographique Générative: un outil puissant pour la visualisation, l'analyse et la modélisation de données chimiques volumineuses

---

THÈSE dirigée par :

    M. VARNEK Alexandre   Professeur, Université de Strasbourg

RAPPORTEURS :

    M. BAJORATH Jürgen   Professeur, Université de Bonn

    M. MONTES Matthieu   Professeur, Conservatoire National des Arts et Métiers

---

AUTRES MEMBRES DU JURY :

    M. BECK Bernd   Docteur, Boehringer Ingelheim Pharma GmbH & Co. KG

    M. ROGNAN Didier   Directeur de Recherche au CNRS

Abstract

This thesis concerns the application of the Generative Topographic Mapping (GTM) approach to the analysis, visualization, and modeling of Big Data in chemistry. The main topics covered in this work are multi-target virtual screening in drug design and large chemical libraries visualization, analysis, and comparison. Several methodological developments were suggested: (*i*) an automatized hierarchical GTM zooming algorithm helping to resolve the map resolution problem; (*ii*) an automatized Maximum Common Substructure (MCS) extraction protocol improving efficiency of data analysis; (*iii*) constrained GTM-based screening allowing to detect molecules with a desired pharmacological profile, and (*iv*) a parallel GTM technique, which significantly increases the speed of GTM training. Developed methodologies were implemented in a software package used in both academic (University of Strasbourg, France) and industrial (Boehringer Ingelheim Pharma company, Germany) projects.

Acknowledgements

Contents

1 Résumé en français

1.1 Introduction

De nos jours, les bases de données chimiques telles que CAS, contiennent des millions de structures chimiques [1], et ce nombre augmente exponentiellement, grâce à l'utilisation de nouvelles technologies de synthèse combinatoire et parallèle, de réacteurs en flux continu ou de micro-ondes, entre autres. De plus, des milliards de structures virtuelles sont aisément énumérées par ordinateur (166 milliards de composés dans la base de données GDB-17 [2]). Ces chiffres restent toutefois modestes comparés au nombre de composés dans l'espace chimique d'intérêt thérapeutique, estimé à 10^{33} [3]. L'exploration de ces espaces chimiques est un défi pour les chimistes souhaitant comprendre leur structure, découvrir les régions inexplorées et analyser les relations structure-activité des molécules qu'ils contiennent.

Les cartes topographiques génératives (Generative Topographic Mapping - GTM) [4] permettent de modéliser, d'analyser et de visualiser de grandes bases de données. Leur contenu est projeté dans un espace bidimensionnel, qualifié d' « espace latent ». Cette méthode a été appliquée avec succès pour comparer des chimiothèques [5] et pour la modélisation de Relations Quantitatives Structure-Activité (QSAR) [6]. Néanmoins, des ajustements technologiques et méthodologiques sont nécessaires pour utiliser cette approche dans le cas des mégadonnées (ou « Big Data »).

Cette thèse est dédiée à l'amélioration de la GTM et à ses applications dans différents contextes de mégadonnées. Cette thèse consiste en 6 Chapitres. Le chapitre 1 est une introduction concernant la méthode GTM et ses applications décrites dans la littérature. Le chapitre 2 présente les améliorations méthodologiques proposées, telles que le zoom hiérarchique, le domaine d'applicabilité double ou encore l'extraction des structures maximales communes. Le Chapitre 3 rapporte les résultats de l'utilisation de la GTM pour établir le profil de composés sur de multiples cibles simultanément, c'est-à-dire pour un criblage virtuel multi-cibles (VS), et des études comparatives de la GTM avec des algorithmes d'apprentissage machine éprouvés. Le Chapitre 4 décrit les résultats de la comparaison de grandes bases de données publiques (PubChem-17 et ChEMBL-17) avec les composés virtuels énumérés dans la FDB-17 [7]. Le Chapitre 5 montre l'application de la GTM pour enrichir les collections de produits de la société Boehringer Ingelheim Pharma (BI) avec des composés originaux, en tenant compte de l'expérience apportée par les projets précédents. Le dernier chapitre (Chapitre 6) est consacré à l'implémentation d'algorithmes parallèles pour accélérer les calculs GTM et aborder de nouveaux problèmes dans le domaine des mégadonnées.

1.2   Résultats et discussions

1.2.1 Criblage virtuel de grandes collections chimiques

Les Relations Quantitatives Structure-Activité (QSAR) sont un domaine clé de la chémoinformatique. Ces modèles visent à sélectionner rationnellement les composés par rapport à une activité biologique ou une propriété. Etant donné que la GTM peut être utilisée pour créer des modèles QSAR, le premier défi était de l'appliquer à du criblage virtuel (VS) sur une cible (mono-cible) puis sur plusieurs cibles simultanément (multi-cible). Ces techniques ont été appliquées à une grande collection de problèmes de classification appelée DUD (Directory of Useful Decoys) [8]. A cette fin, les GTM *universelles* décrites par P. Sidorov et al. [9] ont été utilisées. Ces cartes sont entrainées pour modéliser une grande base de données (ChEMBL v23 dans cette étude) et ont été

choisies pour leur capacité à prédire plusieurs centaines de propriétés biologiques. La méthode a aussi été comparée à d'autres approches d'apprentissage machine éprouvées : la recherche par similarité (avec et sans fusion de données), des réseaux de neurones, et une forêt aléatoire. Pour mesurer la performance d'une méthode, la moyenne des aires sous la courbe ROC (Receiver Operating Characteristic), $<AUC>_{1/2}$, a été utilisée. Les résultats de la validation sur les centaines de cibles utilisées pour choisir les cartes sont présentés en Figure 1.
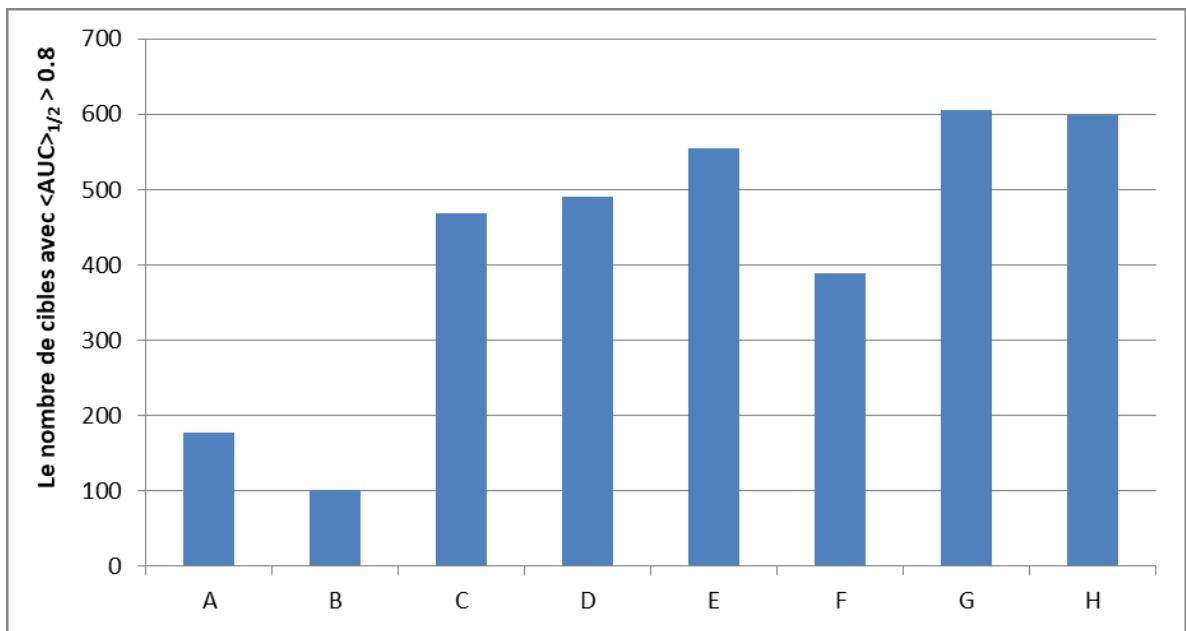
Figure 1. Le nombre de cibles pour lesquelles le meilleur modèle sur les quatre espaces de descripteurs retourne $<AUC>_{1/2} > 0.8$. A – Recherche par similarité dans l'espace initial, B – Recherche par similarité dans l'espace des responsabilités (description des données par la GTM), C – GTM universelle, D – GTM mono-cible, E – Recherche par similarité avec fusion de données dans l'espace initial, F – Recherche par similarité avec fusion de données dans l'espace des responsabilités, G – Réseau de neurones, H – Forêt aléatoire.

La validation effectuée sur les 9 cibles de la DUD en utilisant des données jamais utilisées pour entraîner ou sélectionner les cartes, a montré des performances similaires (Figure 2). Les résultats de cette étude ont été publiées [10].
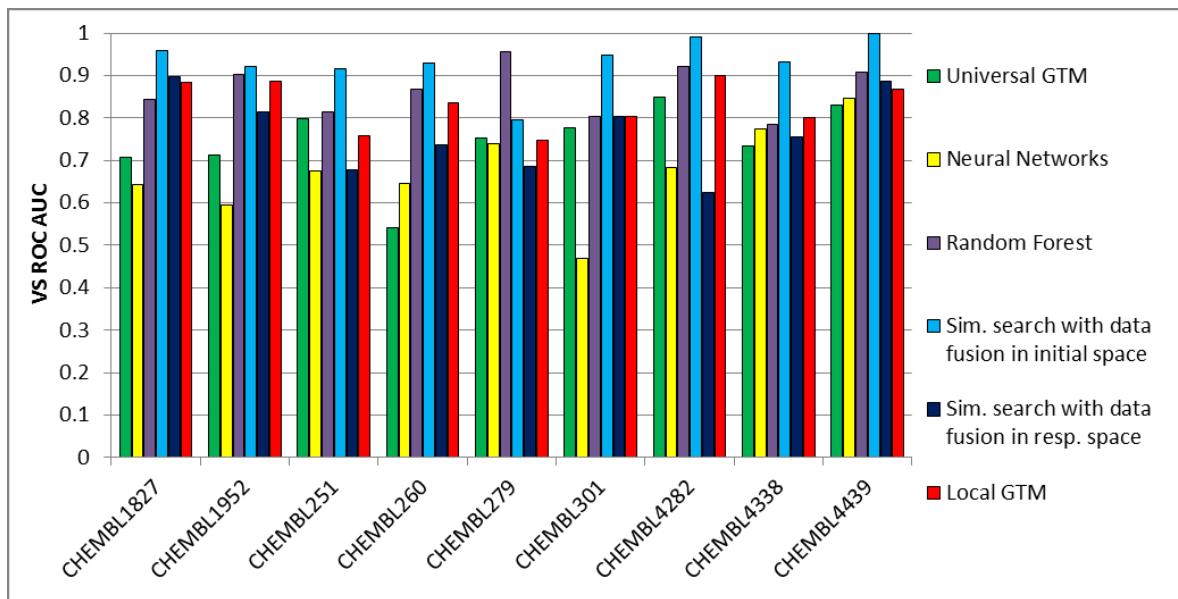
Figure 2. Comparaison des méthodes de criblage virtuel. Les GTM ont été entraînées et validées sur ChEMBL v23. Les cartes utilisées sont celles qui ont montré les meilleures performances en termes de ROC AUC, obtenues en validation croisée.

Ensuite, l'approche de la GTM universelle a été testée dans l'environnement industriel de Boehringer Ingelheim. Tout d'abord, des GTM ont été entrainées sur 25K structures chimiques représentatives des collections internes de l'entreprise (le « *frame set* »). Les descripteurs moléculaires et les paramètres de la méthode GTM les plus pertinents ont été déterminés en échantillonnant systématiquement leurs valeurs sur une grille (le nombre de nœuds est 20*20 ÷ 50*50 avec un pas de 5, le nombre de RBF est 40 ÷ 70% du nombre de nœuds avec un pas de 10, le coefficient de régularisation est 1.0 ÷ 5.0 avec un pas de 0.5, et la largeur des RBF est 1.0 ÷ 5.0 avec un pas de 0.5).

Plus de 230K combinaisons de paramètres ont été essayées, et les 5 meilleures cartes ont été sélectionnées (Table 7 ; chapitre 5.2.4).

Ces cartes ont été validées par validation croisée en 3 paquets sur 2371 problèmes de classification concernant l'activité de composés sur des cibles biologiques. Pour mesurer la performance d'une carte, la moyenne des aires sous la courbe ROC ($<AUC>^{3cls}$ pour les problèmes à 3 classes et $<AUC>^{bin}$ pour les problèmes à 2 classes) a été utilisée (Table 8).

La validation croisée montre que ces cartes sont prédictives dans plus de 50% des tests proposés (1318 tests), avec une $<AUC>^{3cls} \geq 0.7$. Ces cartes ont été utilisées pour prédire l'activité sur 42 nouvelles cibles biologiques. Pour 4 d'entre elles, la précision balancée (Balanced Accuracy, BA) était supérieure à 0.7.

1.2.2 Comparaison de bases de données chimiques publiques

Une base de données couvrant l'espace chimique de composés contenant au plus 17 atomes lourds a été publiée par J.-L. Reymond et al. [2] (GDB-17). Des molécules contenant également au plus 17 atomes lourds ont été échantillonnés dans les bases de données ChEMBL (ChEMBL-17) et PubChem (PubChem-17) pour être comparées à un échantillon de 10M de composés de la GDB-17, la FDB-17 [7]. L'objectif était d'identifier les chémotypes particuliers appartenant à l'une ou à l'autre base en exclusivité. Comme la FDB-17 contient des structures chimiques virtuelles énumérées par un algorithme, la comparaison avec de véritables composés chimiques (ChEMBL-17, PubChem-17) pourrait donner lieu à la découverte de nouveaux chémotypes, qui n'ont encore jamais été synthétisés. Une GTM a donc été entrainée sur un *frame set* de 100K structures, sélectionnées au hasard mais avec un ratio égal pour chacun des 3 jeux de données. Puis, les données (21.1M de composés) ont été projetées sur cette carte. Les cartes ont été annotées en fonction de la prévalence d'une base par rapport à une autre dans une région de l'espace chimique représentée par la carte. Ces cartes annotées sont appelées *paysages*, dans la suite.

Les jeux de données ont été comparés en utilisant *(i)* des métriques de dissimilarité (le coefficient de Bhattacharyya, les distances Euclidienne et de Soergel), *(ii)* des paysages comparant FDB-17 avec PubChem-17/ChEMBL-17, et *(iii)* des propriétés moléculaires (nombre d'atomes lourds, chiralité, LogP, nombre d'atomes aromatiques, etc.) Les résultats de l'étude ont été publiés [10]. Pour résumer, la comparaison a montré que les bases de données PubChem-17 et ChEMBL-17 sont très similaires, ce qui est expliqué par le fait que la première inclut la seconde (Figure 3).
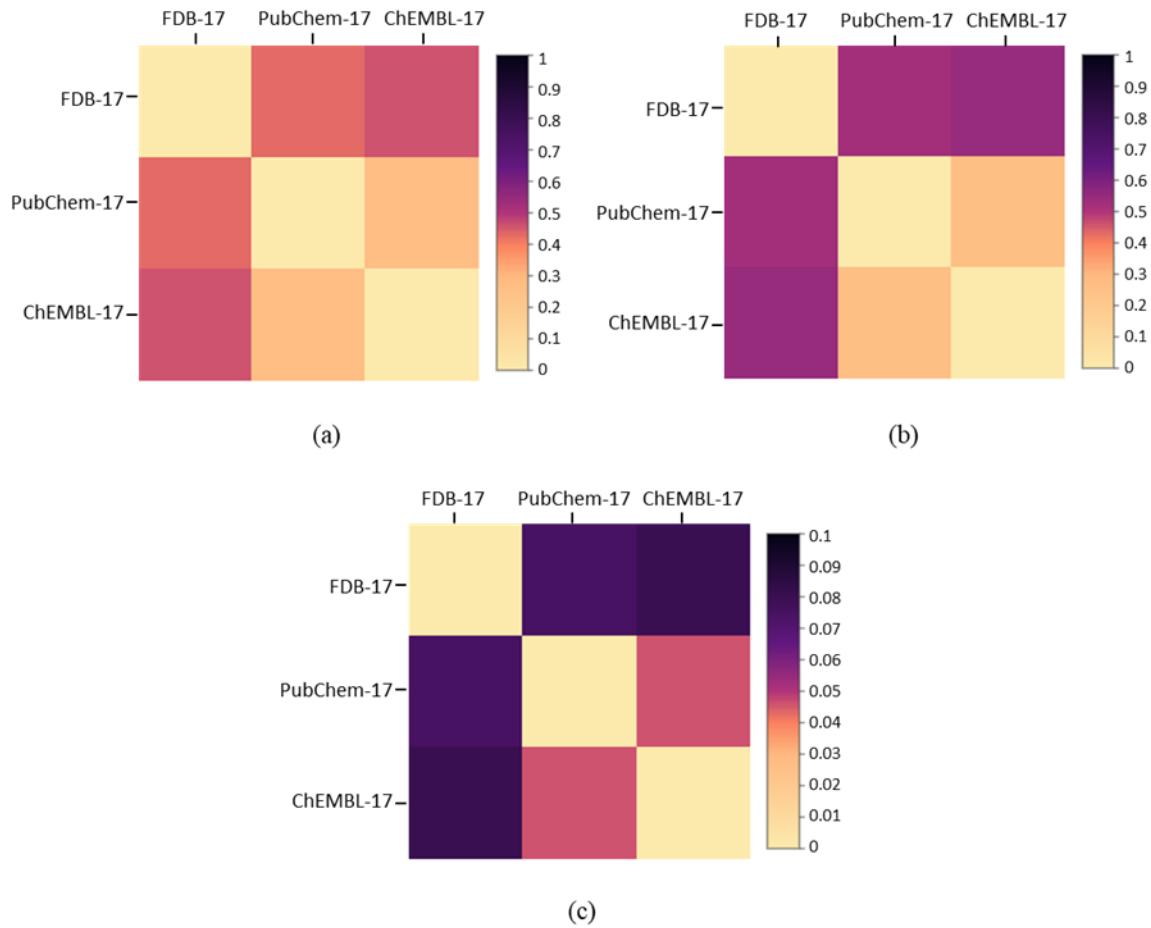
Figure 3. Diagramme de chaleur représentant les similarités entre trois chimiothèques sur la base de GTM. Les métriques utilisées sont (a) le coefficient de Bhattacharyya (1-$S_{Bhattacharyya}$), (b) le coefficient de Tanimoto (1-$S_{Tanimoto}$) et (c) la distance Euclidienne.

Par contraste, la PubChem-17 diffère significativement de la FDB-17. Le paysage résultant, illustré par la Figure 4, montre que la PubChem-17 est dominante dans plusieurs zones de la carte dans lesquelles les composés avec des groupes nitro attachés à un système aromatique et/ou des groupes carboxyl sont localisés (zones rouges). L'absence de ces structures dans la FDB-17 est expliquée par les règles que les auteurs de la base de données ont appliquées au cours de l'énumération des structures pour restreindre l'espace chimique virtuel à des composés qu'ils ont jugés intéressants pour des applications pharmaceutiques [7].

16
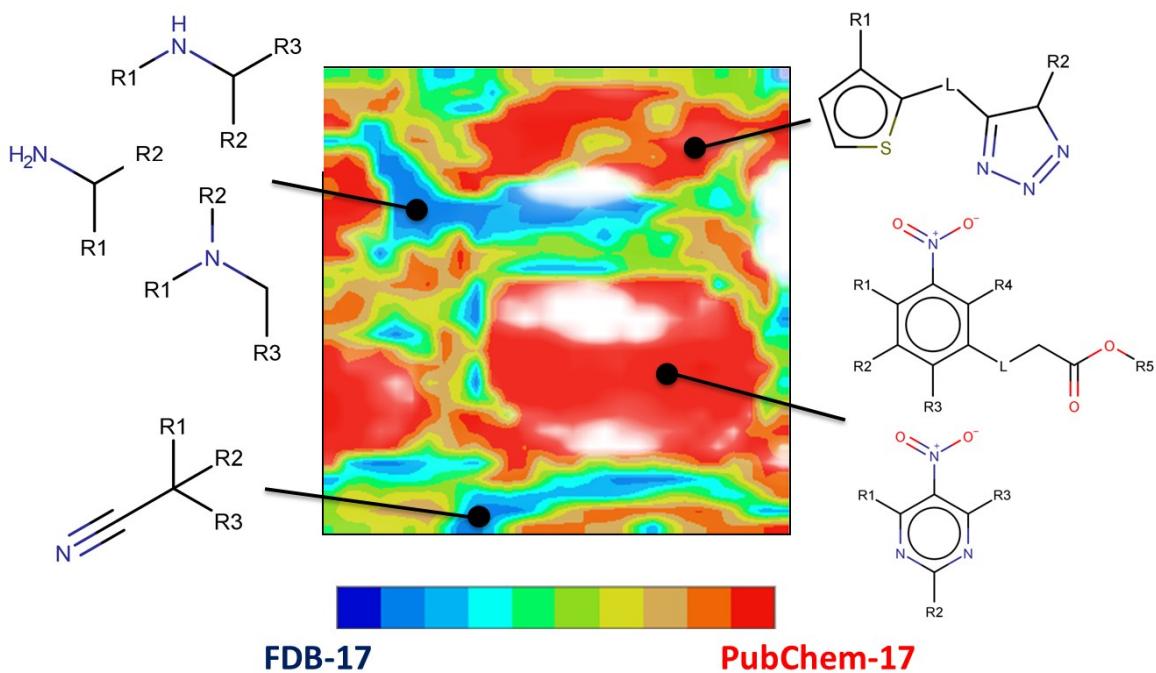
Figure 4. Paysage comparant les bases de données FDB-17 et PubChem-17.

Au cours de ce travail, un écueil était que les cartes représentaient un si grand nombre de composés que chaque élément en couvrait des centaines de milliers, ce qui en compliquait l'analyse. Pour résoudre ce problème et analyser plus finement les composés dans les zones de l'espace chimique où la FDB-17 se recouvre avec la PubChem-17 (zones vertes et jaunes), une technique appelée zoom hiérarchique de GTM (proposée auparavant par Nabney et al. [11]) a été appliquée. Elle consiste à extraire les composés d'une région de l'espace chimique représentée par une zone délimitée sur la carte et d'entrainer une nouvelle GTM en utilisant les mêmes paramètres que ceux de la carte principale (Figure 5). Cette technique a permis d'identifier de nouveaux châssis moléculaires absents de la base de données PubChem. Les structures contenant ces châssis et présentées en Figure 5 ont été extraites de la collection FDB-17. Aucune molécule similaire n'est présente dans la base de données PubChem.
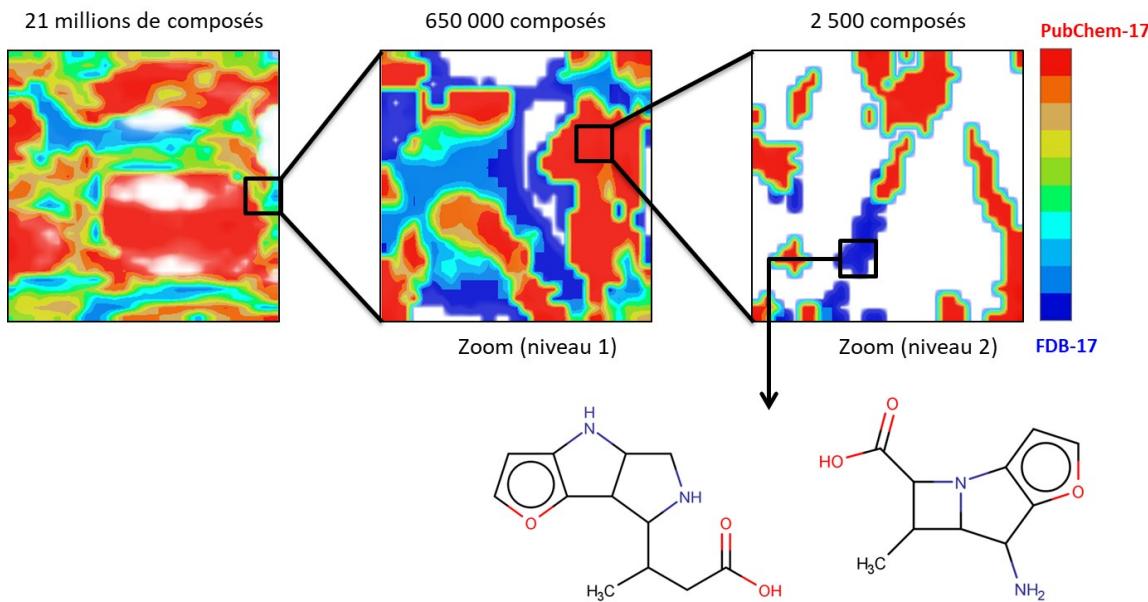
Figure 5. Zoom hiérarchique de GTM sur l'espace chimique occupé par la FDB-17 (en bleu) et la PubChem-17 (en rouge). Pour une zone délimitée sur une carte, un modèle local de GTM est reconstruit en utilisant uniquement sur les molécules y résidant. Sous la carte zoomée sont montrés des exemples de composés extraits d'une zone peuplée exclusivement par des composés de la FDB-17 sur une carte zoomée. Ces composés n'ont pas d'analogues dans la base de données PubChem.

Pour finir, les bases de données ont été comparées en termes de 6 propriétés calculées sur les structures chimiques à l'aide du logiciel MOE : l'entropie de la distribution des éléments composant la molécule (a_ICM), le nombre d'atomes lourds (a_heavy), la chiralité (chiral), la lipophilicité (LogP), le nombre d'atomes aromatiques (a_aro), et le statut de quasi-fragment ASTEX (ast_fraglike_ext) [12]. Les résultats sont représentés sur la Figure 6. Les paysages de propriétés correspondants au nombre d'atomes lourds dans les molécules de ChEMBL-17 et de PubChem-17 (Figure 6) sont similaires. Toutefois, PubChem-17 contient un excès d'entrées de plus haut poids moléculaire (en rouge sombre). Ceci résulte de deux biais de composition des bases de données : d'une part, PubChem est composé de structures chimiques sélectionnées pour être à priori bio-actives puisqu'elles sont soumises à des bancs de tests biologiques. Les très petits composés ne pouvant pas former de complexes très stables avec des protéines (et en dépit de leur éventuelle efficacité en tant que ligand) sont rares dans PubChem.
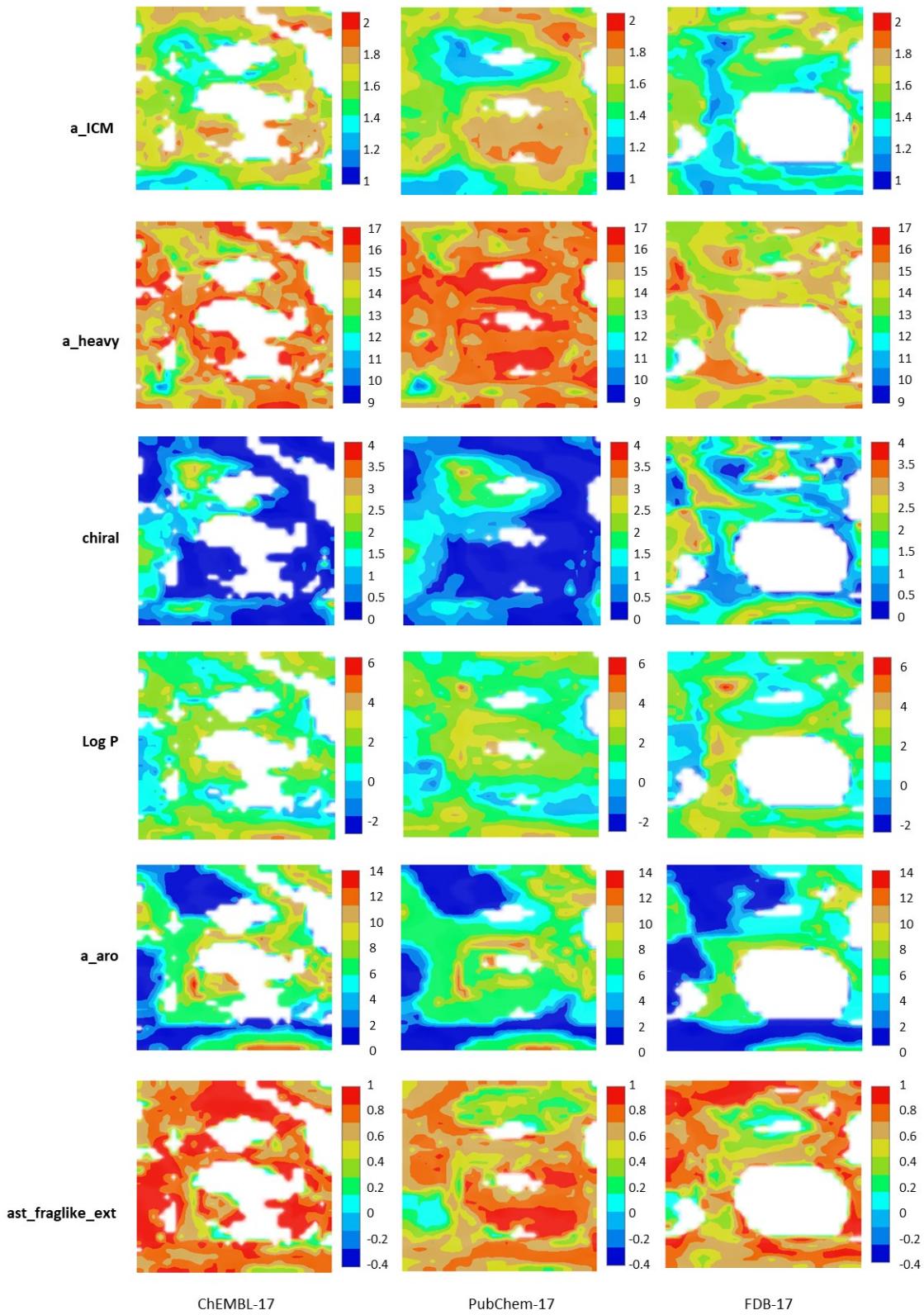
Figure 6. Paysages de propriété pour a_ICM (entropie de la distribution des éléments de la molécule), a_heavy (nombre d'atomes lourd), chiral (chiralité), LogP (lipophilicité), a_aro (nombre d'atomes aromatiques), et ast_fraglike_ext (Satut de quasi-fragment ASTEX) [12].

D'autre part, on peut remarquer que l'échantillon de la FDB-17 a été spécifiquement conçu pour équilibrer le nombre d'entrées correspondant à des molécules de tailles différentes. Les composés ayant un nombre d'atomes lourds intermédiaire ont été volontairement sur-échantillonnés. Autrement, pour des raisons évidentes de combinatoire, l'énumération systématique des composés ayant au plus 17 atomes lourds est dominé par les structures contenant exactement 17 atomes lourds.

Le paysage de l'entropie de la distribution des éléments (indice a_ICM de MOE) dans les molécules est similaire pour les jeux de données ChEMBL-17 and PubChem-17, alors que FDB-17 contient des structures moins diverses, au sens qu'il y a un biais de composition en faveur des chaines hydrocarbures en comparaison de fonctions chimiques plus élaborées. Des règles élémentaires de stabilité chimique empêchent la concaténer des hétéroatomes dans les structures de la base de données GDB-17, ce qui explique que les chaînes carbonées soient prédominantes. Mais, les chimiothèque de molécules effectivement synthétisées incluent des groupes fonctionnels chimiques élaborés qui apportent de la réactivité et des propriétés physico-chimiques intéressantes. Ces biais sont bien mis en évidence sur les cartes.

1.2.3 Enrichissement de librairie structurale pour Boehringer Ingelheim

En prenant en compte l'expérience apportée par les projets précédents, la GTM a démontré une bonne efficacité en criblage virtuel et pour la comparaison de chimiothèques. Dans cette étude, cette technique a été utilisée pour augmenter la diversité chimique de la collection interne de composés de Boehringer Ingelheim (BI). Pour ce faire, une carte GTM a été utilisée pour comparer cette collection BI au catalogue de l'entreprise Aldrich-Market Select (AMS) référençant plus de 8M de produits. Pour entraîner la carte, un jeu de données représentatif de 25,000 structures de diversité chimique contrôlée (ne présentant pas plus de deux structures chimiques plus similaires qu'une valeur seuil) a été constitué à partir de la base de données AMS. Pour commencer, un paysage de classification a été construit pour comparer les distributions des composés dans chaque chimiothèque (Figure 7).
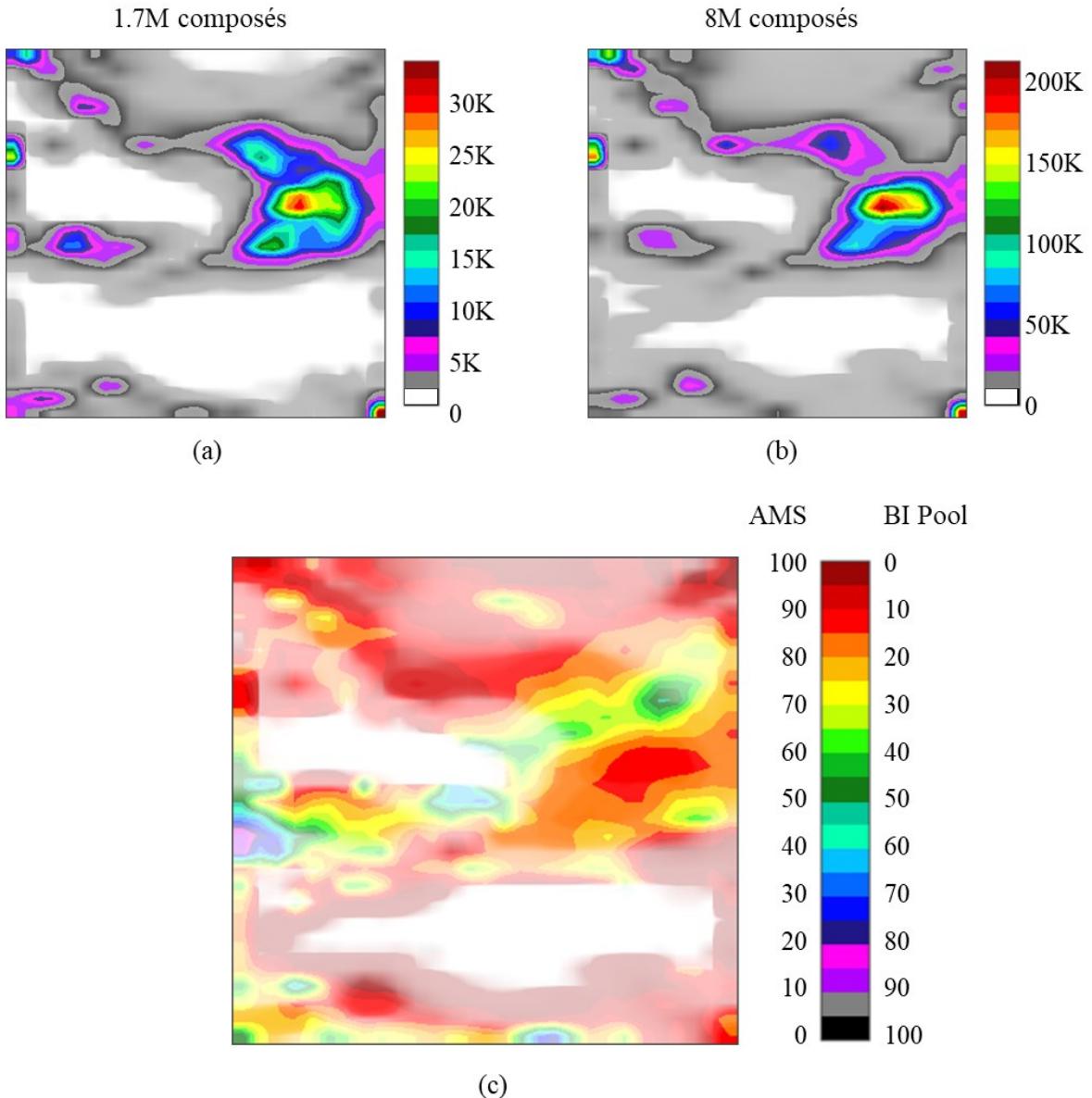
Figure 7. Comparaison des bases de données BI Pool vs AMS: (a) paysage de densité BI Pool, (b) paysage de densité AMS, et (c) paysage de prépondérance AMS contre BI Pool. Les régions blanches sont non peuplées, et la transparence est proportionnelle à la densité de population.

Afin de découvrir de nouveaux châssis moléculaires, l'approche du zoom hiérarchique de GTM a été automatisée pour être appliquée systématiquement sur les zones de la carte dans lesquelles les composés AMS étaient le plus surreprésentés. Les collections ainsi identifiées ont été analysées pour en extraire les sous-structures maximales communes (Figure 8).
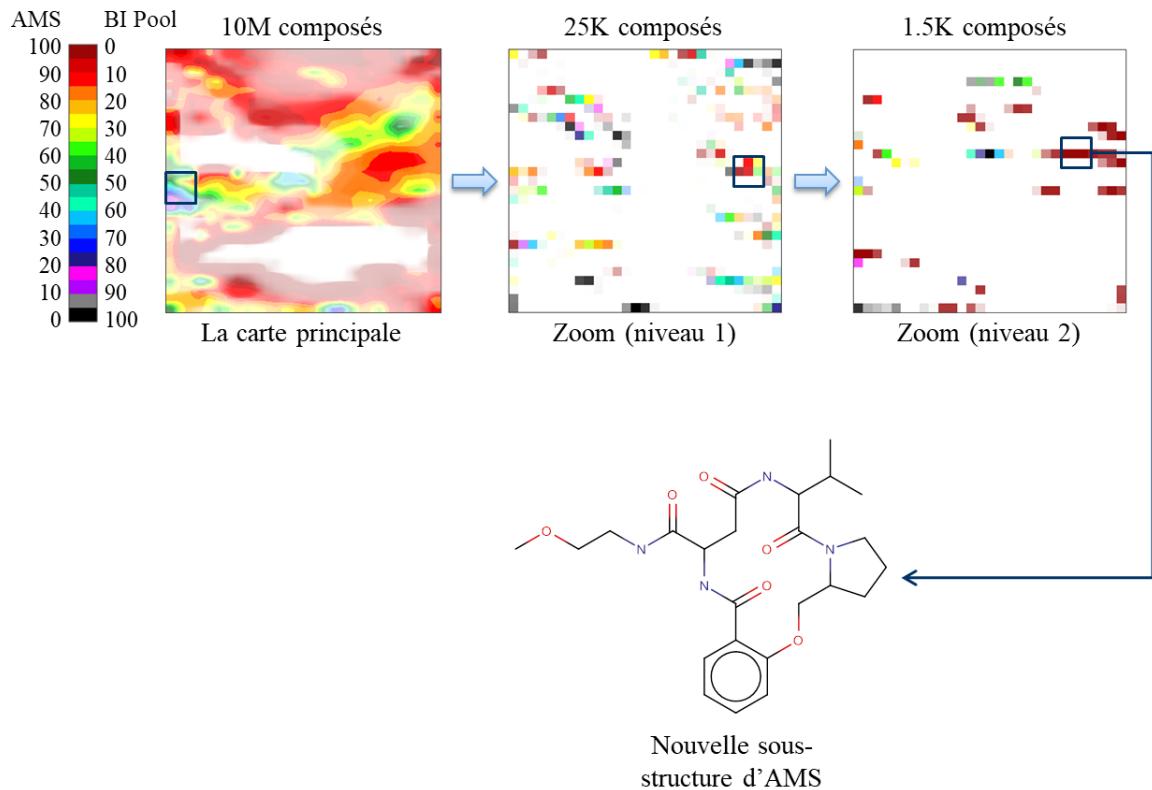
Figure 8. Un exemple d'analyse de zoom hiérarchique de GTM. Ici, une nouvelle sous-structure de la collection Aldrich-Market Select (AMS) a été découverte en utilisant un zoom à 2 niveaux. L'espace blanc indique des zones non peuplées, et la transparence correspond à la densité de la population.

De la sorte, un total de 45.5K nouvelles sous-structures ont été extraites de la base de données AMS ce qui a permis d'identifier 401K composés dans ce catalogue. La plupart de ces composés sont conformes aux règles de Lipinski et peuvent donc être considérés comme biodisponibles par voie orale (Figure 9). De plus, des GTM universelles entrainées sur la version 24 de la base de données ChEMBL ont été appliquées pour estimer le profil biologique de ces structures pour 749 cibles. Plus de 1.2K composés ont été identifiés pour avoir une activité potentielle sur différentes cibles avec une probabilité supérieure à 80%.
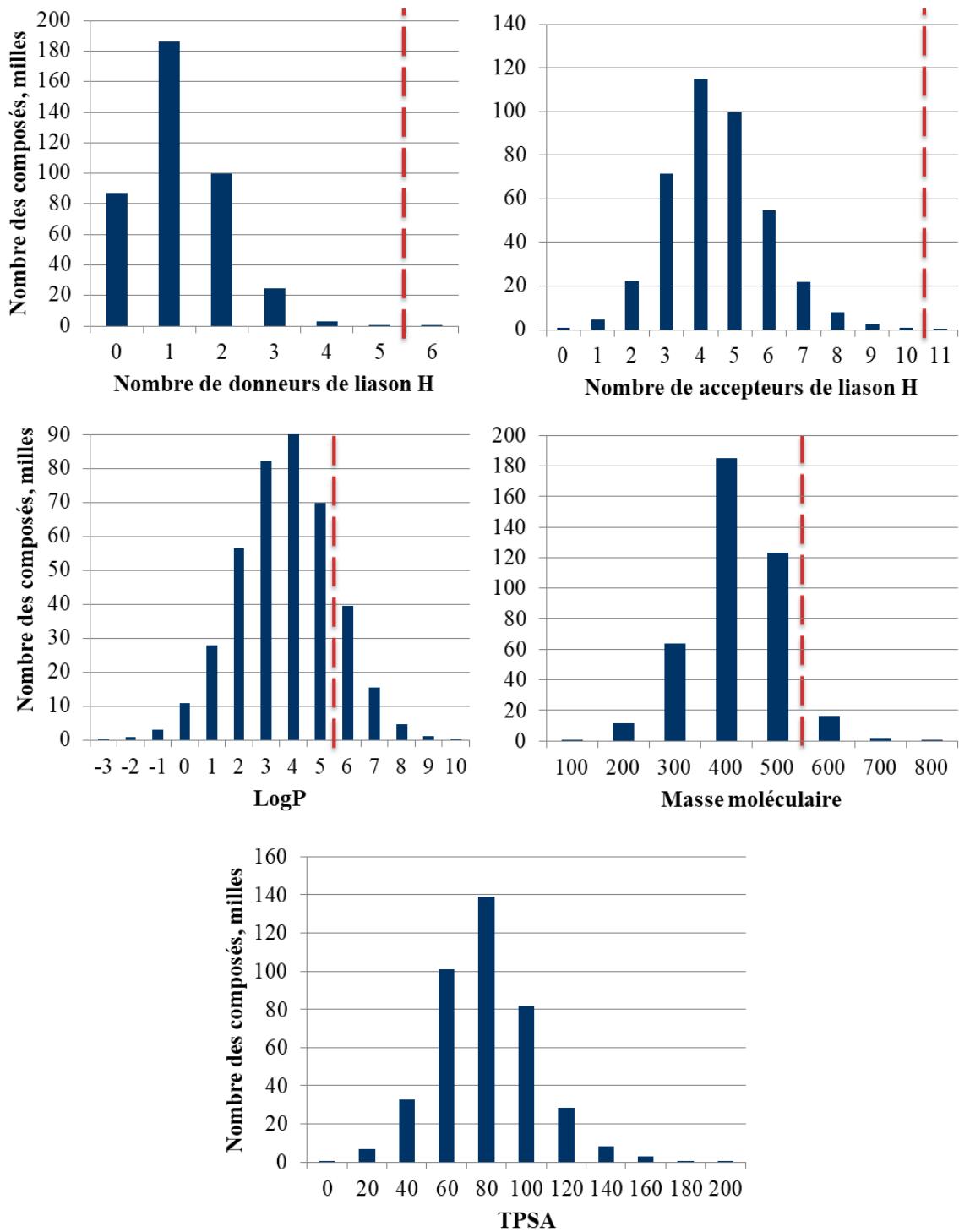
Figure 9. Histogrammes représentant le nombre de donneurs et d'accepteurs de liaison hydrogène, de lipophilicité (LogP), de poids moléculaires, et de surface polaire topologique (TPSA) calculés pour l'extrait de 401K composés de la base de données AMS. Les lignes pointillées rouges matérialisent les règles de Lipinski [13].

Des exemples de ces touches virtuelles sont montrés en Figure 10.
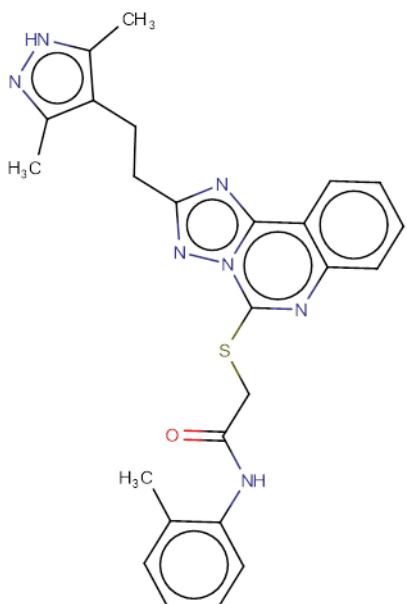
ID de structure AMS 41419963
Cible: Photoreceptor-specific
nuclear receptor
La probabilité d'être actif est de 93%
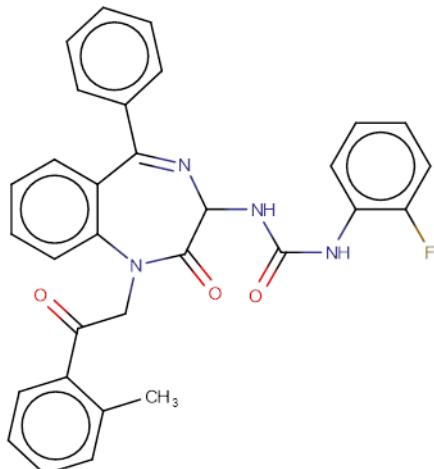
ID de structure AMS 414778192
Cible: Cholecystokinin B receptor
La probabilité d'être actif est de 92%
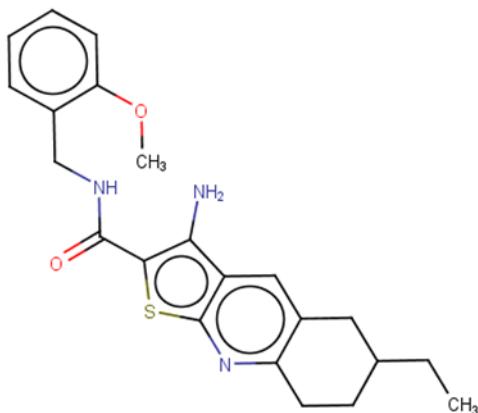
ID de structure AMS 29149085
Cible: Muscarinic acetylcholine
receptor M4
La probabilité d'être actif est de 91%
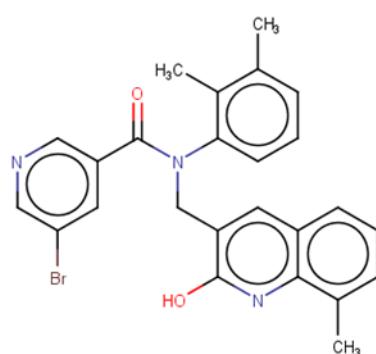
ID de structure AMS 48316039
Cible: Pyruvate dehydrogenase kinase
isoform 1
La probabilité d'être actif est de 90%

Figure 10. Exemples de structures prédites actives et identifiées dans l'extrait de 401K de la base de données AMS.

Les structures découvertes ont été recommandées à l'entreprise afin d'être achetées pour alimenter leurs collections. Le papier rapportant les résultats de cette étude a été accepté à la publication « *Journal of Computer-Aided Molecular Design* ».

1.2.4 GTM parallèle

Les avantages de la GTM ont été montrés dans différentes applications dans le contexte des mégadonnées. Cependant, il reste encore quelques limitations techniques et méthodologiques qui en restreignent l'usage à des quantités de données plus grandes que quelques dizaines de millions de molécules. Pour surmonter ces limites, le concept de GTM parallèle a été proposé. Le concept général est décrit par la Figure 11.
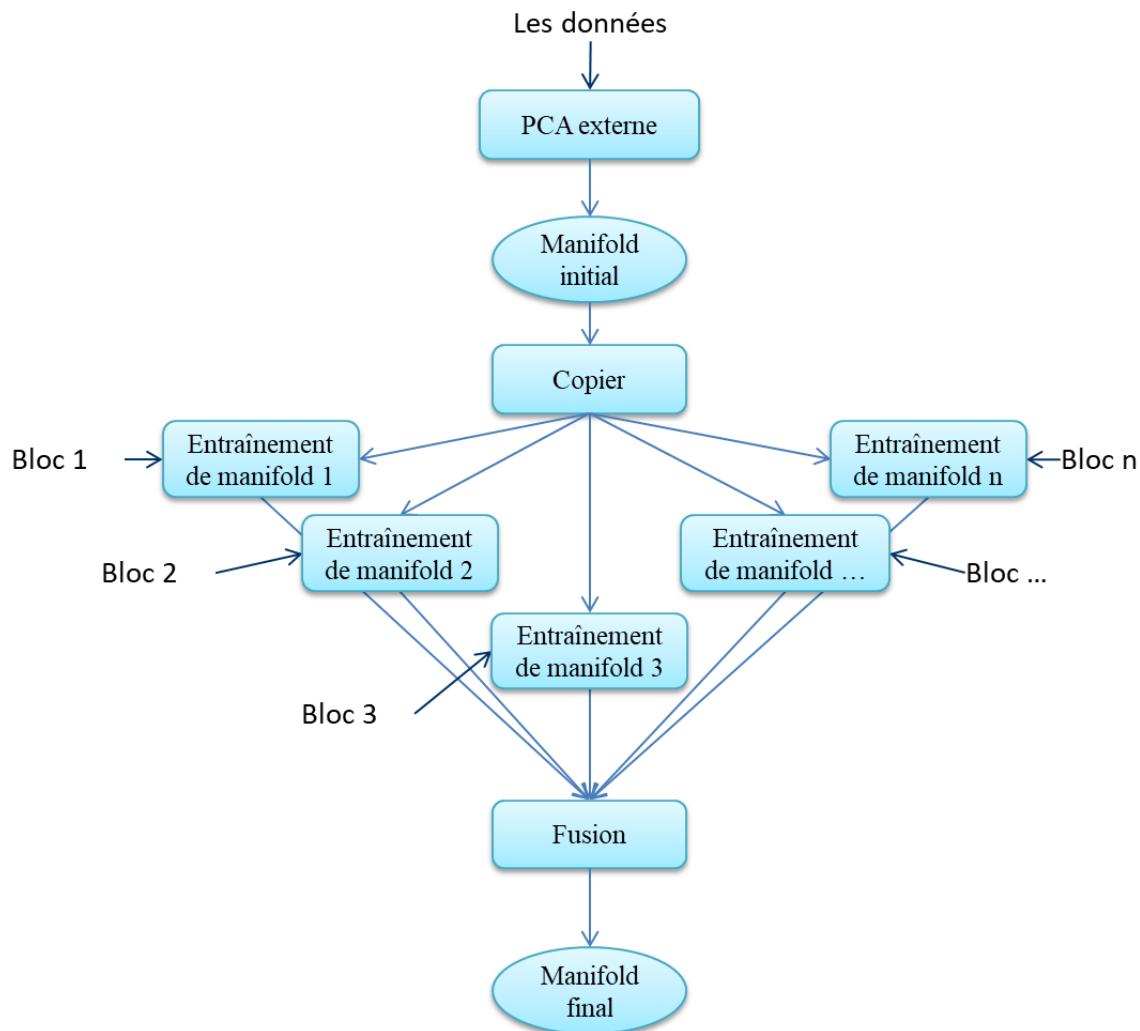
Figure 11. Représentation schématique de l'algorithme GTM Parallèles.

Il consiste à entrainer des GTMs sur différentes parties du jeu de données en parallèle. Une fois que les nappes intermédiaires ont été ajustées à leurs données respectives, elles sont fusionnées dans une nappe unique. A cette fin, trois stratégies sont envisagées: 1) moyenner les matrices de paramètres décrivant chaque nappe et moyenner la largeur de la distribution gaussienne, 2) faire des moyennes pondérées par la vraisemblance issue de chaque nappe, et 3) faire des moyennes au travers d'une GTM. Celle-ci consiste à entraîner une nouvelle GTM à partir d'un jeu de données artificiel composé par les nœuds des GTM intermédiaire dans l'espace initial.

Cette approche a été testée en utilisant un jeu de composés extraits de la base de données ChEMBL (v24), pour lesquels les valeurs d'IC50 sur la prothrombine (CHEMBL204) étaient connues. La GTM parallèle a aussi été comparée à l'algorithme classique et incrémental de la GTM (telle que décrit par C. Bishop et al. [4]). La qualité des modèles obtenus a été mesurée sur leur capacité prédictive concernant l'activité biologique sur la prothrombine et le temps d'exécution. Les résultats de cette étude comparative ont montré que la GTM parallèle produit des modèles aussi prédictifs (les précisions balancées sont similaires avec une déviation d'environ 0.02) mais que les temps de calculs sont divisés par un facteur 2. En comparaison, les GTM incrémentales et parallèles utilisent des jeux de données bien plus gros (plus de 100,000 composés) et bénéficient d'une réduction des temps de calcul d'un facteur pouvant aller jusque 6.

1.3   Conclusions

1)   La méthode GTM (Generative Topographic Mapping) a été testée pour le criblage virtuel (VS) mono-cible et multi-cible. Les études comparatives ont montré que les modèles GTM ont des performances similaires aux autres méthodes d'apprentissage machine. Mais elle possède plusieurs avantages comme la possibilité de visualiser l'espace chimique.

2)   La méthode GTM a été testée avec succès pour comparer de grandes bases de données de composés réels et virtuels (PubChem-17, ChEMBL-17, FDB-17). Il a été

montré que la GTM permet de visualiser facilement des millions de points de données et de localiser les zones de l'espace chimiques où ces ensembles de molécules se recouvrent.

3)   La technique de zoom hiérarchique de GTM a été proposée comme une solution pour analyser plus finement le contenu des zones de l'espace chimique les plus peuplées. Elle augmente la capacité de la GTM à distinguer différents chémotypes. Ceci donne lieu à une extraction plus efficace de châssis et de sous-structures maximales communes.

4)   Un nouveau protocole d'extraction de sous-structures maximales communes a été proposé. Ce protocole a été intégré à la technique de zoom hiérarchique de GTM. L'outil développé a été utilisé avec succès pour enrichir la collection interne de la société Boehringer Ingelheim Pharma (45.5K nouvelles sous-structures, 401K molécules analysées et une liste de composés recommandés pour être achetés ou synthétisés par la société).

5)   Le concept de GTM parallèle a été proposé. Il a été testé sur un jeu de données extrait de la base de données ChEMBL. Il a été montré que la GTM parallèle propose à l'utilisateur des modèles dont les performances sont conservées tout en divisant par 2 les temps de calcul.

1.4   Liste des presentations

1) <u>Lin Arkadii</u>, Alexandre Varnek : BigChem first Autumn School "Introduction to Chemoinformatics", Munich, Allemagne (17-21 Octobre 2016). **Oral.**

2) <u>Lin Arkadii</u>, Alexandre Varnek : Second School "Chemical databases", Barcelona, Espagne (19-21 Avril 2017). **Oral.**

3) Arkadii Lin, Dragos Horvath, <u>Fanny Bonachera</u>, Valentina Afonina, Gilles Marcou, Alexandre Varnek : 8es journées de la Société Française de Chémoinformatique, Orleans, France (12-13 Octobre 2017). **Poster.**

4) <u>Lin Arkadii</u>, Alexandre Varnek : Third School "Computer-Aided Drug Discovery", Modena, Italie (25-27 Octobre 2017). **Oral.**

5) <u>Lin Arkadii</u>, Bernd Beck, Alexandre Varnek : 32nd Molecular Modelling Workshop, Erlangen, Allemagne (12-14 Mars 2018). **Oral.**

6) <u>Lin Arkadii</u>, Bernd Beck, Alexandre Varnek : Third International School-Seminar "From Empirical to Predictive Chemistry", Kazan, Russie (5-7 Avril 2018). **Oral.**

7) <u>Lin Arkadii</u>, Dragos Horvath, Gilles Marcou, Alexandre Varnek, Bernd Beck : Chemoinformatics Strasbourg Summer School 2018, Strasbourg, France (25-29 Juin 2018). **Oral et Poster.**

8) <u>Lin Arkadii</u>, Alexandre Varnek : Fifth BIGCHEM School in Mölndal, Mölndal, Suède (6-10 Mai 2019). **Oral.**

1.5   Liste des publications

1) Arkadii Lin, Dragos Horvath, Valentina Afonina, Gilles Marcou, Jean-Louis Reymond, Alexandre Varnek: Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. ChemMedChem 13(6), 540-554 (2018). DOI: https://doi.org/10.1002/cmdc.201700561

2) Arkadii Lin, Dragos Horvath, Gilles Marcou, Bernd Beck, Alexandre Varnek : Multi-task generative topographic mapping in virtual screening. Journal of Computer-Aided Molecular Design 33(3), 331-343 (2019). DOI : https://doi.org/10.1007/s10822-019-00188-x

3) Arkadii Lin, Bernd Beck, Dragos Horvath, Gilles Marcou, Alexandre Varnek : Diversifying Chemical Libraries with Generative Topographic Mapping. Journal of Computer-Aided Molecular Design (2019), acceptée.

2 Introduction

The number of synthesized chemical structures increases exponentially because of the implementation of parallel and combinatorial synthesis approaches, as well as new experimental techniques like flow or microwave reactors. CAS Registry is the largest chemical database of registered compounds that have been synthesized since the 1800s, and it already contains 154 million organic and inorganic substances [14]. Yet, it covers just a part of chemical space. Thus, Reymond et al. [2] virtually enumerated a new database (*GDB-17*) of 166 billion small molecules containing no more than 17 heavy atoms. According to the estimation made by P. Polishchuk et al. [3], the drug-like chemical space includes at least 10^{33} molecules. These studies demonstrated that modern chemistry enters the era of Big Data.

Among various definitions of "Big Data", the most pertinent, to our opinion, belongs to A. De Mauro et al. [15] who defined this as "the information asset characterized by such high Volume, Velocity, and Variety to require specific technology and analytical methods for its transformation into value". Lusher et al. [16] included in this description "Veracity" and "Value" criteria thus completing the 5 "V's" definition. Specifically for chemical data, Bajorath et al. [17] suggested also to use the Complexity and Heterogeneity criteria.

The value of Big Data in chemistry is determined by the knowledge which can be extracted via large chemical databases analysis and modeling. In this context, data visualization and analysis plays an important role in modern chemistry and, especially, in

drug-discovery. This helps a chemist to decide by combining human and artificial intelligence.

Nowadays, three groups of methods are used for chemical data analysis, visualization and modeling: (*i*) graph-based, (*ii*) descriptors-based, and (*iii*) combined methods. The graph-based approaches represent a molecule as a graph where the nodes represent atoms and the edges play a role of chemical bonds. A general way to analyze graph-based chemical space stands on the concept of a molecular framework (scaffold) defined as the part of a structure which remains after all terminal chains have been removed [18]. Scaffolds can be used to group structures in a hierarchical scaffold tree which allows to visualize data and even to model structure-activity relationship (*SAR*) [19]. Maximum Common Substructure (*MCS*) – based algorithms are used in chemoinformatics to extract the largest connected or disconnected subgraph shared by a pair or a group of structures. Its application can be also found in data clustering and SAR studies [20]. Matched Molecular Pairs (MMP) method [21] represents another popular way for SAR analysis.

In contrast to the graph-based methods, the descriptors-based approaches consider a molecule as a vector of numbers (descriptors) that describe a compound in terms of structural and/or physical or chemical properties (e.g., structural fragments, molecular weight, LogP, etc.). These descriptors vectors are used as input in various machine-learning approaches, among which the dimensionality reduction techniques reside a huge variety of multi-dimensional data visualization and modeling. Nowadays, dozens of dimensionality reduction methods are reported in the literature [22]: Multi-Dimensional Scaling (*MDS*) [23], Sammon mapping [24], Principal Component Analysis (*PCA*) [25–27], Self-Organizing Maps (*SOM*) [28], Laplacian Eigenmaps [29], Canonical Correlation Analysis [30], Independent Component Analysis [31], Exploratory Factor Analysis [32], Isomaps [33], Locally Linear Embedding [34], Auto-encoder based dimensionality reduction [35], etc. These methods became popular due to their efficiency and capabilities. For instance, SOM is providing the user with a nice 2D map which is based on a non-linear model, whereas PCA is able to represent the data in 2D or 3D PC space. However, these popular

methods have some clear drawbacks. Thus, PCA can efficiently be applied to process huge datasets with linearly dependent features, but it is less effective with nonlinear data distributions [36]. As a consequence, this approach fails to represent the cluster structure of vast multidimensional data [37]. MDS is also a linear technique, which for the case of Euclidean distances gives equivalent results to PCA [38]. Sammon maps have no explicit mapping function and, therefore, do not allow one to place any new data on an already existing map. In that case, a new map must be rebuilt from scratch [39]. Besides, calculation and storage of all inter-point distances are required; this imposes severe restrictions on many practical applications dealing with large amounts of data or incremental data flow. The SOM approach has no well-defined objective function to be optimized during the training procedure [40, 41] and, therefore, no theoretical framework to prove its convergence and to select the method's parameters can be defined. This leads to some ambiguity in the selection of the "best" SOMs.

In an attempt to overcome the drawbacks mentioned above, a probabilistic extension of SOM named Generative Topographic Mapping (*GTM*) [4] was proposed. Unlike its predecessor, GTM considers the likelihood of training data points as the objective function. Also, a data point is not associated with one particular node but it is represented as a probability distribution over the entire latent space. Cumulating the probabilities over the data set, it is possible then to create continues chemical landscape which might serve for data sets visualization and comparison as well as for the building of regression and classification models.

The last group of methods can be illustrated on the example of Chemical Space Networks (*CSN*) [42] which combines both graph- and descriptors-based approaches. The idea is to represent chemical space as a huge graph where the nodes represent individual molecules, and the edges between the nodes are created as a function of either pairwise molecular similarity threshold or Matched Molecular Pair relations. CSN can be used to visualize a target-specific data set as an interactive graph where active and inactive molecules are grouped. These networks can efficiently be used for SAR exploration, and

they provide a depiction of target promiscuity, scaffold hopping [43] and/or similarity cliffs [44], where a single target exhibits activity for more than one class of compounds.

Despite the availability of a large number of various tools of chemical space analysis, only a few of them are suitable to be applied to Big Data. In our work, we focused on GTM possessing clear advantages over other methods because of its versatility, easy implementation and the possibility to combine options of data visualization, analysis, and modeling. A detailed description of GTM is given in the next section.

3 Generative Topographic Mapping (GTM) Overview

GTM is a dimensionality reduction algorithm well described in the literature [4, 5, 45]. Briefly speaking, the algorithm injects a 2D hypersurface (*manifold*) into an initial *D*-dimensional data space. The manifold is fitted to the data distribution by the Expectation-Maximization (*EM*) algorithm which minimizes the log-likelihood of the training data. Once the fitting is done, each item from the data space is projected to a 2D latent grid of *K* nodes. In the latent space, the objects are described by the corresponding vector of normalized probabilities (responsibilities). In turn, the entire data set can be represented by cumulative responsibilities. These cumulative responsibilities can be further visualized as a GTM Landscape or used to create regression or classification model.

3.1 Basics

3.1.1 Original GTM Algorithm

The algorithm was proposed by C. Bishop et al [4] in 1998. As it was already mentioned, GTM is a probabilistic extension of SOM where log-likelihood is utilized as an objective function. The manifold used to bind a data point $\mathbf{t}^*$ in the data space and its projection $\mathbf{x}^*$ in the latent space (Figure 12) is described by a set of *M* Radial Basis Function (*RBF*; Gaussian functions are used in the current implementation) centers.
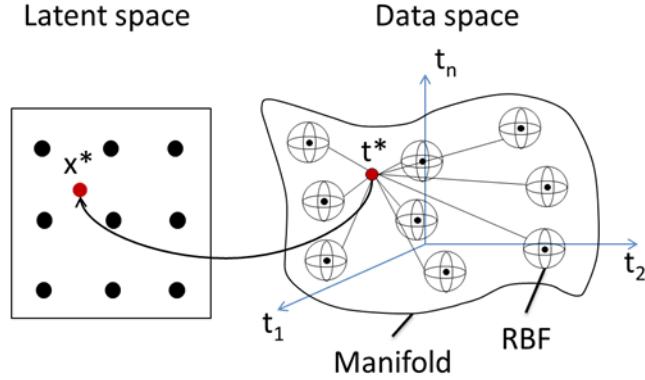
Figure 12. The basic idea of the GTM. Here, the data point **t*** from the multi-dimensional data space (right) is projected to **x*** the 2D latent space (left) using the manifold which is injected into the data space and described by a set of Radial Basis Functions (RBF).

To map the items from the initial space to the latent grid, the mapping function **Y** is used. It is described by K x M matrix (**Φ**) containing the RBF positions in the latent space with respect to the nodes, and the M x D parameter matrix (**W**) characterizing the position of the manifold in the initial space:

$$\mathbf{Y} = \mathbf{\Phi W} \tag{3.1}$$

The first step of the GTM training procedure is parameter matrix (**W**) initialization which can be done by randomization of the initial values or application of PCA where the first two principal components are used:

$$\mathbf{W} = \mathbf{\Phi}^{-1}(\mathbf{XU}) \tag{3.2}$$

Here, **U** is 2 x D matrix of the first two eigenvectors, and **X** is K x 2 matrix of nodes' coordinates in the latent space. The initialized manifold is inserted to the data space, and the initial log-likelihood value LLh($\mathbf{W}$, β) is computed using the 3rd eigenvalue as an initial guess of β^{-1}:

$$\text{LLh}(\mathbf{W}, \beta) = \frac{1}{N} \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{t}_n | \mathbf{x}_k, \mathbf{W}, \beta) \right\} \tag{3.3}$$

$$p(\mathbf{t}_n | \mathbf{x}_k, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi} \right)^{-D/2} \exp \left(-\frac{\beta}{2} \|\mathbf{y}_k - \mathbf{t}_n\|^2 \right) \tag{3.4}$$

On the second step, the EM algorithm is run which, first, computes the corresponding responsibilities $\mathbf{r}_n$, and then updates the parameter matrix $\mathbf{W}$ and β^{-1}:

E-step

$$r_{kn} = \frac{p(\mathbf{t}_n|\mathbf{x}_k, \mathbf{W}, \beta)}{\sum_{k'=1}^{K} p(\mathbf{t}_n|\mathbf{x}_{k'}, \mathbf{W}, \beta)} \tag{3.5},$$

$$g_{kk} = \sum_{n=1}^{N} r_{kn} \tag{3.6},$$

M-step

$$\widetilde{\mathbf{W}} = (\mathbf{\Phi}^T\mathbf{G}\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{RT} \tag{3.7},$$

$$\frac{1}{\widetilde{\beta}} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{kn}\|y(\mathbf{x}_k, \widetilde{\mathbf{W}}) - \mathbf{t}_n\|^2 \tag{3.8}.$$

In the equation (3.7), $\mathbf{T}$ is N x D matrix describing N data points in the initial D-dimensional space, λ is the regularization coefficient, and $\mathbf{I}$ is M x M unit matrix. The algorithm recomputes the LLh($\widetilde{\mathbf{W}}$, $\widetilde{\beta}$) using the updated $\widetilde{\mathbf{W}}$ and $\widetilde{\beta}$, and compare it with the LLh($\mathbf{W}$, β) obtained in the previous iteration. It can be seen from the equation (3.4) that the algorithm uses gradient descent minimizing the distance between the nodes and the data points. The manifold is considered to be trained enough when the EM algorithm achieves a certain threshold of convergence (e.g., LLh$_{new}$ − LLh$_{old}$ ≤ 0.001). Then, each data point is described on the 2D latent grid by its LLh and corresponding vector of responsibilities $\mathbf{r}_k$.

3.1.2 Incremental GTM Algorithm

The "Big Data" term is used to describe data sets of millions of data points. Such data sets can hardly be handled by the classical GTM algorithm due to the huge matrix of responsibilities ($\mathbf{R}$, equation (3.5)). In the case of large data sets (e.g. more than 50K compounds) it cannot be fully stored in the computer's RAM. In order to solve this issue, C. Bishop et al. have proposed to use an incremental GTM [40]. Within this approach, the manifold is initialized by a randomly chosen subset. Next, the data set is split into a series of blocks of a certain size which are used to train the manifold sequentially. In this scenario, the M step described in 3.1.1 is changed (equations (3.7) and (3.8)), and $\widetilde{\mathbf{W}}$ and $\widetilde{\beta}$ are

computed using two types of responsibilities: 1) new ($\mathbf{R}^*_{new}$) and old ($\mathbf{R}^*_{old}$) responsibilities of N^* structures produced for the new data block $\mathbf{T}_{new}$, and 2) responsibilities $\mathbf{R}_{old}$ computed for N structures from the previous block $\mathbf{T}_{old}$:

$$\widetilde{\mathbf{W}} = (\mathbf{\Phi}^T\mathbf{G}\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\{\mathbf{R}_{old}\mathbf{T}_{old} + (\mathbf{R}^*_{new} - \mathbf{R}^*_{old})\mathbf{T}_{new}\} \tag{3.9},$$

$$\frac{1}{\widetilde{\beta}_{new}} = \frac{1}{\beta_{old}} + \frac{1}{DN^*}\sum_{n=1}^{N^*}\sum_{k=1}^{K}(r^*_{new,kn} - r^*_{old,kn})\|y(\mathbf{x}_k, \widetilde{\mathbf{W}}) - \mathbf{t}^*_n\|^2 \tag{3.10}.$$

The next block of compounds is taken into the process only if convergence for the current one was achieved ($LLh_i - LLh_{i-1} \leq 0.001$). The incremental GTM algorithm was implemented by H. Gaspar et al. and tested in a compound library comparison project [5]. Its performance is discussed in chapter 3.4.2.

3.1.3 GTM Landscapes

To visualize and model chemical data, the GTM landscape is used [6, 45, 46]. With respect to different types of information, one can define three types of landscapes: 1) class landscape, 2) property landscape, and 3) density landscape. The examples are illustrated in Figure 13.
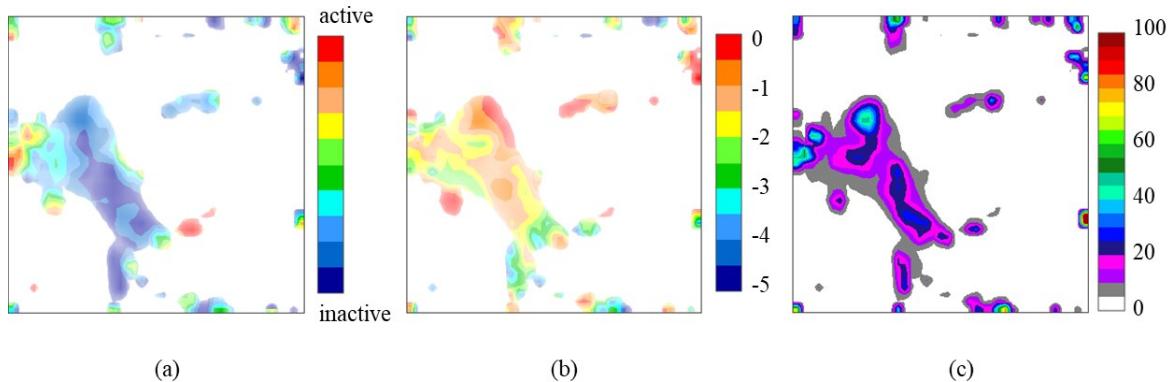
Figure 13. The example of class, property and density landscapes. The map was trained on vascular endothelial growth factor receptor 2 (CHEMBL279) data set containing 6.7K compounds. Here, (a) represents class landscape which demonstrates the distribution of molecules of two classes (active, inactive), (b) – property landscape (solubility, LogS), and (c) – density landscape providing the information about the nodes' population.

The class landscape represents a combination of classes' probabilities c_i computed as:

$$P(c_i|\mathbf{x}_k) = \frac{P(\mathbf{x}_k|c_i) * P(c_i)}{\sum_j P(\mathbf{x}_k|c_j) * P(c_j)} \qquad (3.11),$$

$$P(\mathbf{x}_k|c_i) = \frac{\sum_{n=1}^{N} r_{kn_{Ci}}}{N_{c_i}} \qquad (3.12),$$

$$P(c_i) = \frac{N_{c_i}}{N_{total}} \qquad (3.13),$$

where N_{c_i} is the number of items for the class c_i, N_{total} is the total number of training items, and r_{kn} is the responsibilities of the members of the class c_i in the node k computed according to the equation (3.5). To predict a class for a new compound q, the equation (3.14) is used:

$$P(c_i|\mathbf{t}_q) = \sum_{k=1}^{K} P(c_i|\mathbf{x}_k) * r_{kq} \qquad (3.14),$$

To visualize the landscape, normalized probability of the class c_2 is used as a color code (only a binary class landscape can be visualized at the moment). To consider the density of the nodes' population, transparency is added. In the case of a multi-class task (more than 2), GTM projections (the average positions of the items in the latent space) can be used instead of fuzzy GTM landscapes.

The second type of the GTM landscape is the property landscape which is used to visualize the distribution of a property over the latent space and which might serve as a regression model. The property landscape is defined by a list of property values corresponding to a particular node:

$$p_k = \frac{\sum_{n=1}^{N} p_n * r_{kn}}{\sum_{n=1}^{N} r_{kn}} \qquad (3.15),$$

where p_n is the property value for the compound n, and p_k is the mean property value for the node k.

The prediction of a property p for a new structure q is done similar to class prediction:

$$p_q = \sum_{k=1}^{K} r_{kq} * p_k \qquad (3.16).$$

To visualize the property landscape, p_k value is interpreted as a color code.

The last type of the GTM landscape – density landscape – is a special case of the property landscape where p_k is represented as a sum of responsibilities in the node k. This landscape is used to analyze the data distribution over the map which is not always obvious via the landscape's transparency.

3.2   GTM Parameters Tuning

GTM has four parameters (number of nodes, number of RBFs, regularization coefficient, RBF's width) needed to be optimized according to some scoring function. Besides these parameters, a "suitable" descriptors space and the frame set (usually a subset of representative compounds used to train the manifold; *FS*) size should be chosen. Two approaches are applied: grid search (brute force) and Genetic Algorithm (*GA*) [47]. The former investigates all possible combinations of 4 parameters. This approach is deterministic but it takes too much time and computational power. In contrast, GA is a stochastic approach but it allows the user to reach maximal fitness trying just a range of combinations which might lead to different endpoints in different runs. The workflow of the GA used to tune the GTM parameters and to select the suitable descriptors space and the frame set size is illustrated in Figure 14.

The details of the algorithm are already described in several publications [48–50]. Briefly speaking, GA generates a set of chromosomes composed randomly. All the attempts are cross-validated using the "selection" set (a set which differs from the FS and possesses activity/property values), and the mean Balanced Accuracy (*BA*) is computed. Next, the crossover and mutation of some attempts are applied, and the new attempts are computed.

The algorithm stops in case if it achieved the convergence (there is no attempt with larger BA during the two last generations) or the total number of attempts is exceeded.
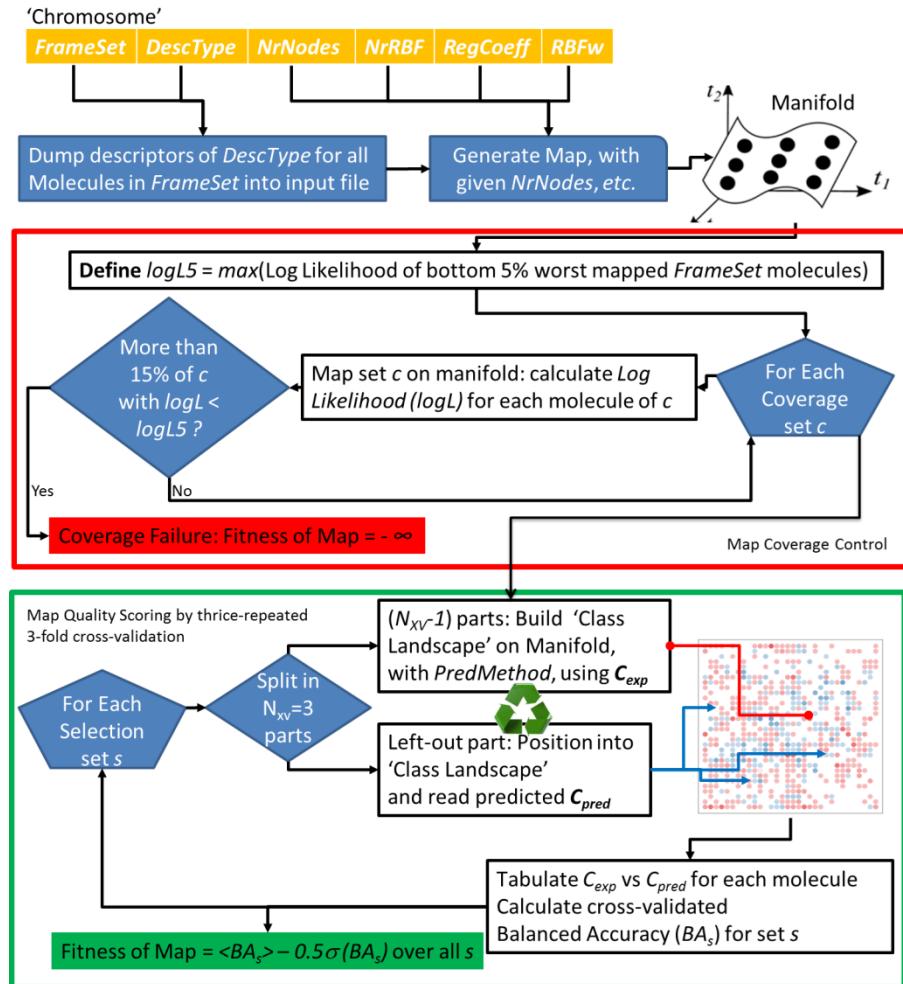
Figure 14. Evolutionary map selection scheme.

3.3 GTM-based Applicability Domain

Applicability Domain (*AD*) plays an important role in any machine-learning method. It allows the researcher to avoid costly wrong predictions in prospective virtual screening. For GTMs, five AD definitions were reported [46, 51]: 1) likelihood-based, 2) density-based, 3) class-dependent density, 4) predominant class AD, and 5) class entropy AD.

Within the likelihood-based concept, an item is considered out of AD if it is too far from the manifold in the initial data space. To filter such items, the LLh cutoff is determined. The approach to compute this cutoff is quite straightforward: the compounds

from the frame set are ordered accordingly to their LLhs, and it is assumed that the last n% of compounds are out of AD. Thus, the LLh cutoff is taken as the highest LLh out of this bottom n%. The density-based AD discards the nodes on the GTM landscape where the cumulative responsibility is below a certain threshold. This allows using only populated zones to make the predictions. The class-dependent density (*CDD*) AD is similar to the density-based AD. The difference is that the CDD AD checks only the density of the winning class c_{best} in the node, which has the highest conditional node probability $P(x_k|c_{best})$ (equation (3.12).

The predominant class AD is based on the selection of a dominant class in a node to which the maximal probability in this node corresponds. To control the predominance, a new class prevalence factor (*CPF*) was introduced. The idea is to discard the nodes in the latent space where the ratio of the classes' probabilities in a node is below the CPF. Herewith, the CPF becomes an additional degree of freedom which should be optimized to obtain a good model in terms of predictive performance.

The last approach is the class entropy-based AD. The class entropy S of the qth molecule is computed as:

$$S_q = -\sum_i P(c_i|q) \log(P(c_i|q) \tag{3.17}.$$

The entropy of the molecule is compared to the maximal entropy $S_{max} = \log(N_c)$ where N_c is the number of classes. The decision to discard the compound is made using the class-likelihood factor (*CLF*) computed as S_q / S_{max}. Thereby, CLF is high for the compounds with similar $P(c_i|q)$ for all classes, and low for the compounds with some dominant class (i.e. the $P(c_i|q)$ for this class is about 0.8-1.0). Thus, the compound is considered as out of AD if its CLF is above some threshold varying between 0 (all compounds are out of the AD) and 1 (all compounds are in AD).

3.4 Maps Application and Analysis

GTM is in practice a Swiss army knife of chemoinformatics, because it may serve in applications ranging from data visualization to libraries comparison, (multi-task) predictive modeling and AD control, *de novo* design, conformational space analysis, etc. (Figure 15). Here, we discuss some of them that were described in the literature so far.
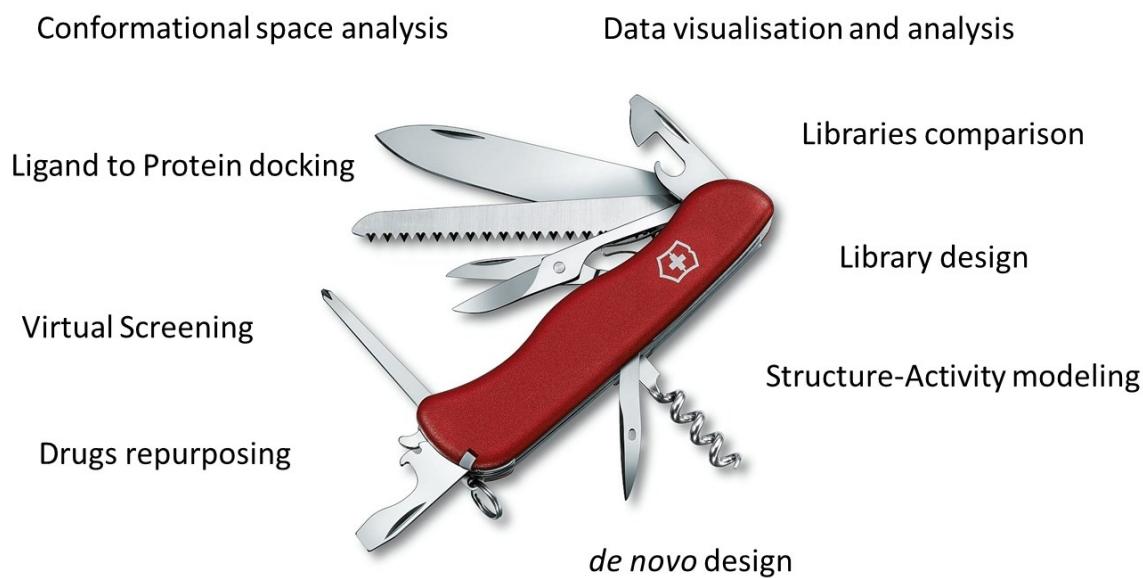
Figure 15. Areas of GTM application.

3.4.1 Obtaining of Classification and Regression Models with GTM

GTM has been already successfully applied as a tool for QSAR and QSPR modeling in many projects. In the paper by N. Kireeva et al. [52], the authors have demonstrated the application of the classification GTM to predict the melting point of ionic liquids. Three data sets were modeled, and the mean accuracy of the models in 5-folds cross-validation varied from 0.81 to 0.87. H. Gaspar et al. [6] have applied the regression GTM to model stability constants for metal binders, aqueous solubility, and activity of thrombin inhibitors. The authors compared the predictive performance of the regression GTM models to other machine-learning approaches, namely Self-Organizing Maps [41], Random Forest (*RF*) [53], k-nearest neighbors [54], M5P regression tree [55], and partial least squares [56]. External validation showed that RF overcomes the GTM in some cases (the difference of

the determination coefficients in cross-validation ΔQ^2 is up to 0.24). At the same time, the likelihood-based applicability domain (chapter 3.3) improved the performance and reduced the ΔQ^2 down to 0.1. A similar trend was demonstrated in the paper of T. Gimadiev et al. [57] where the authors applied GTM to model 21 inhibition activity for efflux and influx transporters.

Across many projects, it was demonstrated that GTM produces target- and property-specific models which quality is comparable to other methods. However, in contrast to other popular machine-learning approaches, GTM is an unsupervised method that trains its manifold using the unlabeled chemical data. Therefore, it can build a map not for a particular activity/property but for a given database which includes thousands and millions of compounds. This idea was extended and tested by P. Sidorov et al. [9] which have proposed a concept of a universal map. The authors aimed to cover a large chemical space of around 1.3M compounds (ChEMBL database of version 20) using a single map. The descriptors space and the GTM parameters were selected using the Genetic algorithm described in chapter 3.2. The results showed that the universal approach is able to cover efficiently large range of chemotypes. Several tests ("challenges") were done to prove its performance. For instance, the best map selected by GA was cross-validated on 410 ChEMBL targets, and about 80% of the targets were predicted with the mean Balanced Accuracy of 0.7.

The universal approach described in [9] has demonstrated that GTM is ready to model Big Data, and it can be also used in multi-target machine learning where the universal map can predict several activities/properties without training a new model. This also opened the door to large-scale Virtual screening (*VS*). In the context of the given work, Virtual Screening is defined as an application of QSAR to model and predict Big Data. Very recently, GTM was shown as a nice tool for VS [58]. The authors trained GTMs in different descriptors spaces on ChEMBL data. It was established that one descriptors space is not sufficient, and at least 7 fragmentation schemes are needed. It was also shown that

the consensus approach made on several maps gives better accuracy than single-map predictions.

3.4.2 Data Analysis and Chemical Libraries Comparison

Besides QSAR/QSPR studies, GTM was applied to visualize and analyze chemical data. For instance, GTM was used to visualize and cluster the data on motor unit action potential [59]. The authors of the study trained GTM on nine data sets and then used the latent grid as a basis for data clustering. In the paper of D.M. Maniyar et al. [37], the authors applied hierarchical GTM [11] to visualize the distribution of active and inactive classes for five data sets (GPCRs and Kinase) obtained from different high-throughput screens. They trained a manifold using these five data sets, and, if the map resolution was not sufficient to distinguish the compounds from different classes, they extracted the compounds from such a "mixed" area and retrained a "child" manifold. GTM has even been proposed for nonlinear fault identification in a chemical process [60].

Also, an attempt to combine the GTM method with Chemical Space Networks (*CSN*) [42] was done [61]. The authors proposed the two-layered SAR visualization concept for SAR exploration of increasingly large compound data sets. The underlying idea is to first generate global "bird's eye" views of the activity landscapes of large data sets to identify SAR-informative regions for more detailed analysis. Then, selected regions were further analyzed by the CSN at the level of individual compounds. The GTM-CSN technique was applied to analyze three relatively small activity-specific compound series (up to 2.2K compounds) extracted from BindingDB [62, 63] and big antimalarial screening (up to 13K compounds) data set [64]. The authors checked structural modifications resulting in potency changes and discussed it in the example of several analogs where such modifications increased the pKi value (e.g. from 6.1 to 8.1 pKi).

Despite a large number of different GTM applications, yet, it was used to analyze only relatively small data sets (up to 20-30K compounds). The first attempt to visualize large data sets (2.2M compounds) was done by H. Gaspar et al. [5]. The authors applied the

incremental GTM (chapter 3.1.2) to compare 36 commercial libraries and the NCI database in terms of molecular properties (molecular weight, number of H-bond donors and acceptors, chirality, logP, TPSA, etc.), similarity (Tanimoto coefficient), and compounds distribution over the 2D latent space. The libraries were also compared using meta-GTM where a map was trained on all 37 libraries. Each library was considered as a single object represented by cumulated responsibilities or property landscape values at nodes x_k. The authors also showed that some regions of interest can be detected in the landscape using the desired property landscapes. This brought us closer to Big Data, but still, the analysis of the structures residing the nodes was done manually.

To automate that, the Responsibility Pattern (*RP*) term was introduced by K. Klimenko et al. [65]. The idea was to group structures that reside neighboring nodes on the map using their responsibilities. RPs allowed to detect and to extract compounds that are similar in the latent space automatically to search then for privileged structural motifs (*PSM*).

The concept of "privileged substructures" was originally introduced by B.E. Evans et al. [66], referring to core structures that are recurrent in compounds active against a given target family and, therefore, associated with that biological activity. Privileged substructures are thought to be selective toward a given target family but not individual family members. Most of the earlier studies focused on the exploration of molecular core structures or scaffolds, and some privileged scaffolds have been proposed for drugs and natural products. However, it was shown in [65] that common structural motifs may vary from precisely defined scaffolds or even substituted scaffolds, to fuzzier ensembles of related, interchangeable scaffolds, to even fuzzier 'pharmacophore-like' patterns.

The PSM approach allowed chemists to relate a particular activity/property to a certain chemical pattern. The PSM technique was also applied in modeling and analysis of antimalarial compounds [49]. The authors highlighted some of the specific privileged patterns linked to antimalarial activity (e.g., naphthoquinones and 4-aminoquinolines). Later, the method was modified by the application of retrosynthetic rules (*RECAP*) [67].

The authors tried to extract the "frequent" RECAP cores to identify PSMs for inhibitors of protease, kinase, and GPCRs. However, the workflow where the PSM was implemented still includes some steps that must be done manually (PSM are extracted by hands). This limits the workflow and restricts it in the analysis of larger data sets.

3.4.3 GTM for Conformational Space Analysis

Another application of GTM was found in the analysis of conformational space. Conformational sampling is the key to the fundamental understanding of molecular properties. It plays an important role in medicinal chemistry since different conformations may possess different biological activities (in terms of IC50, EC50 or Ki). Several techniques are applied in conformational sampling [68–70]. However, GTM has a clear advantage in the context of conformational space visualization.

The general idea of GTM application in conformational sampling was described by D. Horvath et al. [71]. One can train a map using "contact" or "interaction" fingerprints as well as torsion angles as descriptors to predict total, non-bonded and contact energies, surface area or fingerprint darkness. For this purpose, a set of (previously generated) conformers with known score values (e.g. total energy computed by AMBER force field [72]) can be used to prepare frame, color and test sets. Next, the Genetic Algorithm (see chapter 3.2) is run to tune the GTM parameters. Once the algorithm achieved convergence (e.g. root mean square deviation does not change a lot), the obtained map can be used to visualize and analyze the corresponding conformational space as well as to predict the energy of a new conformer or to sample conformers using the property landscape as a basis in the reverse task (projection from the latent space back to the initial space).

The described approach was evaluated by the authors in the task of monitoring the conformational space of dipeptides [73]. Later, it was applied to the docking problem [74]. The concept was illustrated by a docking study into the ATP-binding site of CDK2. The maps trained on contact fingerprints and hybrid descriptors (contact fingerprints in combination with ligand fragment descriptors) were used to discriminate native from non-

native ligand poses and to distinguish ligands by their potency. It was shown that the maps trained on hybrid descriptors possess higher prioritization performance (the Area Under the Receiver Operating Characteristics Curve is above 0.8) and, thus, they can be efficiently used in Virtual Screening campaigns.

3.4.4 GTM in *De Novo* Design

Besides data analysis and modeling, GTM is also used in *de novo* design of new structures. In 2014, K. Mishima et al. [75] applied GTM in a loop of biological activity assessment of virtually enumerated structures. The seed structures were selected from the activity landscape and modified in various ways to generate new structures. The generated structures were filtered after by the same GTM activity landscape and used (in case of success) as new seeds. The loop stops when enough structures are generated. This algorithm was also applied by S. Takeda et al. [76] to generate a set of drug-like molecules.

Another attempt to use GTM in the generation of chemical structures with desirable activity(ies) was made by introducing the Stargate GTM [77]. Here, GTM was used to bind descriptors and activities spaces by training two manifolds in both spaces in parallel. The defined "reverse" mapping function allowed to "jump" from the activities space back to descriptors space and, hence, to determine the desirable descriptors vectors. Next, one can generate structures with high similarity to the returned vectors assuming that these new structures will possess the requested activity profile.

Besides, GTM was also combined with auto-encoder where the map was trained on the generated latent descriptors. B. Sattarov et al. [78] analyzed the binding potency of automatically generated 394 ligands for the Adenosine A2a receptor. These ligands were docked to the binding site using S4MPLE docking method [79]. It was shown that the average docking score of the generated structures is even better than the average docking score of real active molecules.

3.5   Conclusion

In recent studies carried out in the Laboratory of Chemoinformatics, Generated Topographic Mapping designed by C. Bishop as a data visualization approach was significantly extended on the modeling and analysis of chemical data. This PhD project represents a continuation of these studies. Our main challenge concerned the further extension of GTM toward Big Data, which, in turn, may require using large frame sets (FS) in combination with large dimensionality of the initial data space for manifold construction. Since the capacity of earlier reported classical and incremental algorithms for manifold construction was limited, our goal was to design a new more efficient algorithm.

In earlier studies, relatively small FSs were used to build GTM for large chemical databases. However, a systematic investigation of GTM performance as a function of FS size was never performed. This question was considered in our work.

In this thesis, we also tackled some other methodological problems. The first one concerned a rational determination of the log-likelihood threshold used for defining the applicability domain of GTM-based models. The second one dealt with an automatized protocol of Maximum Common Substructures extraction from the ensemble of structures populated selected area on the map.

Some earlier reported options of GTM-based data analysis were fully automatized in this work. It concerns (*i*) selection of zones of interest [5] and, (*ii*) hierarchical GTM zooming [11, 37].

Developed algorithms and tools were used in three projects: (*i*) application of GTM to virtual screening (VS), (*ii*) comparison of large databases, and (*iii*) enrichment of proprietary library.

4 Methodological Developments

4.1   Descriptor normalization for GTM

The Generative Topographic Mapping (GTM) method is sensitive to the descriptors and its preprocessing. For instance, the PCA, which is the first step of GTM, requires the descriptors to be centered. Therefore, it is needed to find a suitable scheme of descriptors preprocessing which provides the user with a better map. For this purpose, five preprocessing schemes were compared to each other and the scenario when no preprocessing was done:

1) No preprocessing;

2) Standardization (centering and division by its standard deviation);

3) Centering;

4) Scaling to [-1;1];

5) Scaling to [-1; 1] and centering.

To see the impact of different preprocessing schemes, a set of 98 compounds active against the tyrosine kinase inhibitors (*SRC*) and 980 decoys were extracted from the Directory of Useful Decoys (*DUD*) [8]. The structures were standardized (aromatized, explicit hydrogens were removed, common chemical groups like nitro group were transformed, etc.), and ISIDA descriptors were generated (atom-centered sequences of atoms and bonds with a length of 1 to 3 atoms) [80]. The descriptors were preprocessed according to 5 scenarios mentioned above, and a GTM was trained using the following parameters: 625 nodes, 144 RBFs, RBF's width is 2.82, and the regularization coefficient is

1.0. The 2/3 part of the data set was used to build the class landscape, and the rest was used as a test set to assess the predictive performance in terms of Balanced Accuracy (BA) and Area Under the Receiver Operating Characteristics Curve (ROC AUC).

Table 1. Validation results of the GTMs trained for the SRC data set with different preprocessing schemes. A probability threshold of 0.5 was used for BA assessment.

Preprocessing scheme	BA	ROC AUC
No preprocessing	0.71	0.88
Standardization	0.72	0.88
Centering	0.74	0.66
Scaling to [-1;1]	0.49	0.72
Scaling to [-1;1] and centering	0.52	0.91
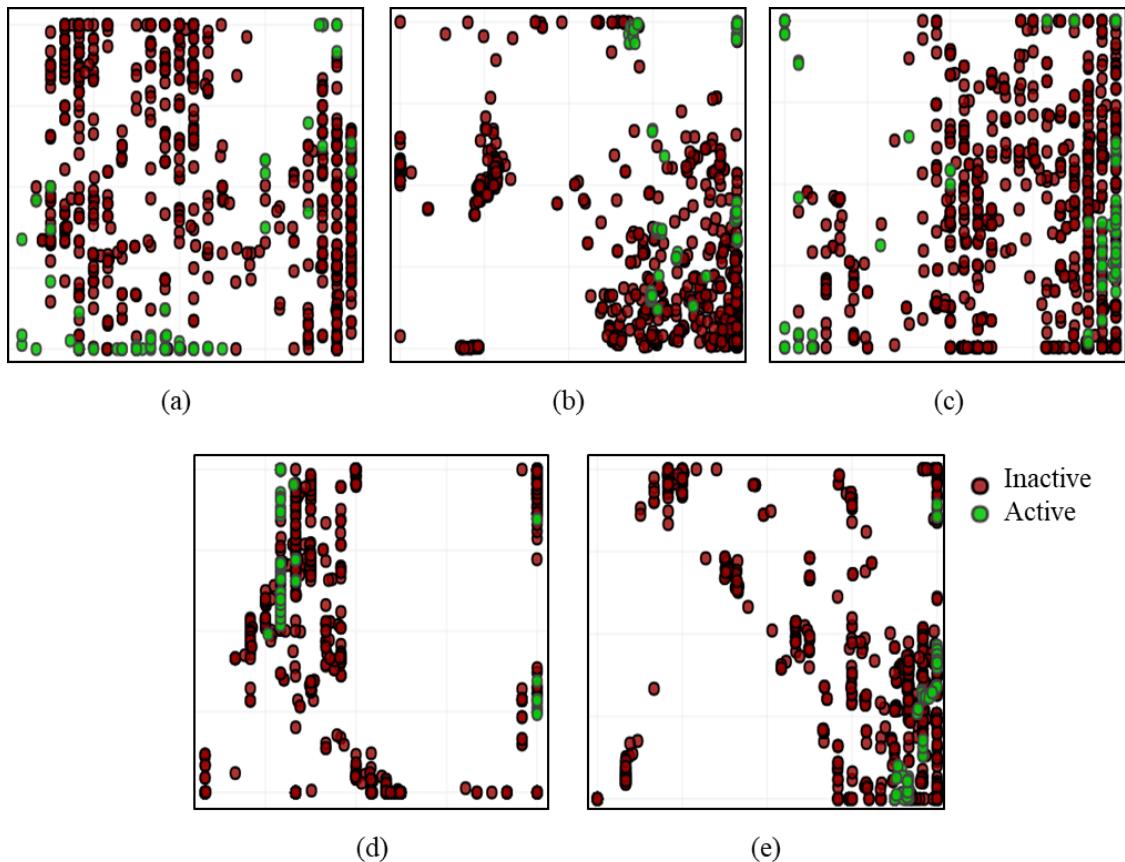
Figure 16. GTM projections of the SRC data set with (a) no descriptors preprocessing, (b) descriptors standardization, (c) centering of the descriptors, (d) scaling the descriptors, and (e) scaling and centering the descriptors.

The results in Table 1 and Figure 16 demonstrate that the GTM trained with the original descriptors performs similarly to those built on standardized descriptors. On the other hand, the items are better spread on the former map (Figure 16a) than on the others. Notice that the above results correspond to a particular data set and descriptors type.

4.2   GTM Applicability Domain (AD)

The Applicability Domain (AD) topic was already discussed in chapter 3.3. The approaches described in [51] use tunable parameters which bring an additional degree of freedom to the model optimization procedure. So far, the predominant class AD needs the class prevalence factor (CPF) for each GTM landscape to ignore the mixed nodes which, in turn, decreases the density of the landscape. The class entropy AD needs a threshold for the class-likelihood factor (CLF). These ADs make the GTM tuning procedure described in chapter 3.2 more complicate.

In the author's opinion, the likelihood-based AD described in chapter 3.3 is the most simple and intuitive approach. Predictions made for the compounds which are away from the manifold will be worse in terms of confidence than for the compounds which are closer to it. The shape of the LLh distribution of the frame compounds (the axis X represents the LLh, and the axis Y represents the number of compounds) is similar to the shape of a shifted Gaussian distribution. The LLh values vary from $-\infty$ to 0, and the peak of this distribution corresponding to the major part of the frame set situates near 0. The right part of the distribution is very short since no compounds can be predicted with LLh>0. In contrast, the left part possesses a very long "tail" (the blue line in Figure 17).

If the LLh distribution would perfectly follow the normal distribution, the top 95% (i.e. 5% beyond the threshold) of the frame compounds would form an area under the Gaussian curve where the last one is cut in the $\mu \pm 2\sigma$ range. However, this LLh distribution is not perfectly normal (besides the fact that it is shifted). Many attempts to fit a Gaussian to the LLh distribution minimizing the root mean square error (RMSE) were done. The schematic example is shown in Figure 17, and RMSE was computed as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{NL}\left(Nc_i^{GTM} - Nc_i^{Gauss}\right)^2}{NL}} \qquad (4.1),$$

where NL is the number of unique LLh with a non-zero number of the frame compounds, and Nc_i^{GTM} and Nc_i^{Gauss} are the numbers of the frame compounds given by GTM and fitted Gaussian at particular log-likelihood value LLh$_i$ ($Nc_i^{GTM} - Nc_i^{Gauss}$ is named "deviation" in Figure 17). It was found that the RMSE is always above zero. Therefore, to determine the meaningful AD, a Gaussian approximation is needed.
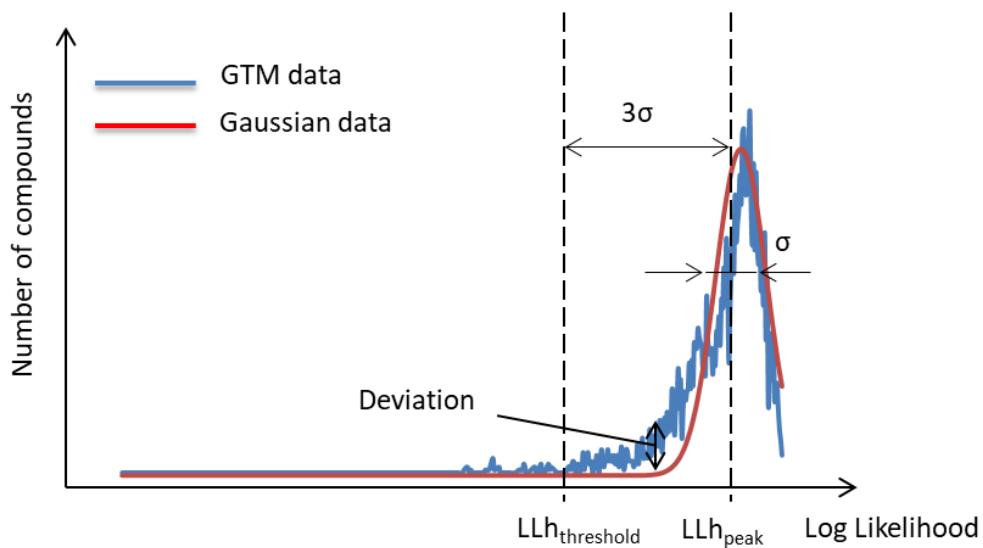
Figure 17. An example of a Gaussian (red line) fitted to the log-likelihood data distribution (blue line) of "Thrombin" (CHEMBL204) data set. GTM Applicability Domain is identified here by log-likelihood threshold LLh$_{threshold}$ = LLh$_{peak}$ − 3σ. Here, LLh$_{peak}$ and σ are, respectively, the peak position and the width of the Gaussian function.

4.3   Automatized Hierarchical GTM Zooming

The map resolution is a known problem of GTM in Big Data. The molecules of different classes might be projected to the same zone on the map. This makes the zone uncertain (mixed). As it was described in chapter 3.3, an attempt to discard such mixed zones was already made considering them as out of the applicability domain. This removes the uncertainty but it also reduces the number of populated nodes on the landscape.

I. Nabney and P. Tino [11] suggested solving the resolution problem by training a new GTM manifold using the items of a selected area as a training set. The compounds used to train the "child" manifold are selected manually using projections on a "parent" map where each structure is represented as a single point. The authors created a multi-level hierarchical GTM tree and tested it on toy data sets. It was also tested in a task of analysis of GPSR activities [37]. In this project, we propose an automatized GTM zooming approach where individual projections are replaced by a class landscape (see chapter 3.1.3). Thus, a compound is extracted from a zone of interest (e.g. a square cluster of nine nodes) basing on the sum of its responsibilities in this zone which has to be larger than a certain threshold (e.g. 0.8). The child manifold is trained then using these compounds as a frame set with the same descriptors and GTM parameters. The likelihood-based AD described in chapter 4.2 can be then applied if needed. The approach was tested in the project of private chemical collection enrichment (see chapter 7; Figure 18).
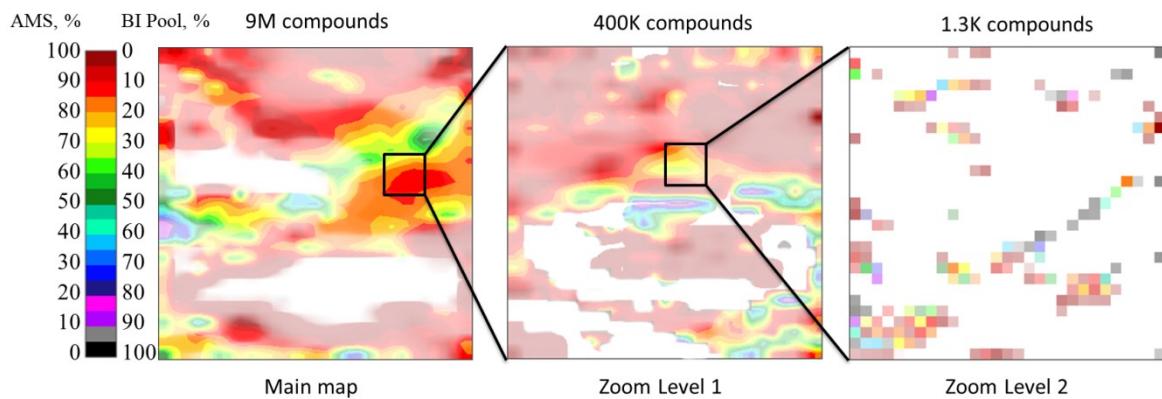
Figure 18. An example of the hierarchical GTM zooming approach applied to large public and private chemical databases comparison. Here, the map is trained to cover Aldrich-Market Select (AMS, 8.5M compounds) data set and the in-house collection of Boehringer Ingelheim (BI Pool, 1.7M compounds; see chapter 7).

It is shown in Figure 18 that the second level of zooming discovered some areas populated exclusively by the private compounds (black nodes), whereas the *parent* area was shown in red (mostly public data).

Within the automated procedure, the zones can be selected accordingly to two scenarios: 1) the grid of nodes can be simply divided into a set of joined square clusters of 3x3 nodes (Figure 19a), or 2) the zones can occupy the grid sharing the nodes on the borders between each other (Figure 19b). The advantage of the second scenario is that the items which locate on the border of a zone and are not considered as members of this zone due to the responsibility threshold, they will be taken by the neighboring zone. This generates more zones than the simple strategy but it can be easily reduced by increasing the zone size. In turn, the second strategy brings more items to the subsets than simple division, and, thus, more chemotypes can be analyzed further.

(a)   (b)
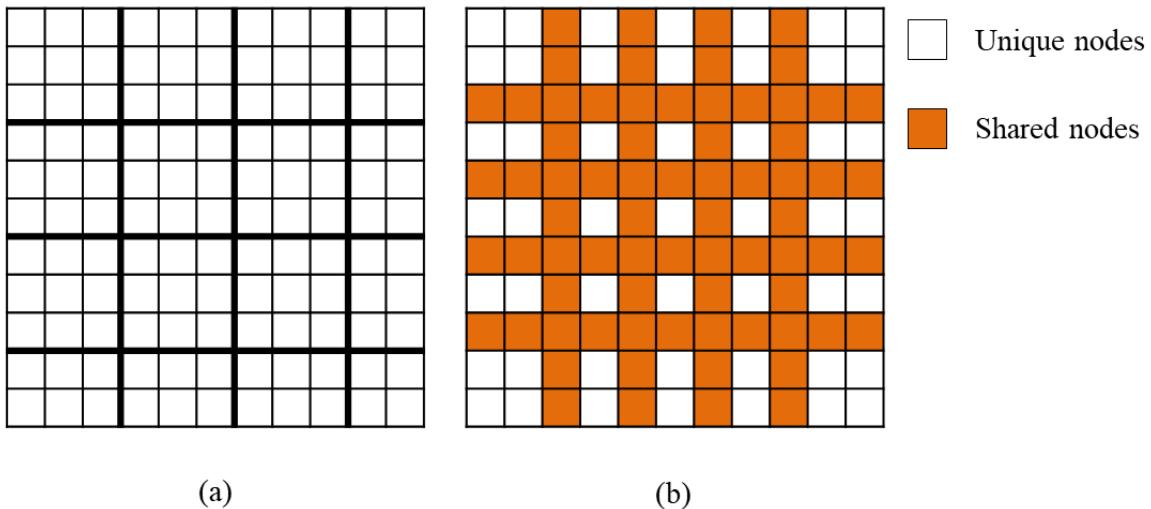
Figure 19. Zones selection schemes: (a) simple division of a grid of nodes (GTM landscape) into a set of square clusters of 9 nodes where the zones' borders are highlighted by orange lines; (b) zones selection using overlap. The zones on the scheme (b) have their own nodes in the white-areas as well as the nodes on the borders shared with the neighboring zones (orange).

As soon as the zones are delineated, the decision to zoom or not to zoom is made based on the number of extracted compounds (for instance, at least 1000 items must be extracted). Child GTMs are trained then using these subsets as frame sets. In the case of large subsets (i.e. larger than 10,000 items), the frame set size should be controlled. Therefore, not the entire subset but only 10% of it (but not less than 1000 items) are used to

train the manifold. After, the analysis of zones of a child manifold is repeated, and if the population of some zone is still too high, the zooming procedure repeats.

4.4   Automatized Maximum Common Substructures Extraction from GTM

The GTM provides chemists with a chemical landscape that can be visualized and analyzed. However, no relation between structural patterns and particular zone on the map is provided. For this purpose, the responsibility patterns (*RP*) method has been proposed to group the compounds which were then analyzed by the Scaffold Hunter tool to identify common scaffolds/substructures [49, 65]. Compounds sharing the same RP will typically share some common structural features that are further manually processed to annotate the map. This is a tedious and error prone-task. As an alternative, we propose to exploit the Maximum Common Substructure (*MCS*) search to automatically highlight shared features. Our solution is based on ChemAxon's JChem engine [81]. The MCS extraction protocol is described in Figure 20.

Here, an arbitrarily selected structure in the list of N items is compared to the other N-1, resulting in N-1 connected MCSs. A size filter keeps only the MCS covering at least 30% of the heavy atoms in both structures of a pair. Then, duplicate MCSs are grouped and the unique MCSs are sorted according to their occurrence in the list. The most frequent MCS is selected. Structures featuring the selected MCS are removed from the list, and a new iteration is started.

K. Klimenko et al. [65] demonstrated that common structural motifs may range from precisely defined scaffolds or even specifically substituted scaffolds, to fuzzier ensembles of related, interchangeable scaffolds, and to even fuzzier 'pharmacophore-like' patterns. Therefore, the perspective here is to use the disconnected MCS which would describe the molecular core as well as the substituents.
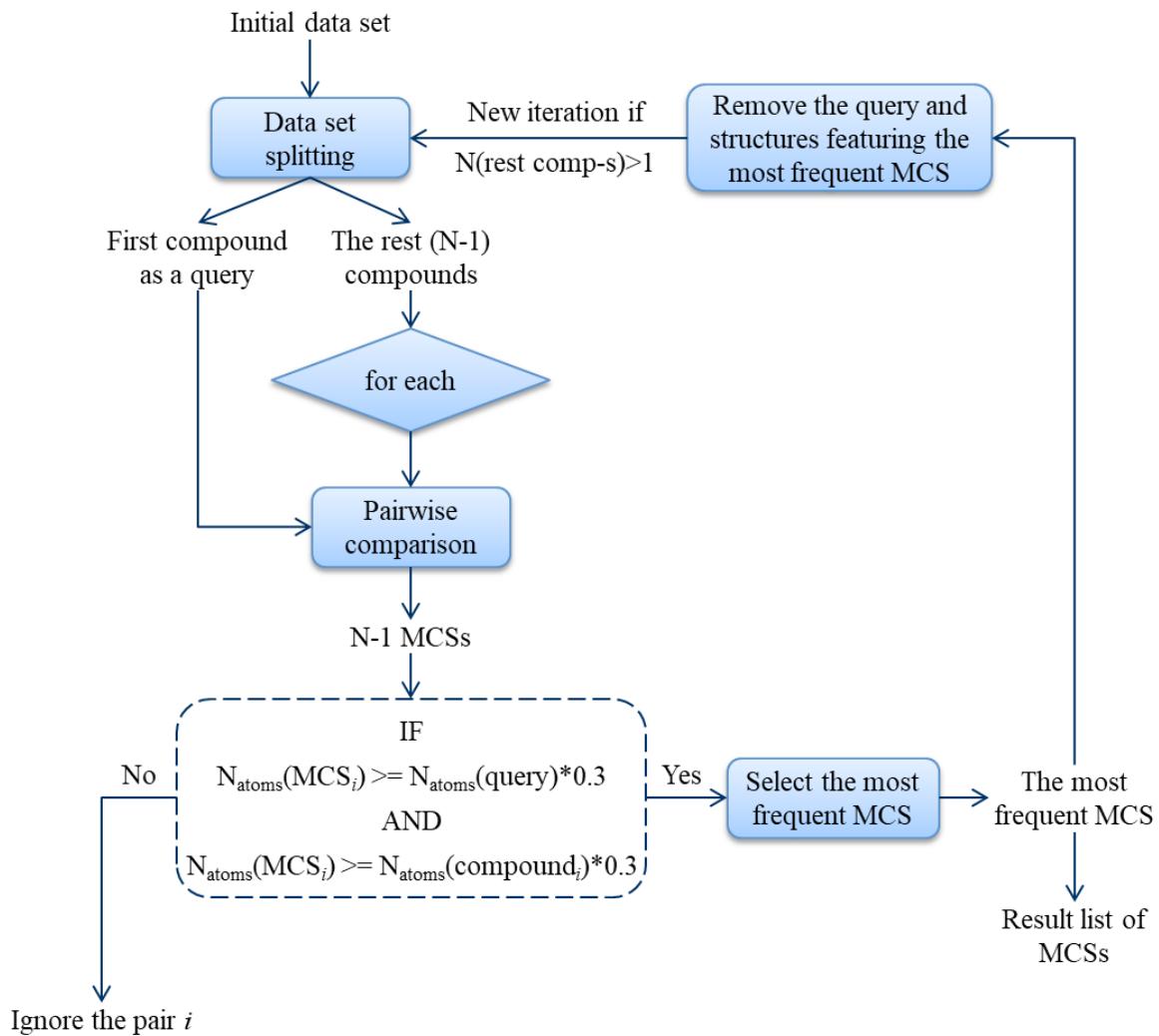
Figure 20. Maximum Common Substructure search protocol.

4.5   Constrained Screening

Nowadays, searching for drug candidates quite often involves screening of chemical libraries of sizes ranging within 10K ÷ 10M compounds. Many different methods of machine-learning are applied to treat big real and virtual chemical libraries [82–86]. In this case, the usual virtual screening (VS) procedure includes many steps where each of them tends to decrease the size of a screening pool, in discarding the unappropriated compounds according to the methodology at that step. Faster and less accurate steps proceed first, operating on the entire library – sophisticated ones later, operating only on subsets passing the fast ones. However, the large size of the potential drug-like space makes us change our

vision of virtual screening. Instead of saving some milliseconds per compound, we should optimize the VS algorithms. The idea of screening the entire pool against the required profile (desirable and/or undesirable activities, ADME properties, etc.) once brings us to the concept of *Constrained Screening* (CS).

CS is based on a universal GTM trained for a large data set (see chapter 3.4.1). The manifold produced by the universal approach covers a wide range of chemotypes and it is applicable to model different biological activities and properties. In particular, on a given GTM landscape describing a property (activity), **P** one can easily select some zones populated by molecules for which the property varies in the range $P_{min} < $ **P** $ < P_{max}$, where P_{min} and P_{max} are the user-defined thresholds. Such zones were named "regions of interest" and described in [5]. As it was mentioned in the paper, to identify the location of molecules possessing desirable profile {**P₁**, **P₂**, ..., **P_N**}, one can superimpose corresponding property landscapes. Then, these regions can be analyzed and/or corresponding compounds can be extracted.

The concept of zones of interest was also applied in [57] where the authors trained a map for human intestinal transporters. It allowed delineating the areas on the map populated either by molecules exhibiting inhibition but not transport activity or vice versa.

In this project, we automatized the zones of interest selection. Since these zones may overlap fully or partially, we also propose a concept of a *Query Landscape* which describes zones populated by molecules possessing desirable profile entirely (all **P**$_i$ are confined in user-defined intervals) or partially (some **P**$_i$ are out of the range).

In Figure 21, an example of the query landscape is shown where the vascular endothelial growth factor receptor 2 (CHEMBL279) data set containing 6.7K compounds was used to train the manifold. For the demonstration purpose, the request on CHEMBL279 activity, solubility (LogS) and the number of H-bond donors was modeled.
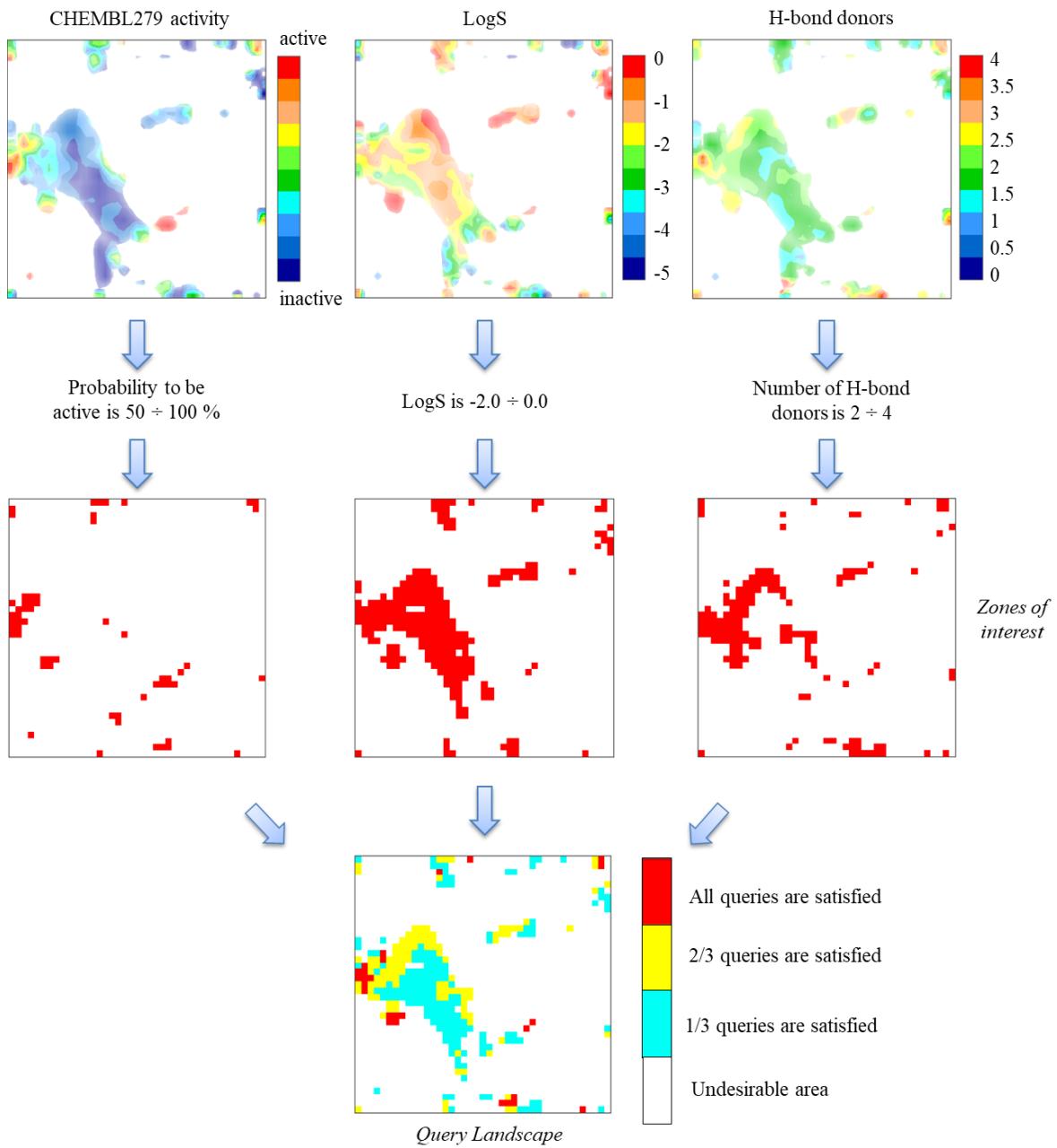
Figure 21. An example of a query landscape where the map is trained on Vascular endothelial growth factor receptor 2 (CHEMBL279) data set (6.7K compounds) using ISIDA fragment descriptors [10, 58]. Here, the query is set to find areas where the probability to be active varies from 50% to 100%, LogS is between -2.0 and 0.0, and number of H-bond donors ranges within 2-4. The first line represents the individual GTM landscapes, the second line represents the areas of interest on the individual landscape, and the last one is the query landscape.

The corresponding GTM landscapes were built, and a query was prepared: the probability to be active in the range of 50-100%, LogS varies from -2.0 to 0.0, and the number of H-bond donors ranges from 2 to 4. Next, the GTM landscapes were filtered according to the query, and the zones of interest were shown (red areas in the middle line of landscapes Figure 21). The overlaying of these zones results in a query landscape where the red areas satisfy all the conditions in the query, yellow ones correspond only to two out of three, and blue areas represent the zones where only one out of three conditions is satisfied. The white areas on the query landscape represent the zones where no training molecules with desirable activities/properties were found.

Query Landscape can be applied (*i*) to select a focused subset from the database used for GTM construction, and (*ii*) for virtual screening of an external database. In the latter case, a satisfaction score is assigned to each compound in the pool which means how well the compound fits the query. The approach was implemented as a web-tool. It is described in chapter 8.5.

4.6   Parallel GTM (PGTM)

Generative Topographic Mapping (GTM) [4] is a perspective tool used to visualize, analyze and model chemical data. Its advantages in comparison to other dimensionality reduction methods were already demonstrated in several projects [6, 45, 65]. The maps trained on data sets of a regular size (up to 10,000 items) as well as the ones trained to describe millions of compounds were presented [5, 9, 10, 50, 58]. The demonstrated results show that GTM can be successfully applied to large chemical databases visualization and comparison as well as in virtual screening campaigns. However, the limitation on the number of training data points restricts GTM to treating millions of structures during the training procedure. To overcome the limit, a frame set (FS) is gathered which is supposed to represent the chemical space sparsely. This FS of few thousand data points (e.g. 25,000 structures) is used to set the initial position and to fit the manifold in the initial data space. Once the manifold is fitted, the entire data set is projected and filtered using the likelihood-

based GTM Applicability Domain (AD). Further, these projections can be used to build a classification or regression GTM landscape which can serve as a QSAR or QSPR model [6].

GTM does not require the chemical space to be dense to train the manifold, and, hence, the chemical space of some million structures can be easily represented by some thousands. At the same time, the potential global chemical space of drug-like molecules is estimated as 10^{33}, and it can hardly be described just by some thousands of structures [3]. Therefore, a new strategy to treat larger frame sets is needed.

FS size is limiting in several ways: by (i) the amount of RAM used to store the large matrix of responsibilities, and (ii) the time spent to perform some matrix operations implemented in the GTM algorithm. An attempt to accelerate the algorithm was already made by parallelization of it using Message Passing Interface (*MPI*) technique [87–89]. To this purpose, the matrix of responsibilities was decomposed and its parts were distributed over the CPUs to be updated by small chunks of the data set iteratively. The disadvantage of this approach is the dependency of the code on the certain architecture of a machine used to run the calculations. Namely, a single machine or a highly organized cluster that supports the MPI technology has to be used for calculations, and the RAM has to be shared between the machines to store the whole matrix of responsibilities. If the first issue can be solved by purchasing a better machine, the second one will limit the calculations as in terms of storing the objects as in terms of speed (the mpi technology will spend some time to transmit the data from one machine to another). Besides that, this does not solve the problem of manifold overfitting which was detected by D. Ormoneit and V. Tresp [90]. It was shown that the Expectation-Maximization algorithm tends to overfit the Gaussians-mixture model.

In this chapter, we present a new attempt to parallelize the GTM which is supposed to speed up the calculations, to solve the problem of overfitting and to support the use of larger frame sets.

4.6.1 Method

The limitation of the classical GTM algorithm is the memorization of the large matrixes of responsibilities (**R**) and descriptors (**T**). To control the size of **R**, incremental GTM was proposed by H. Gaspar et al. [5]. Within the incremental approach (chapter 3.1.2), the equations (3.7) and (3.8) were modified to (3.9) and (3.10), respectively. Thus, the initial data set was divided into a batch of blocks of a certain size (e.g. 10,000 items) and treated sequentially. This solved the problem of the **R** size but the order of the chemotypes coming from different blocks begins to impact the shape of the manifold. So far, the initial manifold position is determined only by the first block, and then the manifold learns the shape of data distribution analyzing each block sequentially. As a result, the impact of the middle blocks on the final shape of the manifold becomes lower in comparison to the later ones. This brings us to the phenomenon when the chemotypes allocating in the middle of a data set might be forgotten by the manifold since the final shape of it is mainly formed by the first and the last blocks.

To overcome the limits of the classical GTM algorithm and to solve the problems of the incremental algorithm, we propose the new Parallel GTM (*PGTM*) approach. The basic idea of it is described in Figure 22.

Within this approach, we distinguish the manifold initialization and manifold training procedures. To initialize the manifold, the incremental Principal Components Analysis (PCA) is applied to the entire data set where the two first components are computed. To do so, the covariance matrix is computed incrementally followed by the Eigenvalue decomposition [91] using a graphical card (the scikit-cuda library in Python was applied) [92]. Once the PCA is done, the initial **W** and β are computed, and the manifold is trained on different blocks of the data set in parallel. The fact that the same initial manifold and the same GTM parameters are used to treat the blocks, the tasks can be independently distributed to different machines with no preferable architecture. In addition, no RAM sharing is needed since the size of a particular matrix **R** is determined only by the size of a block.
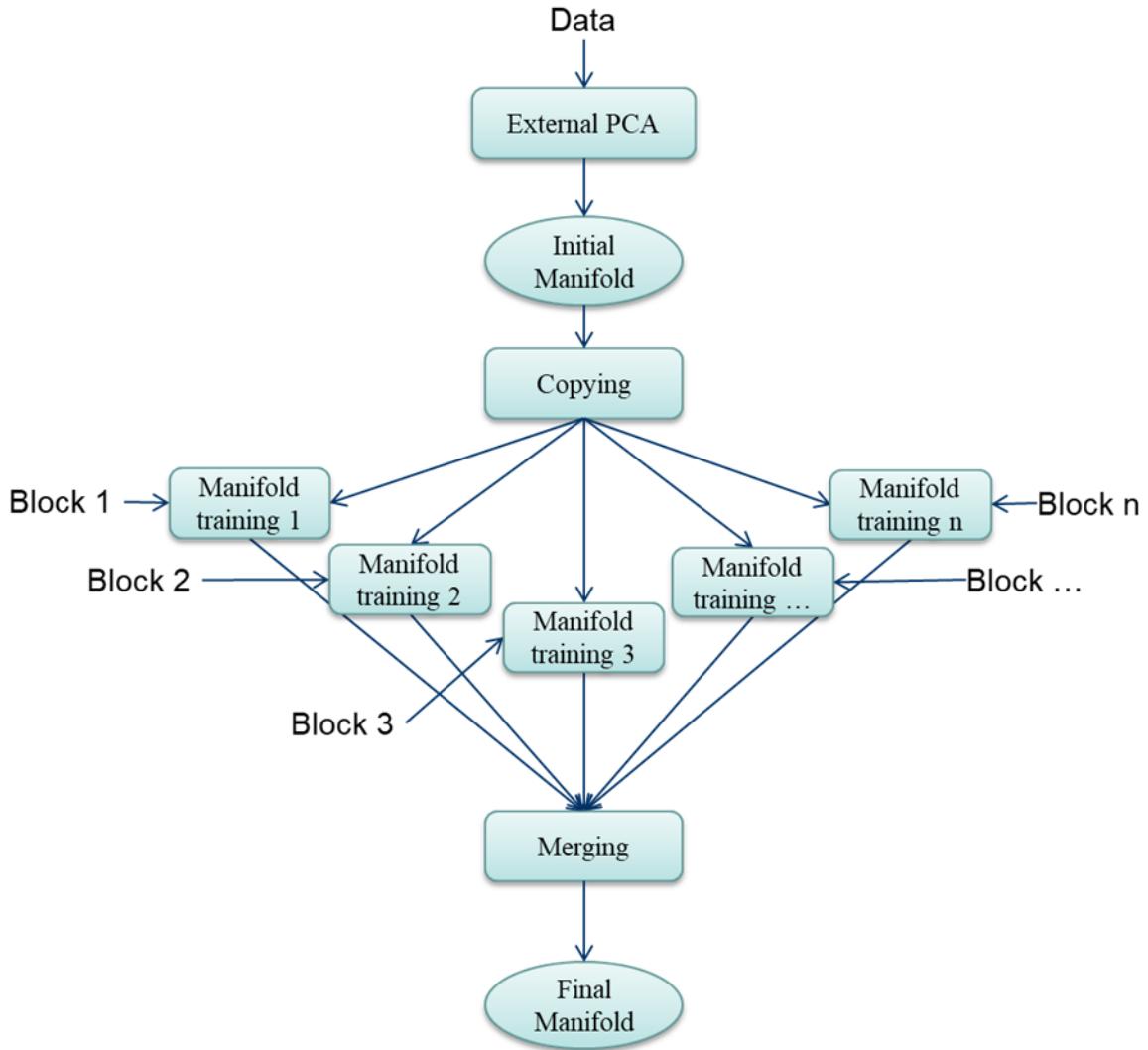
Figure 22. The scheme of the Parallel GTM.

The last step is to merge the produced intermediate GTM manifolds into the global one. For this purpose, simple averaging of **W** and β is used in this study. The output of the method is a single "final" manifold which potentially should cover the given data space.

4.6.2 Data

In this project, ChEMBL database of version 23 was used to perform the benchmarking study [93]. The structures were standardized: removed explicit hydrogens, aromatized using the basic rule, some functional groups were transformed (e.g. nitro group), etc. The ISIDA descriptors that were used to train the first universal GTM in [58] were computed: sequences of 2 and 3 atoms, labeled by their CVFF [94] force field types and

formal charge flag using all paths (IA-FF-FC-AP-2-3) [80, 95]. The descriptors were standardized (centered and divided by its standard deviation) and filtered by its variance (987 out of 5,161 descriptors were kept; the threshold was 2% of the maximal standard deviation in the data set).

To cross-validate the maps, the mean Balanced Accuracy (BA) and the Area Under the Receiver Operating Characteristics Curve (ROC AUC) were used as metrics. The labels "active/inactive" were assigned accordingly to the procedure described in the previous studies [10, 58].

4.6.3 Benchmarking Strategy

The benchmarking study was split into two parts. First, the GTM approaches (classical, incremental and parallel) were compared in terms of execution time and predictive performance (BA) where maps were trained on a target-specific set of compounds (CHEMBL204, Thrombin) with and without "decoys" (100K random compounds with unknown activity). To train the manifold, the GTM parameters corresponding to the first universal GTM described in [10, 58] were used: 41*41 nodes, 23*23 RBFs, regularization coefficient is 1.122018, RBF width is 1.1. To validate the map, a 3-fold cross-validation procedure was run where the number of actives and inactives was controlled (463 actives and 1440 inactives per fold; decoys were not taken for cross-validation). As an additional option, two blocks' sizes were tried: 500 and 1000 compounds. The number of blocks treated in parallel was limited to 14 due to the occupancy of a machine used to run the benchmarking tests.

The second part was devoted to algorithms comparison using Frame Sets (FS) of different sizes: 1K, 5K, 10K, 20k, 30K, 50K, 100K, 200K, 400K, 750K, 1M, 1.7M (entire ChEMBL) compounds. The FSs were gathered controlling the diversity for the compounds using pairwise Soergel distance (1-Tanimoto). The algorithm to collect the compounds was the following: the first compound was selected randomly, and the next compounds were compared to the ones that were already selected. A compound was added to the FS in case

if the minimal Soergel distance among all pairwise comparisons between the compound and others from the FS was larger than a threshold (e.g. 0.95). If the loop finished but the required number of items in the FS was not reached yet, the threshold was decreased (e.g. down to 0.9), and the loop started again. Thus, each FS possessed its own value of dissimilarity. The corresponding minimal pairwise Soergel distances are shown in Table 2.

Table 2. Minimal pairwise Soergel distance corresponding to different Frame Sets.

Frame set size, compounds	Corresponding minimal pairwise Soergel distance (1-Tanimoto) within the FS
1K	0.8
5K	0.7
10K	0.7
20K	0.65
30K	0.6
50K	0.55
100K	0.45
200K	0.4

In the second part, the maps were also compared in terms of data coverage (percentage of compounds passed the log-likelihood threshold), normalized Shannon entropy [5] characterizing the distribution of the compounds over the latent space, number of targets with mean BA ≥ 0.7 and number of targets with mean ROC AUC ≥ 0.7. The protocol used in this work to compute the likelihood threshold is described in chapter 4.2. To cross-validate the maps, more than 600 ChEMBL targets were used.

4.6.4   Results and Discussion

First, the GTM was trained on 5,710 ChEMBL compounds using a target-specific series of compounds with known activities against the Thrombin target (CHEMBL204).

The obtained maps were cross-validated. The results are shown in Table 3. One can see that the classical algorithm produces a better model (the mean BA is 0.73) since no approximations were done. In this context, the incremental and parallel algorithms produce models with comparable predictive performance (BA=0.7±0.015).

Table 3. Benchmarking results using "Thrombin" data set (5,710 compounds).

Description	Block size	Balanced Accuracy				Time, h:m [1]
		Fold 1	Fold 2	Fold 3	Mean	
Classical GTM	-	0.74	0.73	0.73	0.73	3:07
Incremental GTM	500	0.7	0.69	0.69	0.69	2:28
	1000	0.69	0.72	0.69	0.7	0:33
Parallel GTM	500	0.70	0.69	0.68	0.69	0:41
	1000	0.71	0.72	0.72	0.72	0:43

[1] Approximate execution time recorded during manifold training.
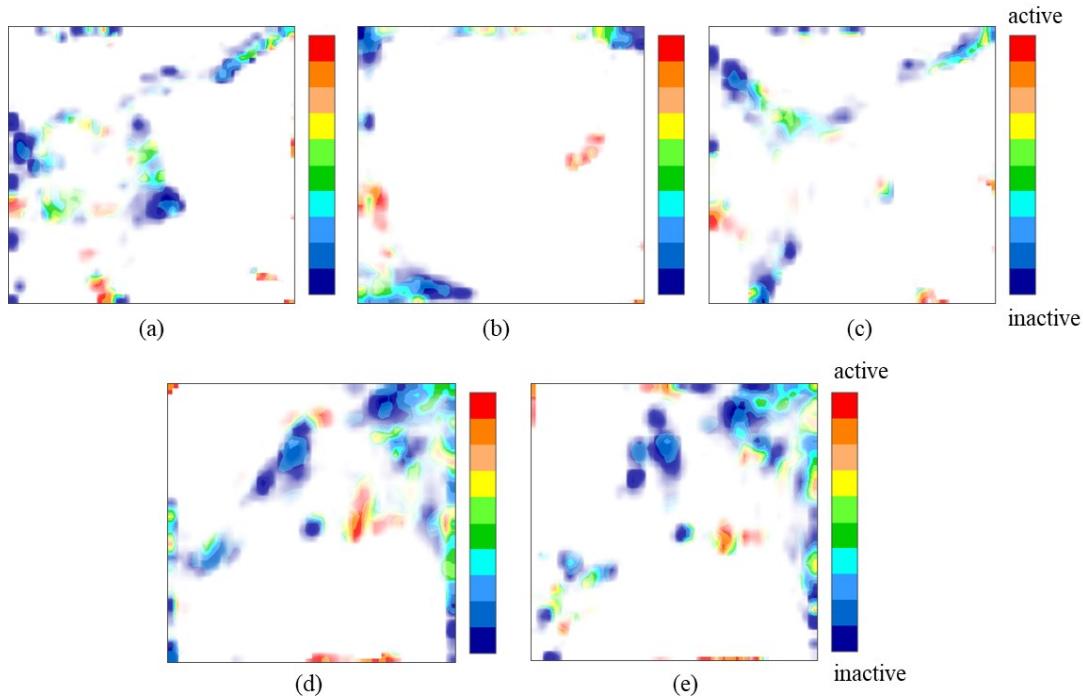
Figure 23. The fuzzy class landscapes for the "Thrombin" data set of 5,710 compounds: (a) the classical GTM, the incremental GTM with blocks of (b) 500 and (c) 1,000 items, and the parallel GTM with blocks of (d) 500 and (e) 1,000 items. Here, the transparency corresponds to the density.

The models trained by the incremental GTM with blocks of 500 and 1000 compounds do not differ significantly in terms of BA as well as the ones trained with the parallel approach. The GTM class landscapes were built and visualized (Figure 23). One can see that the incremental algorithm visualizes the data space differently for the different block sizes, whereas the parallel GTM returns the same landscape for both sizes. A comparison of the likelihood distribution (Figure 24) shows that PGTM covers the data as well as the classical algorithm. In contrast, the incremental algorithm has worse data coverage which can be seen in the GTM landscape (Figure 23b and Figure 23c).
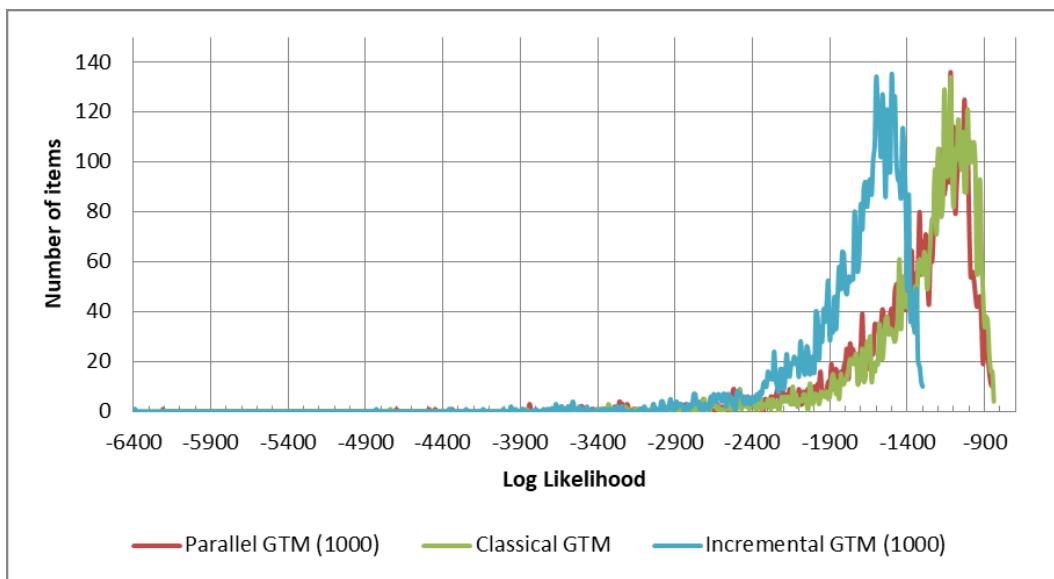
Figure 24. Log-likelihood distribution for the compounds from "Thrombin" data set produced by the classical (green), incremental (blue), and parallel (red) GTMs.

Next, the methods were tested on the larger data set where 100K random "decoys" (ChEMBL compounds with unknown activity) were added. The maps were rebuilt on 105,710 structures. The results of the cross-validation are given in Table 4.

In comparison with the first experiment, the acceleration of GTM by the parallel algorithm now is more significant. The parallel algorithm trained the manifold 5 times faster than the incremental one keeping the same level of the predictive performance (BA=0.67±0.02). The likelihood distribution in Figure 25 demonstrates that the PGTM covers the data similar to the incremental GTM.

Table 4. Benchmarking results where "decoys" were added to the Thrombin data set.

Description	Block size	Balanced Accuracy				Time, h:m [1]
		Fold 1	Fold 2	Fold 3	Mean	
Incremental GTM	5000	0.65	0.65	0.64	0.65	23:57
	10000	0.67	0.67	0.68	0.67	28:52
Parallel GTM	5000	0.65	0.65	0.65	0.65	5:48
	10000	0.69	0.68	0.69	0.69	10:33

[1] Approximate execution time recorded during manifold training.
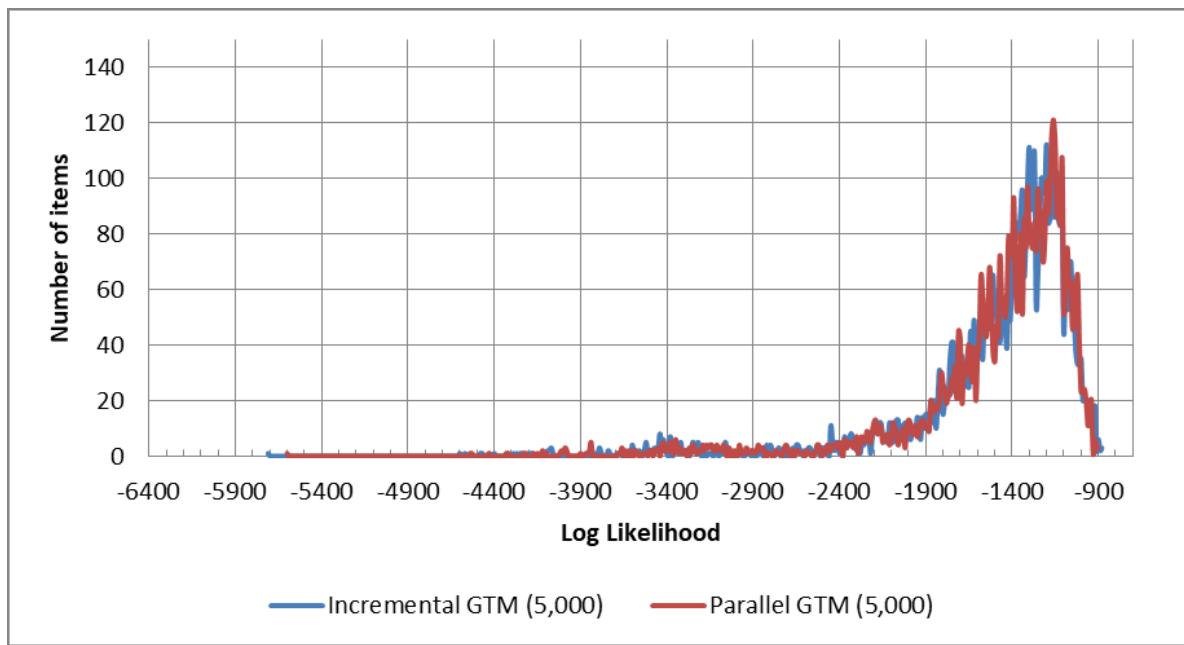
Figure 25. Log-likelihood distribution for the "Thrombin" data set with random 100K decoys produced by the incremental (blue), and parallel (red) GTMs.

Although parallel GTM algorithm leads to similar predictive performance and LLh distribution as incremental GTM, their manifold shapes, and, hence, the data distribution on the maps are pretty different (Figure 26).
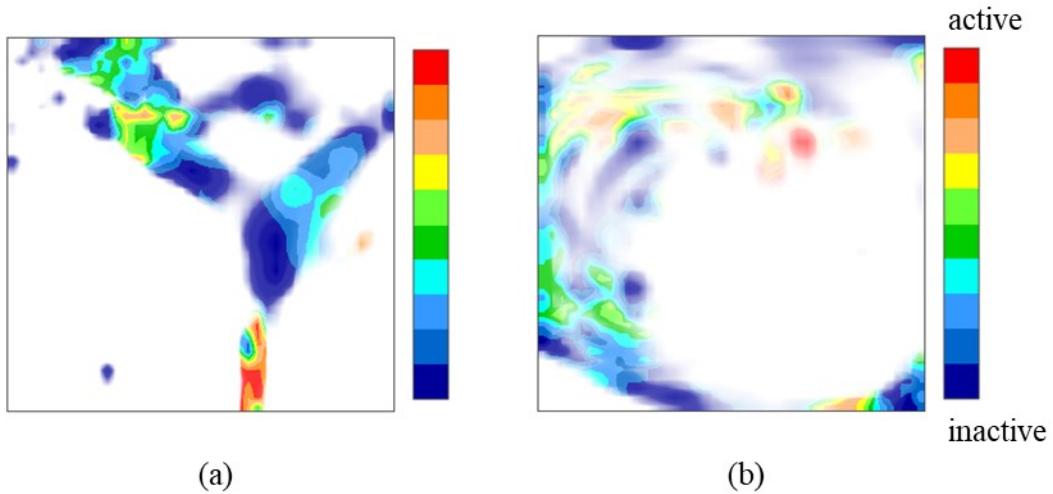
Figure 26. The fuzzy class landscapes where "Thrombin" data set of 5,710 compounds. Here, the manifold were trained by (a) incremental and (b) parallel GTM algorithms using "Thrombin" data set with random 100K decoys (105,710 compounds) as a Frame set.

Finally, the algorithms were compared in terms of mean BA, mean ROC AUC, data coverage, normalized Shannon entropy, number of targets with mean BA ≥ 0.7 and number of targets with mean ROC AUC ≥ 0.7 using frame sets of different sizes. The results are shown in Figure 27 and Figure 28.

One can see that a larger frame set leads to lower data coverage (Figure 27). This can be explained by the Applicability Domain (AD) which is wide in the case of general FS (1K compounds; the most diverse compounds are selected), and, in contrast, it becomes more narrow by adding similar compounds. In the latter case, the map focuses more on the dense groups of compounds which are presented in the FS by a larger number of items. Thus, GTM pays less attention to the chemical families represented by some items, or these families can be even ignored in the case of a huge FS (e.g. 200K). At the same time, the entropy and the predictive performance grow. It can be also seen that the FS of 5K compounds is already enough to describe ChEMBL23 containing 1.7M compounds, whereas it is not clear how big should be the FS in case of larger databases, such as PubChem (96M), Zinc (1.3B), and GDB-17 (166B).
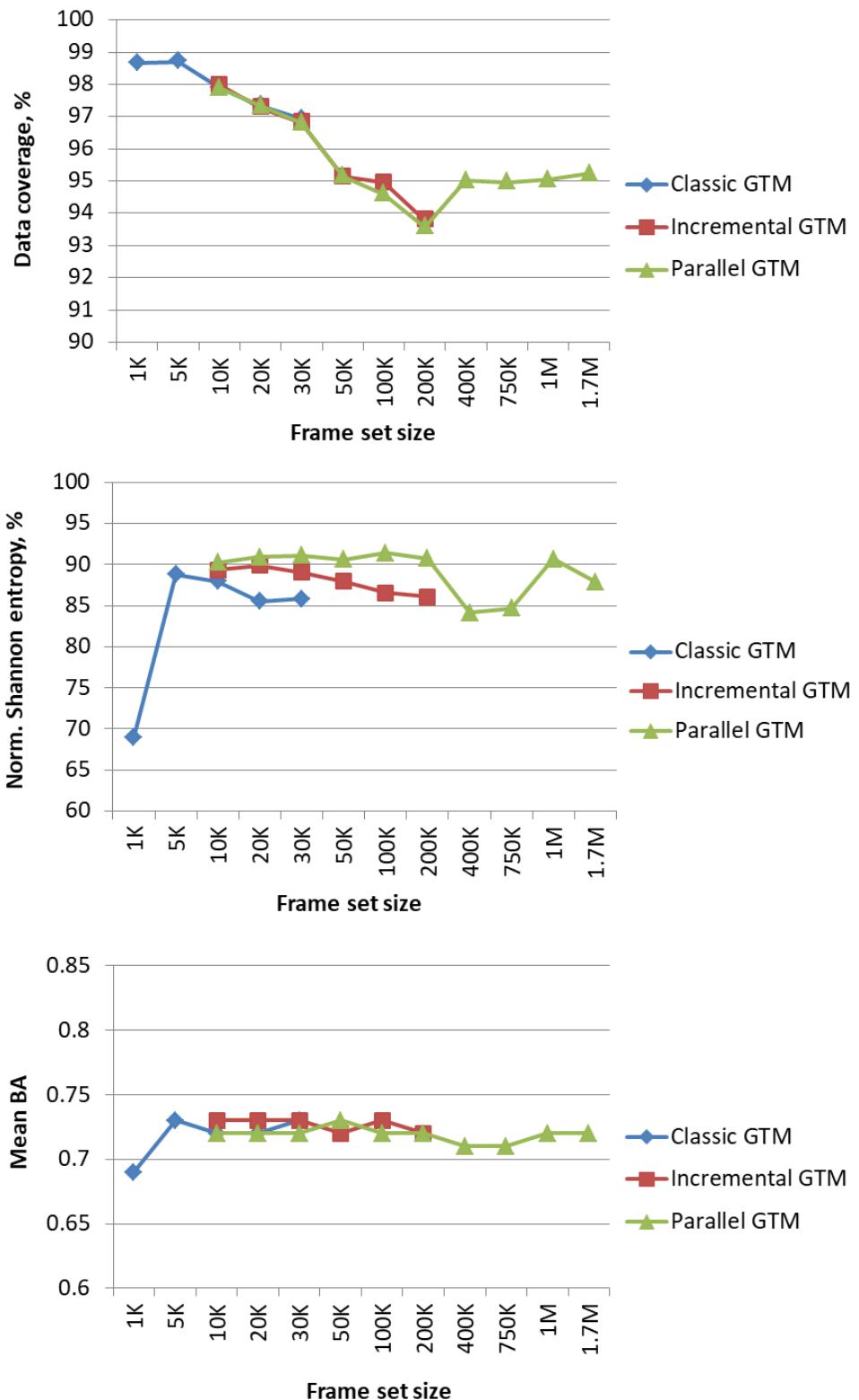
Figure 27. Data coverage, normalized Shannon entropy [5], and mean Balanced Accuracy (BA) computed for classical, incremental and parallel GTMs where frame sets of different sizes were used to train the manifolds.
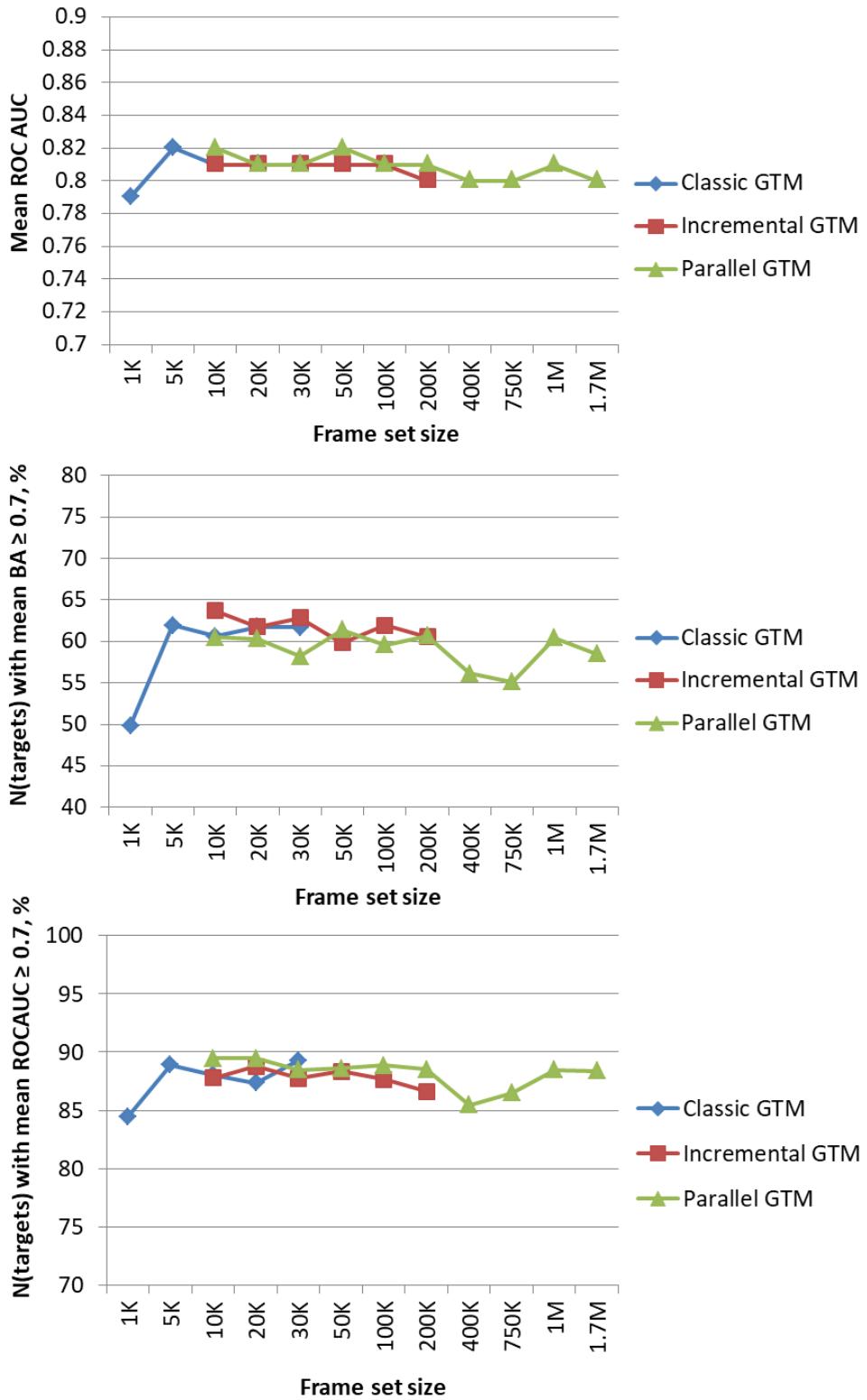
Figure 28. Mean ROC AUC, number of targets with mean BA ≥ 0.7 and number of targets with mean ROC AUC ≥ 0.7 computed for classical, incremental and parallel GTMs where frame sets of different sizes were used to train the manifolds.

Comparing the predictive performance of the GTM algorithms, it is shown that all of them possess the same level of BA and ROC AUC (Figure 28). However, PGTM is much faster than Incremental GTM, and, therefore, it is able to treat larger FSs than both classical and incremental algorithms.

4.7 Conclusion

GTM is an efficient tool applied in different contexts. However, some methodological developments were needed to adopt the method to the Big Data case. First, the impact of different preprocessing schemes was checked using the SRC data set (tyrosine kinase inhibitors). The strategies of descriptors preparation were compared in terms of Balanced Accuracy (BA) and Area under the Receiver Operating Characteristics Curve (ROC AUC). It was demonstrated that the highest predictive performance is achieved by descriptors standardization (centering and division by its standard deviation).

Some applicability domain (AD) concepts have already been proposed for GTM (chapter 3.3), and their drawbacks have been discussed here. For instance, the predominant class AD needs the CPF value to ignore the mixed nodes which, in turn, decreases the density of the landscape. As an alternative, a new approach to compute the log-likelihood cutoff was proposed and applied in this work.

To solve the problem of the map resolution and the problem of the mixed zones, a hierarchical GTM zooming approach was automatized. Two strategies for zones generation were implemented. The developed tool was coupled with a new Maximum Common Substructure (MCS) extraction protocol proposed for zone-specific substructures search. The tool was applied in the project of chemical library enrichment which was done in cooperation with Boehringer Ingelheim company (the results are described in chapter 7).

Finally, the idea of Constrained Screening (CS) and Parallel GTM approaches were presented. As it was described, CS allows screening the database querying not a single activity/property but a desirable profile. The returned compounds possess the satisfaction score which can be used to rank the structures and to select the hits.

Parallel GTM allows training the GTM manifold with larger data sets. It initializes the manifold using the incremental PCA and then trains it on a series of blocks in parallel. The method was compared to the incremental approach in terms of speed of calculations and predictive performance (BA). It was established that Parallel GTM trains the manifold 5-6 times faster producing the models with the same BA.

Implementation of Parallel GTM allowed us to perform a comparison of the predictive performance of classification models as a function of a Frame set size. It has been demonstrated that the FS of 5,000 structures is sufficient to prepare a GTM for the entire ChEMBL23 database containing more than 1.7M compounds.

5 GTM as a Tool for Virtual Screening

Virtual Screening (VS) is a common technique in drug discovery used in different projects [96–98]. Its goal is to select potential hits from the chemical database using knowledge retrieved from the existing data. Usually, the so-called VS funnel has several layers differentiating in terms of accuracy. Thus, the methods with lower accuracy (e.g. similarity filters) but higher speed stand at the beginning and the more accurate methods (e.g., docking) are run at the end since they are restricted in terms of compounds that these methods can handle.

In this chapter, we discuss the application of GTM to virtual screening. The first part of the chapter describes the benchmarking results done for single-target and multi-target VS on public data. Next, the obtained knowledge was applied to industrial data to test the GTM in the industrial drug discovery process.

5.1   Multi-Target Virtual Screening

5.1.1 Introduction

GTM is a data visualization and analysis tool which can successfully be used to train classification and regression models. The benchmarking studies done so far show that GTM provides similar predictive performance to other machine-learning methods (SVM, Random Forest, Neural Networks) [6]. This makes GTM attractive to be used in virtual screening (VS) campaigns.

The predecessor of GTM – Self-Organizing Maps (*SOM*) – was already tested as a VS technique in several studies [84–86]. For instance, it was used to identify several purinergic receptor agonists [86]. Later, SOM was compared to the similarity search with data fusion, and, despite the poor predictive performance, in principle, SOM can be used as a tool for the VS tasks [84]. Since GTM may perfectly mimic SOMs – by narrowing RBF width to ensure that item responsibility focuses 100% on the nearest manifold grid point – but also can outperform it by applying fuzzy logics, GTM is a better VS tool than SOM.

GTM has never been applied to multi-target virtual screening (virtual profiling) where a model is used to select the compounds in terms of several biological activities. This can be achieved on the hand of universal GTMs, a concept introduced by P. Sidorov et al. [9]. Herein, a manifold is optimized not for one single, but with respect to the largest possible panel of target-specific series of compounds (ChEMBL database of v.20 in reference [9]). The obtained map is used then to make predictions for an extended pool of activities/properties (including ones not used for manifold optimization but seen to be properly supported by the manifold nevertheless).

In this project, GTM was tested as a single-target and multi-target virtual screening technique. Its predictive performance was compared to two popular single-target approaches: Random Forest and Neural Network. As a baseline, the similarity search with data fusion was used. The results were published in our article in the *Journal of Computer-Aided Molecular Design* [10] (see below).

Multi-task generative topographic mapping in virtual screening

Arkadii Lin[1,2] · Dragos Horvath[1] · Gilles Marcou[1] · Bernd Beck[2] · Alexandre Varnek[1]

Abstract

The previously reported procedure to generate "universal" Generative Topographic Maps (GTMs) of the drug-like chemical space is in practice a multi-task learning process, in which both operational GTM parameters (example: map grid size) and hyperparameters (key example: the molecular descriptor space to be used) are being chosen by an evolutionary process in order to fit/select "universal" GTM manifolds. After selection (a one-time task aimed at optimizing the compromise in terms of neighborhood behavior compliance, over a large pool of various biological targets), for any further use the manifolds are ready to provide "fit-free" predictive models. Using any structure–activity set—irrespectively whether the associated target served at map fitting stage or not—the generation or "coloring" a property landscape enables predicting the property for any external molecule, with zero additional fitable parameters involved. While previous works have signaled the excellent behavior of such models in aggressive three-fold cross-validation assessments of their predictive power, the present work wished to explore their behavior in Virtual Screening (VS), here simulated on hand of external DUD ligand and decoy series that are fully disjoint from the ChEMBL-extracted landscape coloring sets. Beyond the rather robust results of the universal GTM manifolds in this challenge, it could be shown that the descriptor spaces selected by the evolutionary multi-task learner were intrinsically able to serve as an excellent support for many other VS procedures, starting from parameter-free similarity searching, to local (target-specific) GTM models, to parameter-rich, nonlinear Random Forest and Neural Network approaches.

Abbreviations

GTM	Generative topographic mapping
UGTM	Universal generative topographic mapping
GA	Genetic algorithm
CV	Cross-validation
DUD	Directory of Useful Decoys
NN	Neural network
RF	Random forest

✉ Alexandre Varnek
varnek@unistra.fr

1 Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal Str., 67081 Strasbourg, France

2 Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany

Introduction

Generative Topographic Mapping (GTM) [1] is a dimensionality reduction method corresponding to a probabilistic extension of Self-Organizing Maps (SOM) [2]. In order to project the data onto a 2D latent space, the method injects a 2D hyperplane, called manifold, into the descriptor space, in which each item of the "Frame Set" (FS) spanning this space corresponds to a point defined by its high-dimensional descriptor vector. The manifold is mathematically described by a square grid of reference points (nodes) and a set of Radial Basis Functions (RBF, Gaussian functions). The FS items serve to "bend" the manifold in order to make it visit a maximum of their descriptor space positions. Using a gradient descent, the method tries to fit positions of the RBF centers, in order to maximize Gaussian function levels at all the FS data points. In other words, it tries to fit the data maximizing a LogLikelihood (LLh) value, which is a logarithm of a cumulated probability of a compound to be related to each node of the manifold [3]. When the manifold

is built, each compound is characterized by its LLh value and is described by the vector of its probabilities to "reside" in each node. This vector, R_{nk}, representing the probability of compound k to reside in node n is called the responsibility vector. Since any compound is certain to reside somewhere on the map, $\sum_n R_{nk} = 1, \forall k$. A library of several compounds can be described by the vector of cumulated responsibilities CR of its members k, $CR_n = \sum_k R_{nk}$. Given compounds of known property or bioactivity values, an activity/property Landscape can be created and visualized. This is useful not only for data visualization and analysis but also as a QSAR/QSPR model. After projecting a new compound on it, the class/property value can be easily predicted from the landscape.

Initially, GTM was tested as a tool for Quantitative Structure–Activity Relation (QSAR) tasks on typical structure–property sets [4, 5], where the known actives and inactives of the set were used both as FS and as property set for coloring of the herewith fitted manifold. From this perspective, the initial descriptor space yielding the top predictive manifold could be freely tuned, together with the manifold parameters (number of nodes, number of Gaussians, Gaussian width and Regularization term). The resulting GTM thus represents a predictive model fully dedicated to a specific QSPR problem, and exclusively trained on specific QSPR data. It is the results of a typical single-task learning process, like many other in Ligand-Based Virtual Screening: Decision Trees, Artificial Neural Networks (ANN), Support Vector Machine, Similarity search on binary fingerprints, etc. [6, 7] In addition to this list, SOM method was also tried as a VS technique in many studies [8–10]. For instance, it was used to identify several purinergic receptor agonists [10]. Later, SOM was compared with a Similarity search with data fusion, and, despite a poor predictive performance, the results of such comparison show that in principle SOM can be used as a tool for the VS tasks [8].

However, GTM was also tested successfully as a tool for large public chemical database (PubChem-17, ChEMBL-17 and FDB-17) visualization and analysis [3]. In 2015, Sidorov et al. [11] used GTM in order to create a compound set-independent "universal" map of Chemical Space (CS). The manifold and its underlying descriptor space were not selected with respect to any peculiar property but were aimed at representing the best possible consensus, ensuring a broad "polypharmacological competence", i.e. ability to host predictive property landscapes for a maximum of diverse properties. Conceptually, this is a form of Multi-Task Learning (MTL): based on a generic FS randomly picked to cover the entire ChEMBL CS, structure–activity data from about 100 unrelated target-specific series of ligands of known pK_i values were used to challenge each manifold in terms of its ability to "host" predictive activity landscapes for each of these series. Selection with respect to the mean predictive

performance over all series produced not an optimal manifold dedicated to a given QSPR problem, but a best-compromise manifold of optimal robustness and ability to host any arbitrary property landscape, all while maintaining a certain predictivity level. This ability was eventually validated in showing that it can easily distinguish active from inactive compounds for more than 400 ChEMBL targets (others than the ~ 100 used for selection). Results report an Balanced Accuracy (BA) higher than 0.6 for all the targets (none of which served for map parameter selection).

The above approach is thus related to MTL [12, 13], consisting in learning the choices (descriptors, GTM grid size, etc.) leading to a "consensual" manifold, i.e. learning the choices that are generally relevant to QSPR in drug design, all targets confounded.

MTL is a wide-spread strategy in chemoinformatics and is embodied by numerous distinct approaches from the use of calculated properties by a previously fitted model as input descriptor to a higher-order model (feature nets [14], FN), to multiple-output multilayered ANNs [13] to strategies in which both ligands and targets are descriptor-encoded (computational chemogenomics [15–19]). Conceptually, the "universal" map approach is different from all the above and is closest related to the multiple-output multilayered ANNs. Manifold building conceptually matches the fitting of parameters of the common layers of the ANN, crystallizing the knowledge of the common features that are important to all the learning tasks. Landscape creation by coloring with specific data sets, followed by prediction, matches the task-specific output neurons of the ANNs—with the notable difference that the latter may still be fine-tuned to improve task-specific predictability. By contrast, at given manifold, coloring of a landscape by projection of a property set and thereupon-based prediction is deterministic and parameter-free. Thus, there is no perfect analogy between the "universal" GTM style of MTL and above-mentioned classical MTL methods. Unlike chemogenomics approaches, "universal" manifolds do not require at all any injection of information about the considered targets, which can be of arbitrary diversity. While chemogenomics focusses on groups of related activities (i.e. for biologically related targets) "universal" manifolds were successfully hosting landscapes for completely unrelated chemical and biological properties, ranging from target-specific activities to cell- or organism-based screen results. Learning features that are "universally" important in structure–activity relationships ensures, on one hand, the generality of "universal" GTMs (UGTMs). On the other, generality will unsurprisingly result in lesser predictive propensity for some targets, as the inductive transfer of knowledge operating at manifold construction step basically resumes to a generic ability to span drug-relevant CS.

So far, no comparison of GTMs and—in particular—of UGTM to other VS methods was undertaken. In order

to evaluate the quantitative benefits of building "universal" manifolds, their performance in VS was compared to—firstly—single-task "local" GTMs, dedicated to each biological properties, and also to state-of-art single-task machine learning methods, namely Similarity search and Similarity search with data fusion, Neural Networks (NN), and Random Forest (RF).

Methods

Data

For this project two public databases are used: ChEMBL (version 23) [20] and Directory of Useful Decoys (DUD) [21]. To extract the data, the previously described [11] target-specific structure–activity series extraction protocol has been reenacted on the later release 23 of the ChEMBL database. A total of 618 human single proteins were retained, after "categorization" of ChEMBL-reported activity scores into "actives" and "inactives", respectively. To this purpose, a set of activity classification rules embodied in scripts (available in Supplementary Material of the cited paper) were applied. Compounds with reported percentage of inhibition were considered inactive if values were below 50%, otherwise they were ignored. If dose–response activity measures were available, various cutoffs ranging from low nanomolar to micromolar range were tried out. Compounds better than the threshold were labeled "active" (a minimum of 15 required), the ones of activity weaker that the ten-fold threshold value were "inactives" (at minimum 50), with in-between molecules being ignored (in order to facilitate the separation problem). The actual target-specific cutoff eventually retained was the one ensuring a reasonable balance, closest to one active (or more) for four inactives (but never exceeding parity one active: one inactive—series having, at all considered cutoffs, more reported actives than inactives were discarded). Files (labeled Target-ChEMBLID.smi_ID_class) reporting, for each target, the standardized SMILES string, compound ChEMBL ID and assigned class are now provided as Supplementary Material for the nine targets of the VS simulation, together with their corresponding DUD files. Equivalent data for the remaining 609 targets used in internal validation are available upon request.

Next, DUD data were used to extract independent, external compound series, by focusing on the subset of ChEMBL targets that are also present in DUD and pruning all DUD compounds already encountered in the ChEMBL series. This often meant elimination of virtually all the actives from the DUD series, thus failure to obtain an external data set. However, in nine cases (Table 1) the DUD target-specific series contained sufficiently numerous original actives and

Table 1 A list of nine DUD targets taken for the external validation

Target ID	Target name
CHEMBL1827	Phosphodiesterase 5A
CHEMBL1952	Thymidylate synthase
CHEMBL251	Adenosine A2a receptor
CHEMBL260	MAP kinase p38 alpha
CHEMBL279	Vascular endothelial growth factor receptor 2
CHEMBL301	Cyclin-dependent kinase 2
CHEMBL4282	Serine/threonine-protein kinase AKT
CHEMBL4338	Purine nucleoside phosphorylase
CHEMBL4439	TGF-beta receptor type I

were retained for external validation of ChEMBL-trained models (Table 2).

Structure standardization, assignment of activity classes (active vs. inactive) for structures associated to human targets, and rejection of targets with too small or too imbalanced structure–activity series were employed as already described. DUD compounds were likewise standardized, and their given activity class labels (active vs. inactive = decoy) were adopted as such. At the end, 1.5 M unique ChEMBL compounds and 914K DUD molecules were kept after curation.

Molecular descriptors

One hundred different fragmentation schemes supported by the ISIDA Fragmentor software, [22, 23] and gathered according to the experience of previous works [3, 11] were used as a starting pool for the search of suitable descriptor space. Recall that descriptor space selection is a key meta-parameter of the evolutionary map sampling tool.

Universal (multi-task) GTM manifolds

For technical reasons (the release of a major, faster version of the GTM software), the already published "universal" map selection protocol has been rerun, with another important change with respect to the previously published version; the use of structure–activity class series as selection sets instead of the originally employed (less data-rich) structure-pK_i (continuous) affinity data. Out of the 618 ChEMBL structure–activity series, 236 were randomly designed as selection sets (see file "selection.targets" in the zipped dataset repository in Supplementary Material) for UGTM training (attached "external.targets" enumerates the remaining 382 targets not involved in selection). The FSs were constructed as sets of random ChEMBL samples of different sizes (between 8.5K and 26K compounds). Here, a

Genetic Algorithm [24] was used to optimize GTM parameters, such as the number of nodes, the number of Gaussian functions (RBF), the regularization coefficient and the width of an RBF. In addition to the best descriptors set and the best GTM parameters, GA also has chosen the most suitable descriptors normalization scheme. At a given GTM parameter set, the manifold training procedure is run in incremental mode [25]. The size of each block was 10,000 compounds. Then, for each selection set, a threefold cross-validation of the current manifold was performed, where landscapes are iteratively built based only on 2/3 of the ChEMBL set, while the remaining tier will be projected into the landscape and ranked by a probability to be active, representing the "color" (relative population of actives vs. inactives) in their target area. For technical details about the rigorous formalism to construct and predict with class and activity landscapes, please refer to our previous GTM publications. According to this selection criterion of mean threefold cross-validated BA of prediction, four best universal maps, each based on a different descriptor space, with the mean BA ranging within 0.7–0.75 have been selected (Table 3). Corresponding GTM parameters and FS sizes are presented in Table 4.

Table 2 The datasets used for the screening procedure

Target ID	DUD data sets			ChEMBL data sets			Thresholds[a]
	Actives	Inactives	Total	Actives	Inactives	Total	K_i/IC/EC$_{50}$ (nM)
CHEMBL1827	170	25,334	25,504	691	824	1515	50
CHEMBL1952	63	6113	6176	124	455	579	1000
CHEMBL251	79	28,001	28,080	1303	3618	4921	100
CHEMBL260	100	32,925	33,025	1453	2567	4020	100
CHEMBL279	94	22,595	22,689	2047	4663	6710	100
CHEMBL301	189	25,675	25,864	638	2305	2943	500
CHEMBL4282	52	14,228	14,280	725	2619	3344	500
CHEMBL4338	102	6334	6436	100	111	211	50
CHEMBL4439	82	8013	8095	282	385	667	50

[a]Compounds with dose–response affinity value below or equal to threshold (in nM) are considered active, while those with values exceeding the 10-fold threshold value are inactives. At intermediate activities, compounds are discarded from the ChEMBL set. Note that the DUD definition of "actives" does not comply to the same rules—they routinely include co-crystallized ligands, irrespective of their affinities

Table 3 The best selected descriptors sets [22]

Map	Abbreviation	Definition	Descriptor set size
1	IA-FF-FC-AP-2-3	Sequences of atoms with a length of two to three atoms labeled by force field type and formal charge flag, using all paths	987
2	IIRAB-FF-1-2	Atom-centered fragments of restricted atom and bonds of a length one to two atoms labeled by force field types	1029
3	IAB-PH-FC-AP-2-4	Sequences of atoms and bonds of a length two to four atoms labeled by pharmacophoric atom types and formal charges using all paths	779
4	IA-2-7	Sequences of atoms of a length two to seven atoms	728

Table 4 Selected GTM meta-parameters for the four best chromosomes chosen by the genetic algorithm [24]

Map	FS size	Number of nodes per line	Number of RBF per line	Regularization coefficient	RBF width	Normalization scheme[a]
1	17,000	41	23	1.122	1.1	2
2	17,000	47	29	0.018	1.6	1
3	25,500	37	19	0.017	2.1	2
4	25,500	38	19	3.55	1.9	2

[a]The standardization schemes: 1—centering on the mean value; 2—Z-normalization (centering on the mean value and division by the standard deviation)

Monitored success scores

In this benchmarking study, the mean area under the Receiver Operating Characteristic (ROC AUC) when predicting half of the compound series based on landscapes colored (or models learned, for other methods—*vide infra*) on the other half is used in the internal validation procedure. This further on named $<AUC>_{1/2}$ criterion will be consistently used to compare models (except for single-query similarity searching, where it cannot be defined—see following subsection). The mean is taken over ten independent repeats of the above procedures, where splitting into training and kept-out compounds is fully randomized. No specific care is taken to ensure that each compound is strictly kept out once and only once per iteration.

Internal validation results were alternatively depicted as density distribution plots of the ROC AUC values over the training subsets (Figs. 1, 2, *vide infra*). For each method each ChEMBL target-specific set returns the ten distinct ROC AUC values from the randomized internal validation experiments described in the "Methods" section. Plotting the density (number of targets) in counting each target 10 times,

into the specific bins matching each of its ROC AUC values achieved on the random splits (and followed by a normalization of the density to compensate for multiple counts)—would however produce one "global" histogram, with no information on the expected fluctuation of density bar heights. Estimating those error bars is however of paramount importance, in order to ensure that the histogram shape is not an artefact of the peculiar randomized choice of training/test splits. For this specific purpose, this work proceeds to first generate "splitting accident-prone" histograms, considering each target-specific compound set to be represented by one randomly picked ROC AUC out of the 10. Depending on the pick, the set will be counted in a lower or higher bin, i.e. its localization on the X axis will reflect the intrinsic uncertainly induced by the train/test splitting. Every set is counted exactly once—only its X-axis bin may fluctuate. Therefore, every such "splitting accident-prone" histogram will differ in shape. One thousand of these are generated, which allows a thorough monitoring of the expected fluctuation of bar heights as a consequence of splitting artefacts. Eventually, the plot shows the mean bar heights (which converge to the above-mentioned "global" histogram) with associated error bars (if readable—occasionally, fluctuations are too small).
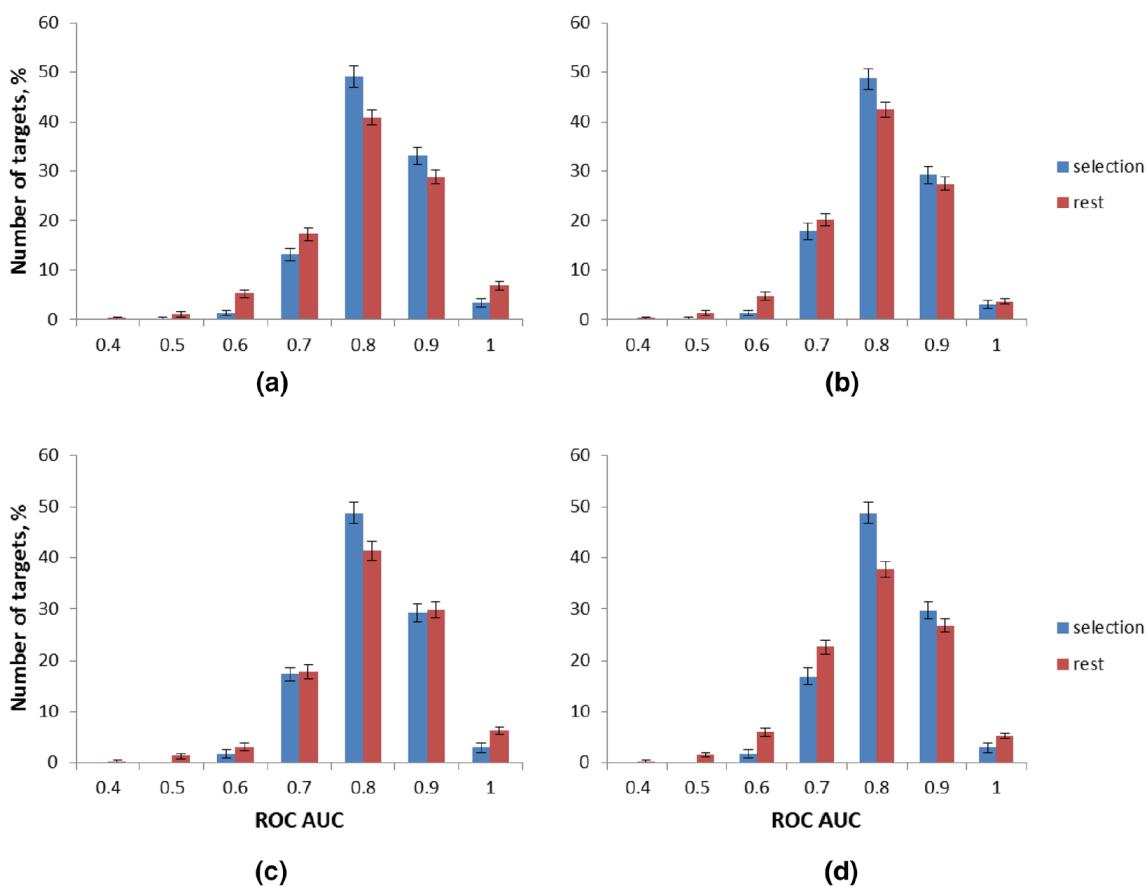
Fig. 1 ROC AUC values for the selection set and rest targets: **a** map 1, **b** map 2, **c** map 3, **d** map 4
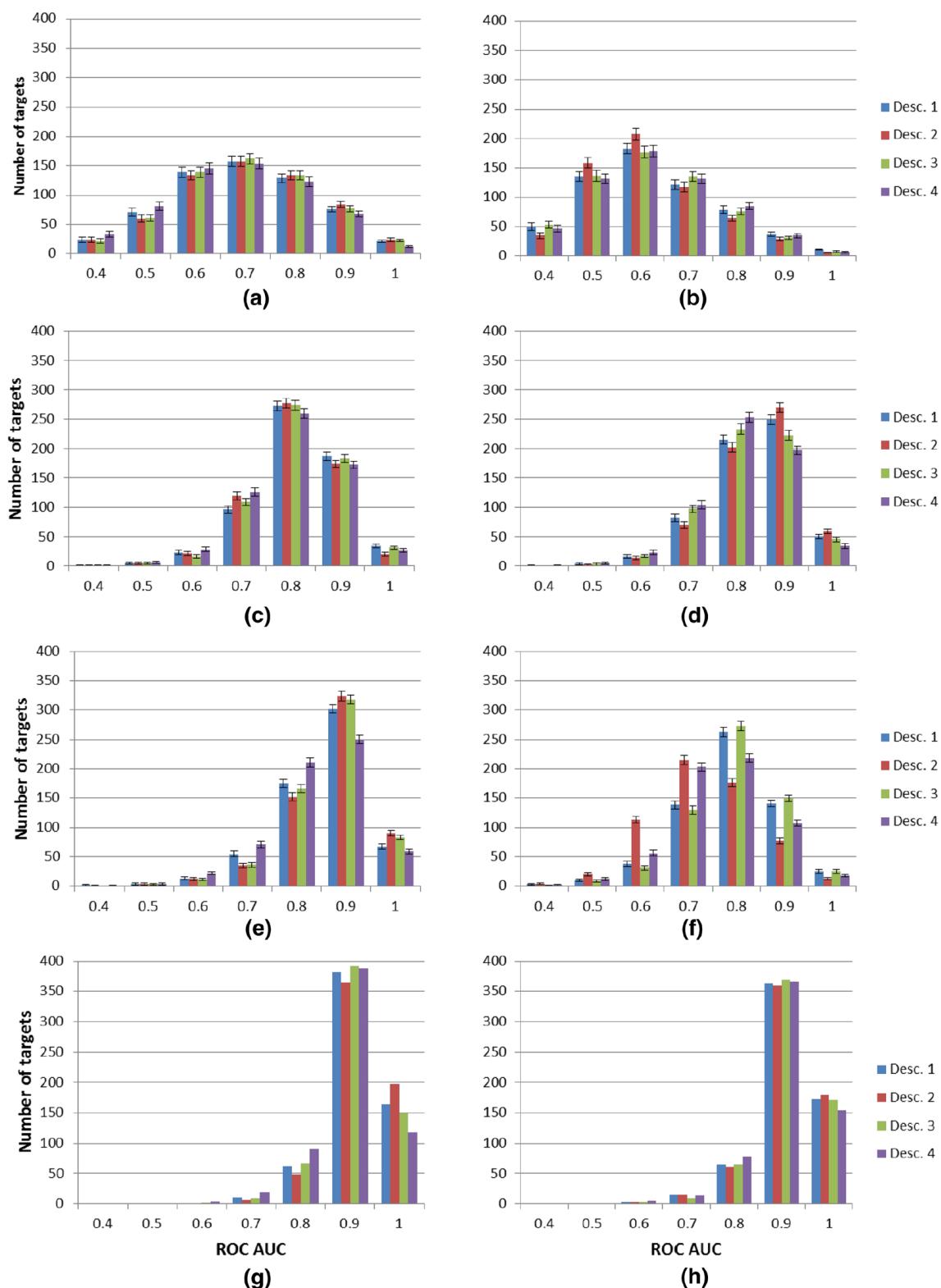
Fig. 2 Internal validation results on 618 ChEMBL targets: single-query Similarity search in **a** descriptors and **b** latent spaces, **c** UGTM, **d** local GTM, Similarity search with data fusion in **e** descriptors and **f** latent spaces, **g** NN, and **h** RF. Here, Desc. 1–4 correspond to the descriptors sets shown in the Table 3

In actual virtual screening, the DUD series is projected onto the "complete" landscape generated from the entire ChEMBL set. To estimate the predictive performance of a particular map, ROC AUC (further on referred to as VSAUC) is computed, after ranking DUD compounds as above-mentioned [26].

Benchmarked models

For each of the 618 targets, single-task (local) models were set up in each out of four descriptors spaces chosen in Table 3 using the following methods:

- Regular (local) GTM
- Similarity search
- Similarity search with data fusion
- RF
- NN

Depending on the nature of the model, setting it up requires distinct protocols, involving parameter selection or fitting (local GTM, PF, NN) or decisions on used similarity scoring, etc. These aspects will be detailed in the dedicated paragraphs below, while the same success score monitoring procedure outlined above was applied to all models. The descriptors normalization scheme was not changed and corresponds to the one that is shown in Table 4.

The parameters of local GTM were not optimized, but were taken by default: the number of nodes is 625 (25×25), the number of Gaussian functions is 144 (12×12), the width of a Gaussian function is 2.82, the regularization coefficient is 1.0. To perform the experiments with NN and RF, SciKit Learn implementations of Multi-Layer Perceptron (MLP) (https://scikit-learn.org/stable/modules/neural_networks_supervised.html) and RandomForestClassifier (https://scikit-learn.org/stable/modules/ensemble.html#forest) were employed [26–29]. Here, the MLP parameters are taken by default: the number of hidden layers is 1, the number of the nodes in a layer is 100, the rectified linear unit function (relu) is used as an activation function [30], and the "adam" solver is used for the weights optimization [31]. Backpropagation approach is applied to train the net [26–28]. In case of RF, an ensemble of trees is built on a random half of compounds where the original ratio actives/inactives is kept. All the parameters are taken by the default, mentioned in SciKit Learn (https://scikit-learn.org/stable/modules/ensemble.html#forest), where the number of trees in a forest is 10.

As a gold standard for the VS tasks, Similarity search and Similarity search with data fusion were chosen. Both these methods are based on a simple similarity principle: similar compounds should share similar activity. Therefore, the idea of similarity searching is to find compounds out of a screening pool which are similar to the reference point with a known label (i.e. active). While there are better suited criteria [32, 33] to specifically monitor neighborhood behavior compliance, herein the generally applicable ROC AUC criterion is used to score the potential predictive performance of the method, after ranking candidates in decreasing similarity order (Tanimoto scores) to the used query. Also, as an alternative to a simple similarity searching, similarity searching with data fusion is taken. Within this approach the screening pool is compared not to one but to N reference compounds (in this project the pool of reference compounds was chosen to embody a randomly picked 50% of all ChEMBL actives available for a target). To rank a candidate, the highest Tanimoto score is taken out of the N computed values. As it was done earlier, in order to ensure reproducible results, averaging out the dependence on the randomly picked query compound(s), all similarity-based calculations were repeated 10 times, and the mean ROC AUC was computed for each target. In single-query searches, the $<AUC>_s$ value resulted from 10 individual similarity ranking simulations using 10 randomly picked active queries. With data fusion, 10-fold repeats of searches employing one half of the pool of actives generate the corresponding $<AUC>_{1/2}$ criterion that will be directly compared with equivalent $<AUC>_{1/2}$ criteria of the other VS methods, and the single-search $<AUC>_s$.

Eventually, the DUD pool was screened to obtain a VSAUC score using only the data fusion-based strategy, i.e. ranked according to their Tanimoto score with respect to their nearest neighbor of the entire corresponding ChEMBL series.

In order to measure the impact of dimensionality reduction/information loss by the GTM transformation of initial descriptors into responsibility vectors, similarity searching was performed in both descriptor and GTM responsibility vector spaces.

Results and discussion

Internal validation of the new UGTM versions

For above-cited technical reasons, this article introduces new, refitted "universal" GTM manifolds using a new GTM software release and extended selection sets of 236 (randomly picked) ChEMBL structure–activity class series associated to as many single protein targets. This undertaking is completely independent of the herein presented VS benchmark, as it focuses on the "multi-task" learning of the optimal compromise in terms of neighborhood behavior compliance over a large panel of targets, and even though this by no means a preparation step of the actual VS, UGTM performance analysis must be briefly discussed here. First, it must not be forgotten that, out of the 618 ChEMBL

target-specific series exploited by this study, 236 have a special status with respect to UGTMs: they served as selection sets for the optimal UGTM manifolds. This concerns two of the nine targets used in the VS simulation are included here (CHEMBL4439 and CHEMBL1952). By contrast, the remaining 382 external sets (including the other seven VS targets) were never used in UGTM tuning. It is thus legitimate to verify whether these 236 targets are favored—better predicted—by UGTMs, with respect to the latter. Figure 1 reports the distribution of "selection" versus "external" target-specific sets with respect to the internal validation ROC AUC values (see density distributions plots, in the Scoring section of methods). While the histograms show the expectable shift in favor of better results for the selection sets, this trend is very limited. Therefore, in the following analysis, no further distinction between selection and external ChEMBL sets will be done—statistics will indiscriminately refer to the set of 618 target-specific series. Furthermore, this observation is interesting, as it proves that MTL over ~ 200 structure–activity sets associated to fully non-related biological properties allows to cartograph the drug-relevant CS with a precision that is sufficient to ensure a same level of prediction accuracy for a large number of distinct biologically relevant targets to date.

Last but not least, let it be noted that even for the two targets CHEMBL4439 and CHEMBL1952 which served at map selection stage, the external validation by VS is no less rigorous than for any other of the herein benchmarked models. Any predictive model issued from supervised learning uses target-related information for calibration, and then is challenged to predict an independent compound set—as is the case here (DUD molecules filtered in order to ensure that they do not include any ChEMBL members). For all the nine targets, "coloring" of UGTM manifolds with ChEMBL data is the prerequisite to predict the likelihood to be active for the external DUD compounds—this is the equivalent of aforementioned model "calibration", except that it occurs in a deterministic and non-supervised manner—the manifold being already given. To resume, for two targets the injection of training information into UGTM models implies both manifold fitting and coloring, whilst for the seven others it implies only non-supervised manifold coloring. In either case, external validation concerns independent, never encountered compounds.

Internal validation benchmark

Comparative internal validation results for the various methods in terms of the above-defined $<AUC>_{1/2}$ ($<AUC>_S$ for single-query similarity screening) are given in Fig. 2. The poorest results come from single-query similarity, which is normal because the quantity of injected knowledge (one active reference) is minimal. Things are even worse

after dimensionality reduction: moving to responsibilities decreases performances even more. Nevertheless, with 50% of the mass of known actives used to color GTM fuzzy class landscapes, predictivity increases dramatically over single-query searches, and in spite of moving into the responsibility vector space.

Local maps are, as expected, better than universal maps. To begin with, they are already based on molecular descriptors known—thanks to the MTL of UGTM hyperparameters—to be generally pertinent choices, for a large pool of targets, Even though their control parameters were set to default values (likewise, the parameters of UGTMS being locked to the ones defining the best compromise neighborhood behavior), the degrees of freedom controlling the "bending" of their manifolds are now free to adjust specifically in response to the dedicated structure–activity series. Local maps might presumably be improved even more if their hyperparameters would be optimized.

Yet, similarity with data fusion, which is comparable to the GTM-based approach in terms of input SAR knowledge—50% of the actives—outperforms the former when driven in the original descriptor spaces: projection on a map inexorably costs in terms of information loss.

Eventually, NNs and RFs, are machine-learning approaches featuring a wealth of tunable parameters—unlike the fixed Universal and local GTM manifolds. Therefore, they are clearly the better performers.

In view of virtual screening of the DUD series, the best map for each target has been selected basing on its $<AUC>_{1/2}$ score. The number of targets for which the best map/descriptors space achieves a $<AUC>_{1/2} > 0.8$ have been counted for each method (Fig. 3).
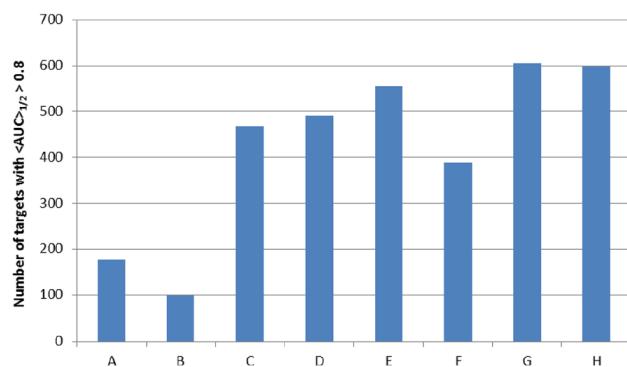
Fig. 3 The number of targets for which the best model over the four descriptor spaces returns $<AUC>_{1/2} > 0.8$. If, for a target, at least one of the four models of given type, based on the four descriptor spaces reaches this threshold, then the target will be added to the type bin: A—similarity search in initial space, B—similarity search in responsibility space, C—UGTM, D—local GTM, E—similarity search with data fusion in initial space, F—similarity search with data fusion in responsibility space, G—NN, H—RF

The bar chart in Fig. 3 keeps the trend seen in Fig. 1 and demonstrates that RF and NN outperform the GTM approach. At the same time, local GTM demonstrates the ability to be used successfully for 490 targets which makes it comparable with Similarity search with data fusion, which successfully handles 555 of the targets.

Virtual screening simulation using DUD compounds

The last part of the project is devoted to the retrieval, by VS, of actives among DUD compounds, with the ChEMBL-data-driven models. As it was described earlier, nine targets were found in common for DUD and ChEMBL (Tables 1, 2), where the smallest series includes more than 6000 compounds from DUD and more than 200 compounds from ChEMBL. The most data-rich target contains more than 33,000 compounds from DUD and more than 6000 compounds from ChEMBL.

Note that the DUD classification into actives and (presumably) inactive decoys is conceptually different from the classifications employed in the training sets. DUD actives may, for example, include co-crystallized ligands of high micromolar to millimolar potency, which are far from qualifying as "actives" by ChEMBL standards. This fact is potentially harmful for the external "prediction" performance monitored here—yet, this class of artefacts generally applies to classification models, which are the last recourse in response to highly heterogenous affinity measures that cannot be directly compared unless they are converted to "classes" according to more or less rigorous criteria. However, relative comparison of method performances should still be possible—if extrapolation from ChEMBL data to the DUD set is successfully accomplished by at least some methods, failure to do so by others cannot be ascribed to classification artefacts. This is the case in the present work.

To screen the DUD pool, the best maps were chosen based on their mean ROC AUC value obtained in internal-validation (Table 5).

In this VS simulation, the QSAR-based approaches were used, with the hypothesis (colored landscape, learned model) being based on the entire ChEMBL series of the nine above-mentioned targets. Single-query similarity searching was not considered here, as its intrinsic limitations due to the poverty of injected knowledge (a single active) were clear from internal validation results. In addition to ROC AUC, an Enrichment Factor (EF) within the 10% of top ranked compounds was added as a second criterion to estimate the quality of the predictions. The results of the external validation are shown in the Figs. 4 and 5.

Here, the predictive performance for the UGTM approach varies within $0.55 \div 0.9$ in terms of ROC AUC and within 0.2–6.2 in terms of the EF. Local GTMs show much better performance (ROC AUC ranges within 0.75–0.9, EF ranges within 2.2–8.2). While NNs were on par with RF and outperformed GTM models in terms of internal validation results, it appears that they are no longer systematically among top performers in VS, where similarity searching, RF and local GTM models are often much more robust. The activity landscapes and the DUD projections done for the target CHEMBL4282 and presented in Fig. 6 show that most of the DUD compounds are within the occupied zones (in other words, within the GTM applicability domain).

It is also seen from the DUD and ChEMBL activity landscapes that active DUD compounds are projected onto active zones of ChEMBL, which makes the ROC AUC and EF very high.

Discussion

The construction procedure of "universal" maps supporting multiple predictive landscapes on a same GTM manifold is a novel strategy in MTL. It is atypical in several aspects:

- First, it includes both operational parameters of the GTM model and hyperparameters. The key hyperparameter here is the choice of the molecular descriptor space,

Table 5 ROC AUC values and corresponding descriptors space for the best models computed within the internal validation

Target ID	UGTM	Local GTM	Similarity search in initial space	Similarity search in latent space	NN	RF
CHEMBL1827	0.89/4[a]	0.88/2	0.92/2	0.82/4	0.97/1	0.97/1
CHEMBL1952	0.88/4	0.84/4	0.85/4	0.76/4	0.92/1	0.92/3
CHEMBL251	0.84/3	0.84/2	0.91/2	0.81/3	0.95/2	0.96/3
CHEMBL260	0.76/2	0.77/2	0.9/3	0.81/3	0.95/3	0.95/1
CHEMBL279	0.74/2	0.71/3	0.89/3	0.76/3	0.93/3	0.93/4
CHEMBL301	0.82/4	0.83/4	0.91/2	0.8/3	0.94/2	0.95/3
CHEMBL4282	0.83/3	0.88/2	0.94/2	0.83/3	0.96/2	0.96/2
CHEMBL4338	0.83/1	0.86/3	0.85/3	0.78/3	0.94/2	0.93/2
CHEMBL4439	0.88/2	0.9/2	0.89/2	0.87/3	0.94/2	0.94/3

[a]Mean ROC AUC/No. of a map/descriptors space corresponded to Table 3

Fig. 4 The comparison of the VS methods, where each column corresponds to the best map in terms of its ROC AUC value computed in the internal validation (see Table 5)
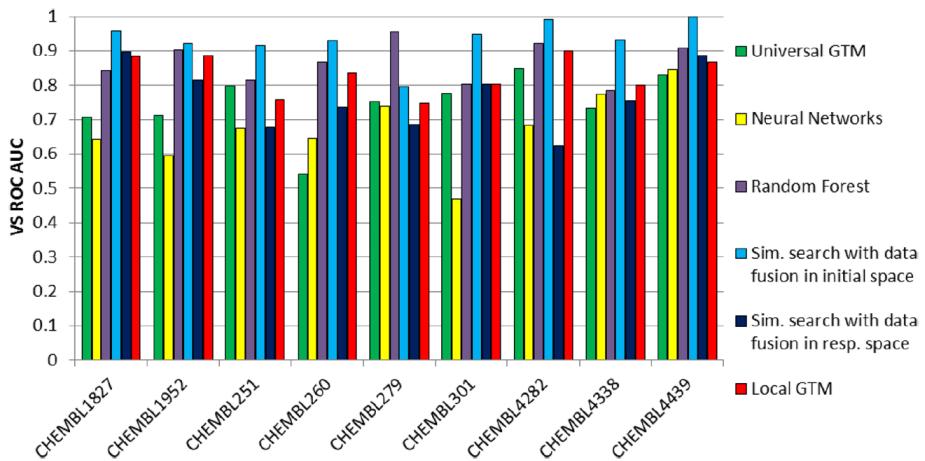
Fig. 5 The EF for different VS approaches where the EF value is given for the map with the highest ROC AUC value computed in the internal validation (see Table 5)
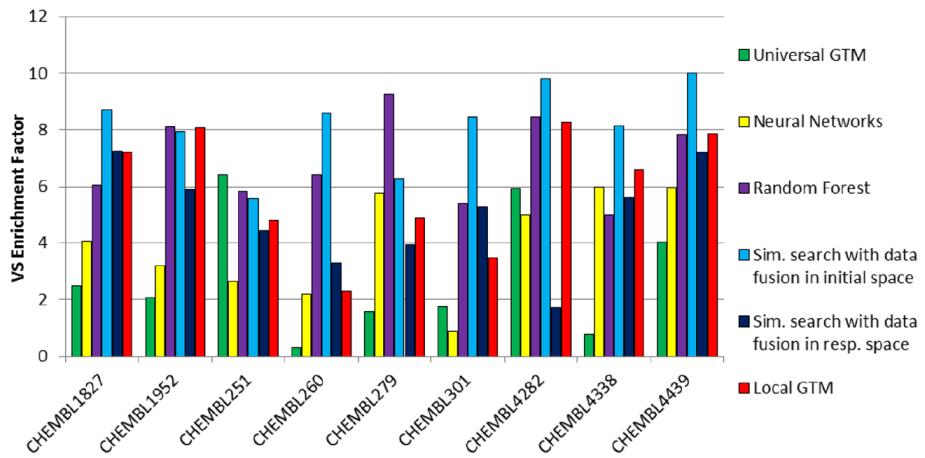
allowing the procedure to select those descriptor spaces which remain neighborhood behavior-compliant after GTM-driven dimensionality reduction

- Second, its multi-task nature is given by the construction of a common manifold, which is, per se, an unsupervised learning process aimed at maximizing the coverage of FS compounds by this manifold. This common manifold is challenged to host fuzzy classification landscapes for many different biological targets. Each of them is a classical single-task model for the property associated to the ligands that were used to color the specific landscape. However, since these landscape-based predictive models do not feature any specific fitable parameters, their quality can be regarded as an intrinsic property of the underlying common manifold. Creation of the manifold implicitly provides access to as many landscape-driven predictive models as available property-annotated ligand series. The MTL—here primarily consisting in selecting optimally suited descriptor spaces and optimally associated GTM grid size, manifold flexibility parameters, etc.—was directed by the goal of discovering (hyper) parameter combinations maximizing the mean quality of

236 distinct "selection" series of target-specific activity-annotated ligands

- Third, it does not focus on specific transfer of knowledge within biologically related targets, such as is the case in computational chemogenomics. This MTL simultaneously addressed the rather exhaustive set of all human protein targets with sufficient activity annotations in ChEMBL, all protein families confounded. Neither the 236 "selection" series of target-specific activity-annotated ligands, nor the remaining 382 series used for external validation (with comparable success rate to the former 236) include any intended family-specific bias in terms of biological targets. Here, MTL would not target typical questions like "What are the common features of kinase binders?", but more general "What are the common features of bioactive molecules, all targets confounded?"

Uncovering the few ISIDA fragmentation schemes that are optimally suited for this endeavor is a first key result of this atypical multitask learning setup. Since descriptor spaces cannot host predictive GTM models unless they are,
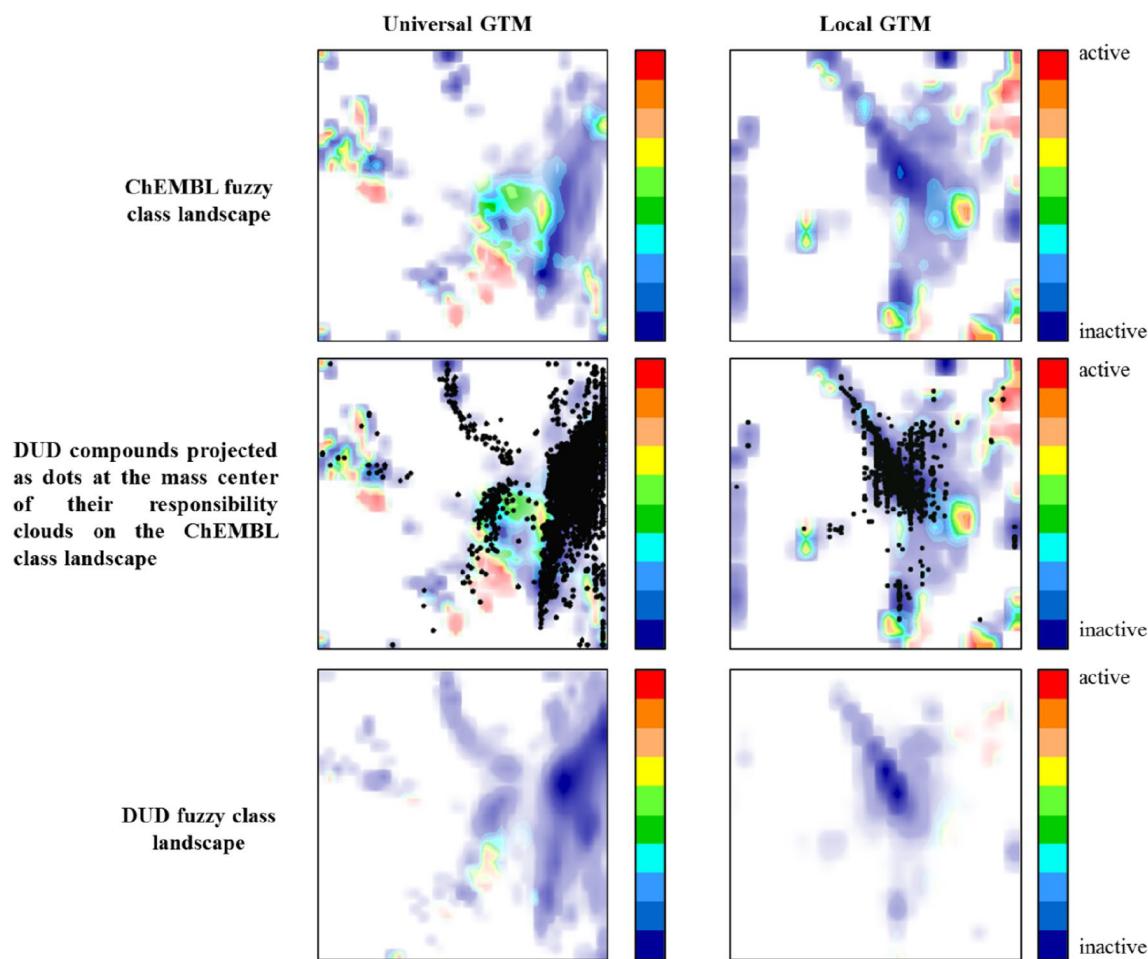
Fig. 6 Fuzzy class landscape representations of the (ChEMBL and respectively DUD) sets associated to target CHEMBL4282 on universal map 3 (left) versus the local GTM (right)

per se, neighborhood behavior-compliant, it is unsurprising to observe that all the alternative approaches—from data-fusion-driven similarity searching to target-dedicated local GTM, RF and NN models—were rather successful, both in terms of internal validation and external VS. There was no need to rescan, for each predictive method, the entire set of available molecular descriptor spaces—the choices of the evolutionary UGTM builder were appropriate. Note that the 100 different descriptor spaces out of which the four herein used were selected have themselves emerged as a historical accumulation of descriptor spaces that were used in the past [3, 11], on rather unrelated problems such as library comparison, and were seen to be successful. In this sense, if we declare all the cases in which knowledge from previous experiences is actively used to restrain the scope of effectively considered working hypotheses as some form of "multi task" learning, then MTL is rather the rule than the exception in chemoinformatics.

UGTM models are remarkably robust in VS—for models with zero adjustable parameters, albeit they are systematically outperformed—in particular with respect to enrichment of the top selection—by the equally parameter-free data-fusion similarity searching, not affected by information loss upon dimensionality reduction. However, UGTM models are specifically failing to rank a significant number of actives among the top 100 candidates—they are not effective in ensuring high EF values in VS. By contrast, their global ROC AUC scores show that they do, overall, manage to eventually rank actives ahead of most of the inactives, only slightly less effective than the other methods—without systematically placing actives at the top of the list.

Responsibility vectors are still maintaining some degree of neighborhood behavior-compliance, but their use in similarity searching is not recommended, as landscape-driven prediction on UGTM manifolds is the more powerful method. Note that data fusion-based similarity screening with Q actives being used as queries would scale like QxN in terms of computational effort required to virtually screen a database on N candidates. By contrast, landscape-based prediction effort is simply proportional to N and does not

depend on the training set size used to create the predictive landscape. Thus, the latter would become computationally more interesting after a given Q value—not to mention all the benefits stemming from intuitive visualization provided by the GTM approach.

Conclusions

The previously reported strategy to generate "universal" maps, able to support predictive models for a broad spectrum of biological activities represents a generic MTL approach, where optimal molecular descriptors are selected alongside with optimal operational parameters of the GTM algorithm. A first important outcome of the approach is uncovering "multicompetent" molecular descriptor spaces that remain neighborhood behavior-compliant even after the dimensionality reduction process—leading to GTM responsibility vectors and ultimately to a (x, y) point in 2D GTM latent space. These tend to correspond to ISIDA fragmentation schemes restricted to rather small fragment sizes but incorporating information-rich atom labels such as pH-dependent pharmacophore types or CVFF force field types.

It could be shown that descriptors herewith selected are not only an excellent support for GTMs, but also for many other predictive models—starting with plain similarity screening. In this sense, all models here implicitly benefitted from the initial MTL, which provided a pool of four descriptor spaces that turned out to be highly relevant for all the envisaged QSAR model building procedures for more than 600 completely independent targets.

Tanimoto-score-based similarity screening (using a data fusion scenario, thus ensuring that the amount of information injected into it—active examples—matches the sizes of the training sets used by other approaches) is actually more successful than UGTM-driven predictions, as information loss upon dimensionality reduction is unavoidable.

Local GTMs, where manifolds are allowed to focus on the chemical subspace populated by a single target-specific ligand series, are unsurprisingly better performers than their universal, consensus-oriented counterparts. Note, however, that the latter would always represent a better choice whenever the activity-annotated data set pertaining to a target of interest is not sufficient to support the fitting of local maps. The same holds true for parameter-rich non-linear RF and NN models.

References

1. Bishop CM, Svensén M, Williams CK (1998) GTM: the generative topographic mapping. Neural Comput 10(1):215–234
2. Kohonen T (1990) The self-organizing map. Proc IEEE 78(9):1464–1480
3. Lin A, Horvath D, Afonina V, Marcou G, Jean-Louis R, Varnek A (2018) Mapping of the available chemical space versus the chemical universe of lead-like compounds. ChemMedChem 13:540–554. https://doi.org/10.1002/cmdc.201700561
4. Kireeva N, Baskin I, Gaspar H, Horvath D, Marcou G, Varnek A (2012) Generative topographic mapping (GTM): universal tool for data visualization, structure—activity modeling and dataset comparison. Mol Inform 31(3–4):301–312
5. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) GTM-based QSAR models and their applicability domains. Mol Inform 34(6–7):348–356. https://doi.org/10.1002/minf.201400153
6. Muegge I, Oloff S (2006) Advances in virtual screening. Drug Discov Today 3(4):405–411. https://doi.org/10.1016/j.ddtec.2006.12.002
7. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20(3):318–331. https://doi.org/10.1016/j.drudis.2014.10.012
8. Hristozov D, Oprea TI, Gasteiger J (2007) Ligand-based virtual screening by novelty detection with self-organizing maps. J Chem Inf Model 47(6):2044–2062. https://doi.org/10.1021/ci700040r
9. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. J Med Chem 50(7):1698–1702. https://doi.org/10.1021/jm060604z
10. Schneider G, Nettekoven M (2003) Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. J Comb Chem 5(3):233–237
11. Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D (2015) Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. J Comput Aided Mol Des 29(12):1087–1108. https://doi.org/10.1007/s10822-015-9882-z
12. Rosenbaum L, Dörr A, Bauer MR, Boeckler FM, Zell A (2013) Inferring multi-target QSAR models with taxonomy-based multi-task learning. J Cheminform 5(1):33
13. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV (2009) Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J Chem Inf Model 49(1):133–144. https://doi.org/10.1021/ci8002914
14. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V (2017) Demystifying multitask deep neural networks for quantitative structure–activity relationships. J Chem Inf Model 57(10):2490–2504
15. Brown JB, Okuno Y, Marcou G, Varnek A, Horvath D (2014) Computational chemogenomics: is it more than inductive transfer? J Comput Aided Mol Des 28(6):597–618. https://doi.org/10.1007/s10822-014-9743-1
16. Heikamp K, Bajorath J (2013) Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations. J Chem Inf Model 53(4):791–801. https://doi.org/10.1021/ci400090t
17. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA (2013) Shifting from the single to the multitarget paradigm in drug discovery. Drug Discovery Today 18(9–10):495–501. https://doi.org/10.1016/j.drudis.2013.01.008

18. Bieler M, Heilker R, Koeppen H, Schneider G (2011) Assay related target similarity (ARTS)—chemogenomics approach for quantitative comparison of biological targets. J Chem Inf Model 51(8):1897–1905. https://doi.org/10.1021/ci200105t

19. Jacob L, Hoffmann B, Stoven V, Vert J-P (2008) Virtual screening of GPCRs: an in silico chemogenomics approach. BMC Bioinform 9(1):363

20. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(D1):D1100–D1107

21. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular Docking. J Med Chem 49(23):6789–6801. https://doi.org/10.1021/jm0608356

22. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA property-labelled fragment descriptors. Mol Inform 29(12):855–868. https://doi.org/10.1002/minf.201000099

23. Ruggiu F, Marcou G, Solov'ev V, Horvath D, Varnek A (2017) ISIDA fragmentor 2017-user manual. http://infochim.u-strasbg.fr/downloads/manuals/Fragmentor2017/Fragmentor2017_Manual_nov2017.pdf

24. Horvath D, Brown J, Marcou G, Varnek A (2014) An evolutionary optimizer of libsvm models. Challenges 5(2):450–472

25. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. J Chem Inf Model 55(1):84–94. https://doi.org/10.1021/ci500575y

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830

27. Ruck DW, Rogers SK, Kabrisky M, Oxley ME, Suter BW (1990) The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Trans Neural Netw 1(4):296–298. https://doi.org/10.1109/72.80266

28. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533

29. Breiman L (2001) Random forests. Mach Learn 45(1):5–32

30. Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), 2013, IEEE, Vancouver, pp 8609–8613

31. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980

32. Horvath D, Koch C, Schneider G, Marcou G, Varnek A (2011) Local neighborhood behavior in a combinatorial library context. J Comput Aided Mol Des 25(3):237–252. https://doi.org/10.1007/s10822-011-9416-2

33. Papadatos G, Cooper AWJ, Kadirkamanathan V, Macdonald SJF, McLay IM, Pickett SD, Pritchard JM, Willett P, Gillet VJ (2009) Analysis of neighborhood behavior in lead optimization and array design. J Chem Inf Model 49(2):195–208. https://doi.org/10.1021/ci800302g

5.1.2 Conclusion

The universal GTM was tested as a tool for single-target and multi-target virtual screening tasks. It was shown that local GTM possesses better predictive performance than the universal approach. Even so, the universal GTM predicted almost 500 ChEMBL targets with ROC AUC > 0.8 in the internal validation. In the external validation, 8 out of 9 targets were predicted with ROC AUC > 0.7. In terms of the enrichment factor, only half of the DUD targets were predicted well.

In contrast, the single-target GTM approach demonstrates high predictive performance which is comparable to other VS techniques described in the paper. Almost 500 ChEMBL targets were predicted with ROC AUC > 0.8 in the internal validation. In the virtual screening of the DUD database, local GTM even overcomes the MLP with one hidden layer, and it is comparable to RF. The same tendency is also demonstrated by the enrichment factor.

The results show that GTM can be efficiently applied as a filter in the VS funnel. Its speed and predictive performance are comparable to other popular VS techniques, whereas it has the advantage of visualization support.

5.2 Virtual Screening in Industrial Context

5.2.1 Introduction

The benchmarking results presented above demonstrate that the universal GTM can be applied in VS campaigns. One or several universal maps can easily work a with wide range of assays and cover different chemotypes. Therefore, it was decided to test GTM in the industrial environment of Boehringer Ingelheim Pharma company (BI). For this purpose, their proprietary database of 1.7M compounds was used to train the manifold. Next, the map is used to predict more than 2.3K assays as well as some ADME properties.

5.2.2 Data

1.7M structures were standardized by ChemAxon Standardizer [81] using the following protocol:

1) Dearomatization;

2) Remove stereo;

3) Remove explicit hydrogens;

4) Remove solvents;

5) Aromatization;

6) Normalize default ChemAxon Standardizer chemotypes (nitro, azide, diazo, phosphonic, etc.).

To validate the GTM models, BI bio profile was used where a list of IC50/EC50 values was given. 6848 assays were presented in the profile but only 3320 assays containing more than 100 records were taken. The labels assignment protocol described in Figure 29 was applied to split the data into 3 classes: active, weakly active and inactive.

First, the algorithm optimizes the threshold for the "active" class to collect at least 15 compounds. The active threshold ranges within 10 and 1000 nM (not systematically; see Figure 29). Next, it tunes the threshold for the "inactive" class maximizing the number of items but keeping the ratio of the thresholds ($\mathbf{Inact_{IC50}}$ / $\mathbf{Act_{IC50}}$) at least 10 folds or greater. Here, the inactive threshold varies from 1 μM to 10 μM with a step of 1 μM. Once 30% of compounds are collected as inactives (at least 15), the ratio of the thresholds is checked again, and, if it is larger than 10, the active threshold ($\mathbf{Act_{IC50}}$) is increased in a way that it becomes to be 10 times smaller than the inactive threshold ($\mathbf{Inact_{IC50}}$).

2371 assays associated with sufficiently large (at least 30 compounds/series) and conveniently balanced (no less than 15 actives and 15 inactives) structure-activity series were selected. The external validation was performed using new data points measured in BI 6 months later.

1st step

Act$_{IC50}$ = 10nM

N$_{act}$ ≥ 15 — Yes / No

Act$_{IC50}$ = 50nM

N$_{act}$ ≥ 15 — Yes / No

– – – – – –

Act$_{IC50}$ = 1µM

N$_{act}$ ≥ 15 — Yes / No

Discard the assay

Act$_{IC50}$ ÷ [10nM, 50nM, 100nM, 300nM, 500nM, 700nM, 1µM]

2nd step

Inact$_{IC50}$ = 10µM

Condition 1 — Yes / No

Inact$_{IC50}$ = 9µM

Condition 1 — Yes / No

– – – – – –

Inact$_{IC50}$ = 1µM

Condition 2 — Yes / No

Discard assay

Condition 1:
N$_{inact}$ ≥ 15 and N$_{inact}$ ≥ 30%

Condition 2:
N$_{inact}$ ≥ 15
On each step, the ratio $\frac{Inact_{IC50}^{new}}{Act_{IC50}} \geq 10$ is checked to be true. If it is false, the cycle stops, and the condition 2 is checked.
Inact$_{IC50}$ ÷ 10-1 µM, step 1 µM.

3rd step

$\frac{Inact_{IC50}^{new}}{Act_{IC50}} > 10$ — Yes / No

Cutoff for actives:
Inact$_{IC50}$/10
Cutoff for inactives:
Inact$_{IC50}$

Cutoff for actives:
Act$_{IC50}$
Cutoff for inactives:
Inact$_{IC50}$

The scenario when the ratio is less than 10 will not occur here since it was discarded in step 2.

Figure 29. Labels assignment protocol which bases on IC50 value of compounds. Here, Act$_{IC50}$ is the threshold on IC50 for active compounds; Inact$_{IC50}$ is the threshold on IC50 for inactive compounds.

5.2.3 Method

To find a suitable universal map(s), a grid search was run. Within this search, 4 GTM parameters (Table 5) and descriptor space were optimized. Here, 100 fragmentation schemes supported by the ISIDA Fragmentor software [80, 95] were used as a starting pool for the search of a suitable descriptor space. These 100 fragmentation schemes were gathered according to the experience of previous works [9, 50].

To build the GTM manifold, a Frame set (FS) of 25K compounds was prepared. Here, the FS is fixed to reduce the number of tunable parameters. To gather the FS, clustering procedure with Tanimoto=0.7 was performed (done by BI earlier). As a result, more than

135K clusters were found. 25K clusters out of it were chosen randomly, where one random compound represents each particular cluster.

Table 5. GTM parameters ranges set for the grid search.

Name of the parameter	Starting value	Ending value	Step
Number of nodes (root value), **k**	20	50	5
Number of RBF centers (root number), **m**	40% out of the number of nodes	70% out of the number of nodes	10
Regularization coefficient, **l**	1.0	5.0	0.5
Width of an RBF center, **w**	1.0	5.0	1.5

Once the descriptors were computed, they were normalized and filtered according to their standard deviation (rare columns for which its standard deviation is lower than 2% of the value range were removed). To train the manifold, the incremental GTM algorithm with 5K items in a block was used (chapter 3.1.2) [5].

The goal of this virtual screening was to distinguish 3 classes: actives, weakly actives and inactives. Therefore, classification models with 3 classes as well as with 2 classes (just active and inactive) were built. To evaluate the models, a 3-folds cross-validation procedure was performed for 500 random assays (the validation on the entire set of assays is time-consuming). As a score, the mean area under the Receiver Operating Characteristic (*ROC AUC*) was computed for each class within one fold: actives against others, inactives against others, and middle compounds against others. The result was averaged over the 3 folds, and then over 500 assays. This ROC AUC was used to estimate the quality of the map(s) ($< \overline{AUC} >^{3\,cls}$ and $< \overline{AUC} >^{bin}$ for 3 classes and 2 classes, respectively).

In addition to the mean ROC AUC values, some other scores were used:

- Number of assays for which the mean ROCAUC ≥ 0.5;

- Number of assays for which the mean ROCAUC ≥ 0.6;

- Number of assays for which the mean ROCAUC ≥ 0.7 (main score used in 3 classes classification to select the best map);

- Number of assays for which the mean ROCAUC ≥ 0.8;

- Number of assays for which the mean ROCAUC ≥ 0.9.

Once the top-5 maps are chosen, they will be checked using all 2371 assays.

5.2.4 Results and Discussion

In the grid search, more than 226K GTMs were trained and cross-validated. The ROC AUC scores obtained for the best maps with different map resolution are shown in Figure 30. One can see that the map with 25*25 nodes is already enough to perform 2 classes classification, whereas for 3 classes higher map resolution is better.
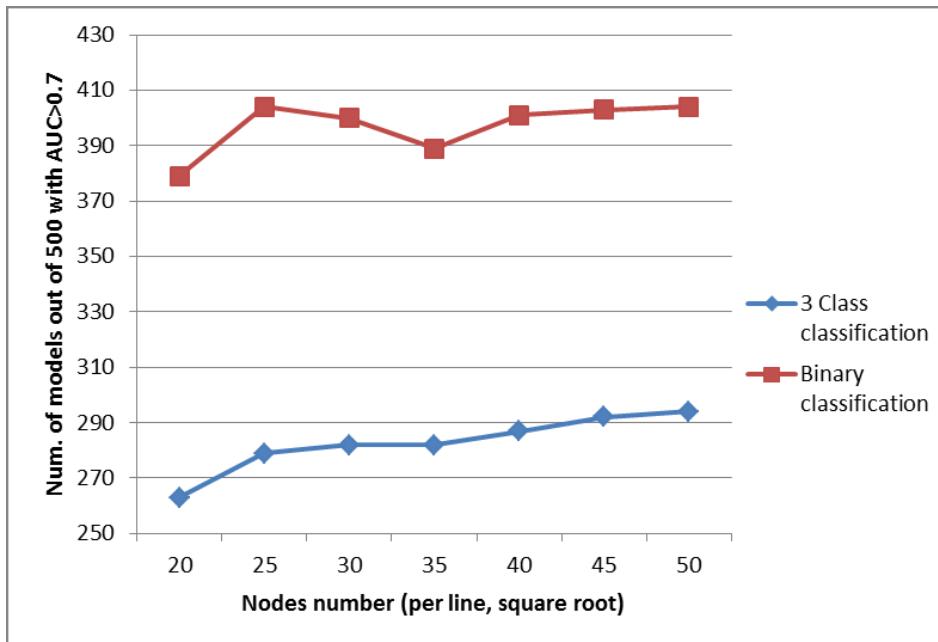
Figure 30. The grid search progress. Here, the number of models aligned along the Y axe corresponds to the best map with the current map resolution.

The maps were sorted according to the number of assays predicted with the mean AUC over the 3 classes ($< AUC >^{3\,cls}$) larger than 0.7. The best 5 maps were selected (Table 7). The explanation of the corresponding descriptors is given in Table 6.

Table 6. Descriptors explanation [80, 95].

Descriptors abbreviation	Description
IB--FC-AP-2-11	Sequences of bonds of length 2 to 11 using formal charges and all paths
III-PH-3-6	Triplets of length 3 to 6 using pharmacophores
IB--FC-2-11	Sequences of bonds of length 2 to 11 using formal charges

These maps were then validated on the entire set of 2371 assays. The results are in Table 8.

One can see from Table 7 that the best map in 3 classes cross-validation successfully predicted 59% of given assays (294 out of 500). In 2 classes validation, the result is even better (80%). The same trend was demonstrated in cross-validation on the entire set (1318 out of 2371 assays were predicted well by the map 1; Table 8).

Table 7. Top-5 maps sorted by ROCAUC$\geq$0.7 in 3-classes task (see the abbreviations in Table 5 and Table 6).

Maps' description	$< \overline{AUC} >^{3\,cls}$	Number of assays where ROCAUC (3 classes)					$< \overline{AUC} >^{bin}$	Number of assays where ROCAUC (2 classes)				
		$\geq$0.5	$\geq$0.6	$\geq$0.7	$\geq$0.8	$\geq$0.9		$\geq$0.5	$\geq$0.6	$\geq$0.7	$\geq$0.8	$\geq$0.9
k=50, m=20, l=2.5, w=1.0, descriptors: IB--FC-AP-FC-2-11 (**map 1**)	0.71	494	448	294	65	2	0.8	493	473	401	297	87
k=45, m=31, l=3.0, w=1.0, descriptors: III-PH-3-6 (**map 2**)	0.71	494	448	294	65	1	0.79	493	473	401	297	87
k=45, m=22, l=5.0, w=1.0, descriptors: III-PH-3-6 (**map 3**)	0.71	494	448	294	65	1	0.8	493	473	401	297	87
k=50, m=20, l=2.0, w=1.0, descriptors: IB--FC-2-11 (**map 4**)	0.71	494	448	294	65	3	0.8	493	473	401	297	87
k=40, m=16, l=3.5, w=1.5, descriptors: III-PH-3-6 (**map 5**)	0.71	494	448	294	65	2	0.79	493	473	401	297	87

Table 8. Top-5 maps (Table 7) validated with 2371 assays.

Maps' description	$<\overline{AUC}>^{3\,cls}$	Number of assays where ROCAUC (3 classes)					$<\overline{AUC}>^{bin}$	Number of assays where ROCAUC (2 classes)				
		≥ 0.5	≥ 0.6	≥ 0.7	≥ 0.8	≥ 0.9		≥ 0.5	≥ 0.6	≥ 0.7	≥ 0.8	≥ 0.9
Map 1 [a]	0.71	2354	2114	1318	311	4	0.8	2341	2212	1898	1305	428
Map 2	0.71	2354	2114	1318	311	4	0.79	2341	2212	1898	1305	428
Map 3	0.71	2354	2114	1318	311	4	0.8	2341	2212	1898	1305	428
Map 4	0.71	2354	2114	1318	311	4	0.8	2341	2212	1898	1305	428
Map 5	0.71	2354	2114	1318	311	4	0.79	2341	2212	1898	1305	428

[a] See Table 7.

To validate the maps in ADME properties, the latter ones were classified, and $< \overline{AUC} >^{3\,cls}$ and $< \overline{AUC} >^{bin}$ were computed (Table 9). The $< \overline{AUC} >^{3\,cls}$ values demonstrate that the map 1 stays at the top in both 3 classes and 2 classes classification. The average $< \overline{AUC} >^{3\,cls}$ for the map 1 varies from 0.65 to 0.72.

Table 9. Validation results for ADME properties.

ADME property	Map 1 [a]		Map 2		Map 3		Map 4		Map 5	
	$<\overline{AUC}>^{3\,cls}$	$<\overline{AUC}>^{bin}$	$<\overline{AUC}>^{3\,cls}$	$<\overline{AUC}>^{bin}$	$<\overline{AUC}>^{3\,cls}$	$<\overline{AUC}>^{bin}$	$<\overline{AUC}>^{3\,cls}$	$<\overline{AUC}>^{bin}$	$<\overline{AUC}>^{3\,cls}$	$<\overline{AUC}>^{bin}$
Caco2_Efflux	0.69	0.76	0.68	0.76	0.68	0.76	0.68	0.77	0.66	0.74
CL_Mouse	0.67	0.75	0.64	0.7	0.66	0.73	0.65	0.75	0.65	0.7
CL_Rat	0.66	0.75	0.64	0.72	0.65	0.73	0.65	0.75	0.62	0.72
HHEP	0.66	0.71	0.69	0.76	0.68	0.77	0.67	0.74	0.68	0.77
HLM	0.65	0.72	0.62	0.69	0.62	0.69	0.63	0.71	0.62	0.67
MDCKBCRP_Efflux	0.66	0.73	0.68	0.75	0.7	0.78	0.65	0.74	0.68	0.75
MDCKPGP_Efflux	0.69	0.76	0.68	0.75	0.68	0.75	0.67	0.74	0.67	0.73
MHEP	0.68	0.75	0.69	0.74	0.68	0.74	0.68	0.75	0.7	0.74
MLM	0.68	0.76	0.66	0.75	0.66	0.74	0.67	0.76	0.65	0.72
PPBhuman	0.72	0.82	0.7	0.8	0.7	0.79	0.7	0.8	0.69	0.79
PPBmouse	0.72	0.82	0.72	0.83	0.69	0.79	0.7	0.79	0.7	0.8
RHEP	0.67	0.75	0.66	0.78	0.65	0.73	0.67	0.75	0.65	0.73
RLM	0.65	0.74	0.62	0.68	0.62	0.68	0.64	0.72	0.61	0.66
SOL68	0.66	0.7	0.63	0.66	0.63	0.66	0.65	0.68	0.62	0.65
Mean	0.68	0.75	0.66	0.74	0.66	0.74	0.66	0.75	0.66	0.73

[a] See Table 7.

The last step was to externally validate the maps using new data for 42 assays. The Balanced Accuracy in 3 classes classification was above 0.5 for 30 assays.

5.2.5 Conclusion

Five GTMs were trained and selected out of 236K maps produced by grid search optimizer. They were cross-validated on more than 2.3K assays from BI. The cross-validation demonstrated that about 55% of the assays are predicted with ROC AUC $\geq$ 0.7. The external validation on 42 assays for which new data were received showed that 30 out of 42 assays are predicted well (Balanced Accuracy $\geq$ 0.5 in 3 class classification).

6 Public Chemical Databases Comparison

6.1   Introduction

Chemical databases are huge, and they grow each year since new records are added to public and private chemical databases. Nowadays, the largest public chemical resources (PubChem, CAS, Zinc) contain millions and even hundreds of millions of compounds. However, the potential of the full chemical space is much larger. So far, P. Polishchuk et al. [3] have guesstimated the drug-like space as 10^{33} compounds.

Analysis of large chemical space is a real challenge that requires suitable chemoinformatics tools. Generative Topographic Mapping (GTM) has been already tested as a tool to analyze big data sets (up to 2M items). In this project, we raise the bar (up to 20M compounds) and test GTM in the task of big chemical libraries analysis and comparison. For this purpose, a data set of existing compounds from PubChem database with no more than 17 heavy atoms were compared to virtually generated compounds from the FDB-17 database [7]. The data sets were compared using *(i)* Bhattacharyya, Soergel and Euclidean distances, *(ii)* GTM class landscapes, and *(iii)* GTM property landscapes. To resolve the problem of GTM resolution and to find unique for a given database chemotype, hierarchical GTM zooming technique described in chapter 4.3 was applied, see below our publication in *ChemMedChem* [50].

Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds

Arkadii Lin,[a] Dragos Horvath,*[a] Valentina Afonina,[a, b] Gilles Marcou,[a] Jean-Louis Reymond,[c] and Alexandre Varnek*[a]

This is, to our knowledge, the most comprehensive analysis to date based on generative topographic mapping (GTM) of fragment-like chemical space (40 million molecules with no more than 17 heavy atoms, both from the theoretically enumerated GDB-17 and real-world PubChem/ChEMBL databases). The challenge was to prove that a robust map of fragment-like chemical space can actually be built, in spite of a limited ($\ll 10^5$) maximal number of compounds ("frame set") usable for fitting the GTM manifold. An evolutionary map building strategy has been updated with a "coverage check" step, which discards manifolds failing to accommodate compounds outside the frame set. The evolved map has a good propensity to separate actives from inactives for more than 20 external structure–activity sets. It was proven to properly accommodate the entire collection of 40 M compounds. Next, it served as a library comparison tool to highlight biases of real-world molecules (PubChem and ChEMBL) versus the universe of all possible species represented by FDB-17, a fragment-like subset of GDB-17 containing 10 million molecules. Specific patterns, proper to some libraries and absent from others (diversity holes), were highlighted.

Introduction

Nowadays, chemical databases include millions of chemical structures, and this number exponentially increases because of the implementation of parallel and combinatorial synthesis approaches, as well as new experimental techniques like flow or microwave reactors. Yet, these databases cover only a small part of chemical space or the "universe" of all possible molecules. The exploration of this chemical space is a challenge for chemists seeking to understand its structure, to discover its unexplored regions, and to analyze the structural relationships between the compounds that it encompasses. Chemoinformatics,[1] representing compounds as molecular graphs and encoding them as vectors of descriptors, is the paramount tool for rational navigation of this chemical space. Associating structures to recorded experimental properties and learning from this big data[2, 3] is a key challenge of chemoinformatics.

The key to successful chemical space mapping is compliance with the similarity principle:[4] similar compounds, which are expected to have similar properties, must appear as neighboring entities on the chemical space map, irrespective of how it was built. There are various strategies to represent the chemical space with chemoinformatics support, and these depend primarily on the manner in which the chemical information is represented: graph-based or descriptor-based. In graph-based approaches, data visualization is based on substructures/scaffolds and their hierarchical relationships (the scaffold tree[5] or scaffold network[6]). In the case of descriptor-based chemical spaces, each molecule is represented by a D-dimensional vector. Based on this fact, two popular approaches can be used: similarity network graphs[7] (with nodes representing molecules that are connected, that is, "neighbors", if their similarity exceeds a user-defined threshold) or dimensionality reduction (objects from the D-dimensional chemical space are transferred into a latent space of two or three dimensions). Unfortunately, the use of large sets of data imposes a limit on the list of methods that can be used for data visualization. Three basic methods fit these restrictions: principal component analysis (PCA),[8] self-organizing Kohonen mapping (SOM),[9, 10] and generative topographic mapping (GTM).[11, 12] The main drawback of PCA is that it is a linear method of dimensionality reduction, and in some cases, a small number of principal components (at most three, in order to obtain a human-readable projection) explains only a small part of data variance. Another problem comes from the low information content of PCA plots, which results from the tendency to concentrate most of

[a] A. Lin, Dr. D. Horvath, V. Afonina, Dr. G. Marcou, Dr. A. Varnek
Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4 Blaise Pascal str., 67081 Strasbourg (France)
E-mail: dhorvath@unistra.fr
varnek@unistra.fr

[b] V. Afonina
Laboratory of Chemoinformatics and Molecular Modeling, Department of Organic Chemistry, A.M. Butlerov Institute of Chemistry, Kazan Federal University, 18 Kremlyovskaya str., 420008 Kazan (Russia)

[c] Dr. J.-L. Reymond
Department of Chemistry and Biochemistry, University of Berne, 3 Freiestrasse, 3012 Berne (Switzerland)

the data points in a certain region in the form of a Gaussian cloud, while the rest of the plot is left poorly populated. The second approach, SOM, does not have this drawback because it is a nonlinear dimensionality reduction method. Unfortunately, in case of this method, the output information is truncated to the assignment of a molecule into its residence node and the indication of how well it fits into this node.

Brief introduction to generative topographic mapping

Generative topographic mapping, as a probabilistic extension of SOM,[10] does not have above-mentioned disadvantages. Intuitively, it consists of inserting a two-dimensional manifold, to be imagined as a flexible "rubber sheet", into the high-dimensional descriptor space. The manifold is mathematically described by a square grid of reference points ("nodes") and a set of radial basis functions (RBFs); the number of nodes and the number of RBFs are key operational parameters to be specified at input. This rubber sheet is then "bent" in order to cover, as closely as possible, the descriptor space points corresponding to the "frame" molecules provided as input. Molecules are thus fuzzily associated to the closest grid nodes of the bent manifold: the degrees of association of each molecule to a node are called "responsibilities". Eventually, the rubber sheet is again "flattened out" as a regular grid of nodes in a 2D plane, and the molecules from the initial descriptor space can now be localized on this map based on their responsibilities. Therefore, GTM could be used not only as a chemical data visualization tool but also to build classification and regression structure–property models.

The formal methodology will be briefly described in the following paragraphs. GTM is a nonlinear dimensionality reduction approach that maps points from the D-dimensional data space to a two-dimensional latent space (the actual "map"). These spaces are connected by a nonlinear, parametric function $y(x,W)$. The latent space is represented by a squared $K \times K$ grid. Every point of the data space is mapped on the latent space with the generation of the corresponding probability distribution, that is, responsibilities with respect to the nodes of grid. This is, as already hinted, achieved by nonlinearly embedding a two-dimensional manifold in the D-dimensional space. The nodes x_k on the regular grid in the latent space are mapped to the corresponding centers of Gaussians y_k in data space, by using a parameterized, nonlinear mapping function $y_k = y(x_k, W)$.[13] Therefore, an instance t in data space (a molecule represented as a point of coordinates t according to a molecular descriptor) will be more strongly associated to any node x_k the closer its point t is situated with respect to the image y_k of the latent space node, as described by Equation (1), in which t_n is a data instance and β is the common inverse variance of the distribution.

$$p(t_n|x_k, W, \beta) = \frac{\beta^{D/2}}{2\pi} exp\left(\frac{\beta}{2}\|y_k - t_n\|^2\right) \tag{1}$$

The manifold may be distorted in order to match frame set compound coordinates t_n as closely as possible, in optimizing the "log likelihood" parameter described by Equation (2).

$$\mathcal{L}(W, \beta) = \sum_{n=1}^{N} ln\left\{\frac{1}{K}\sum_{k=1}^{K} p(t_n|x_k, W, \beta)\right\} \tag{2}$$

After optimization, the normalized probability of association of an instance n (a molecule) represented by descriptor vector t_n to a node x_k is labeled R_{kn} and called the responsibility vector of instance n [Eq. (3)].

$$R_{kn} = \frac{p(t_n|x_k, W, \beta)}{\sum_k p(t_n|x_k, W, \beta)} \tag{3}$$

To enable GTM usage for big data, it is necessary to avoid the need to upload the entire compound descriptor matrix for the entire frame set of N compounds. A set of $N \times D$-dimensional vectors, in which D may be of the order of 10^4 quickly becomes prohibitive in terms of memory requirements. To bypass this bottleneck, the incremental version of the GTM algorithm (iGTM) is used.[11,13] Instead of updating the model with the entire data matrix, iGTM divides the data into blocks and updates the model block by block until convergence of the log likelihood function.

Analysis of a compound library after GTM projection relies on three main tools: class maps, property landscapes, and density maps. In density maps, individual responsibility vectors of library compounds are added, and the local color intensity can be used as marker of cumulated responsibility.

Class maps and property landscapes imply some previous learning to associate specific classes/property values with map coordinates. In this process, the molecular property is "transferred" from the landscape training items to the nodes, which are being assigned the responsibility-weighted mean of properties P of the herein-residing molecules [Eq. (4)].[13]

$$\bar{P}_k = \frac{\sum_{n=1}^{N} P_n R_{kn}}{\sum_{n=1}^{N} R_{kn}} \tag{4}$$

Property P may be any measured or calculated molecular property, leading to the respective "property landscapes", but it could also represent a binary classification label. In the latter case, for example, with $P = 1$ representing "inactives" and $P = 2$ representing "actives", the mean $\bar{P}$ values of the node will make up a "fuzzy" classification landscape, which represents the predicted probability of a node resident to belong to either of the two classes. Note that, if the two classes are highly imbalanced, the most numerous one will implicitly dominate the landscape, and it is preferable to assign node $\bar{P}$ values reflecting the relative enrichment of the node in terms of residents of every class. To this purpose, it is sufficient to scale up the responsibilities of the items of the minority class by a factor equal to the ratio of the population sizes of the majority and minority classes.[14] This normalization, which has no impact unless the two classes are imbalanced, is systematically applied in all fuzzy classification landscapes in this work.

Any density landscape (cumulated responsibility vector) or property landscape as defined above represents a signature (vector) of a compound collection. Therefore, any compound collection no longer needs to be characterized by its individual molecules but may simply be rendered by this synthetic GTM-based characteristic projection pattern vector. The simplest synthetic descriptor of a compound collection C_i is the cumulated responsibility vector $CumR_k(C_i)$, defined as the sum of R_{kn} values of the library members $n_i \in C_i$. Therefore, a comparison of two compound libraries amounts to the extremely fast comparison of two (or several, if several viewpoints based on different landscape coloring schemes are desired) vectors, rather than the calculation of similarity scores between every possible compound pair from the libraries.

To be able to compare databases of different sizes, the cumulated responsibilities $CumR_k(C_i)$ can be normalized by division over n_i, as suggested by Fechner et al.[15] The overlap of libraries in the latent space can be estimated by means of covariance/distance measures of cumulated responsibilities: the Tanimoto coefficient, the Bhattacharyya coefficient,[14] and the Euclidean distance.

So far, GTM has been successfully used in various proof-of-concept studies in chemical space analysis, specifically by exploiting a very strong feature of GTM, that is, the ability to combine chemical mapping and compound property analysis/prediction within the same intuitive framework. Commercial compound library analysis and comparison has been successfully applied to collections with an order of magnitude of 10^6.[13] Eventually, an evolutionary procedure for selecting maps[16] of optimal polypharmacological competence was designed, and this led to "universal" maps of drug-like space. GTM is a unsupervised process that requires the specification of a frame set of compounds encoded by their molecular descriptors. These two essential degrees of freedom will be called "metaparameters" because they are of general relevance for map-building problems. In addition, a series of operational parameters (map size, number of RBFs, etc.) are specific to the GTM algorithm. The key strength of the evolutionary tuning procedure is the simultaneous combined search through both meta- and operational parameter space, which leads to the best "global" GTM construction options. For the winning maps, chosen from the very many possible ways to construct a GTM-based map, the claim of "universality" is supported by their ability to map bioactive compound sets, such as to discriminate actives from inactives, for a vast majority of the distinct and unrelated ChEMBL-reported[17] bioactivities. In other words, those maps were selected for their propensity to correctly predict the property or class of a compound by mapping it onto the "landscape" created on the basis of other examples of molecules of known property or class and then "reading" the prediction from the map, for a vast spectrum of different properties.

Goal of the study

Commercial and bioactive compounds are, however, just a tiny minority of the universe of possible compounds, a fact that, so far, has not been considered in any of the above-mentioned mapping attempts. A groundbreaking enumeration of all possible organic compounds with less than 18 heavy atoms, the GDB-17[18] database represents an opportunity to now expand GTM-driven mapping, with all the benefits emerging from property projections into this much larger realm of chemical structure. Mapping the entire set of 166 billion compounds, albeit technically feasible, would come at computational costs beyond availability. The present work focused merely on a selection of 10 million lead- and fragment-like compounds[27] from GDB-17 this subset is hereafter referred to simply as "FDB-17" unless explicit reference is made to the entire 166 billion compound collection. FDB-17, composed of molecules that are significantly smaller than most typical drugs, could be considered to represent the entire universe of lead- and fragment-like compounds. This was, first, a technical challenge, because the so-far manageable order of magnitude of 10^6 compounds was clearly not sufficient, incremental GTM construction algorithm notwithstanding.[13] The bottleneck here clearly lies at the map construction stage: the manifold needs to be calibrated on the basis of a frame set of compounds, which must be representative of the ensemble of the chemical space zone to be spanned by the map. It is very difficult to a priori predict the minimal size of such a diverse, representative frame set.

Also, by contrast with the above-mentioned construction of "universal" maps of drug-like space, the quality of which was estimated from quantitative property prediction challenges, the universe of theoretically feasible compounds would, by definition, be void of associated experimental data; the question of the relevance of the map for those chemical space regions has to be rethought. The parameter describing how close a given molecule (e.g., its corresponding point in the initial descriptor space) is to the fitted manifold is called the "log likelihood". This was now systematically used to verify whether relevant compounds that were not included in the frame set were properly described by the fitted manifold. The herein defined "coverage" criterion was, however, not necessarily a simple log likelihood cutoff, because the absolute values thereof strongly depend on the choice of the operational parameters of the map. After exploration of several working strategies (not detailed here), a self-adaptive log likelihood cutoff (function of the typical values found for frame set compounds) was employed.

Mapping by coverage-controlled evolutionary strategy

Maps of lead- and fragment-like compound spaces were thus constructed on the basis of frame sets, including subsets of FDB-17, but also of compounds of less than 18 heavy atoms from the two major databases PubChem[19] and ChEMBL.[17] The respective subsets are named PubChem-17 (11 million compounds) and ChEMBL-17 (0.1 million compounds). The inclusion of the latter compounds, some of which form compound series with associated bioactivity data, enabled the usage of the already cited polypharmacological competence criterion to select the most relevant maps.

Descriptor selection concerned the choice of the optimal ISIDA fragmentation schemes,[20–22] out of the large panel of possibilities (sequences/circular fragments/atom pairs, colored by atom label/pharmacophore type/force field type, etc.) supported by the Fragmentor software. By contrast to our typical strategy in modeling drug-like compounds (relying on fragmentation schemes having proven usefulness in previous drug-discovery-related work), the smaller sizes of compounds herein required an ab initio assessment of appropriate fragmentation schemes.

The employed "frame" sets were subsets of varying sizes of above-mentioned molecules, whereas distinct, large "coverage" subsets were used to assess the coverage criterion of the generated manifolds; the log likelihood values of coverage compounds should not be significantly worse than the typical distribution of actual frame set compounds. If the contrary applied, it meant that the manifold specifically spanned the frame set but not other relevant chemical space zones: the GTM build-up attempt as encoded by the current set of operational and metaparameters was aborted.

Eventually, map fitness was assessed by the polypharmacological competence criterion, within the limits of the available structure–activity data for compounds with less than 18 heavy atoms, as above-mentioned.

Confirmation of the generality of the obtained map

The full FDB-17, PubChem-17, and ChEMBL-17 sets were eventually projected on selected "winning" maps, to ensure that the employed coverage sets were significant and that the manifold provided a correct coverage of the entire subsets. In addition, two further 10 million compound samples from the entire 166 billion GDB-17 collection were subjected to the mapping exercise, for validation purposes. One of the two alternative sets, herein named "FDB_bis-17" was sampled according to the same protocol that led to the lead- and fragment-like FDB-17 subset.[27] By contrast, the other set, "GDB_rand-17", was a plain random subset and was thus dominated by 17-atom species, which are by far the most numerous. If the built map is relevant, then library comparison should show that the FDB-17 and FDB_bis-17 subsets have a virtually identical coverage of chemical space, whereas GDB_rand-17 will display a different space coverage pattern.

Library comparison and characterization by mapping

The ultimate topic of this work was to use the above-generated maps to directly compare the mentioned compound collections and to learn from the discrepancies between the unbiased "universe" of molecules, as represented by FDB-17, and the chemical subspace populated by the to-date existing molecules from the public databases PubChem-17 and ChEMBL-17. Several methods were used in this sense:

1) Direct comparison of the libraries, represented by their cumulated responsibility vectors on the map;

2) Construction and visual comparison of property landscapes (size, chirality, aromaticity, hydrophobicity, and (predicted) solubility) and library density profiles, and the detection of "diversity holes" (lowly or unpopulated regions specific to each library);

3) Extraction of specific or "privileged" map zones populated by above-expectation levels of compounds from one specific library, and their chemical interpretation.

Methods

Compounds and standardization

All compounds, irrespective of their source (see below) were submitted to the same "classical" standardization protocol implemented on our virtual screening web server and parallelized on the cluster of the Strasbourg High-Performance Computing Center in order to cope with tens of millions of compounds. This protocol included counterion strip-off, standardization[23] to ChemAxon basic aromatic forms and consistent representation of N-oxides (including nitro groups) with split charges, generation of the ChemAxon-predicted major tautomer[24] and major microspecies[25] at pH 7.4, and conversion to a stereochemistry-depleted representation. The herein-used molecular descriptors were not stereochemistry sensitive, so assessment of structural uniqueness after removal of stereochemical information was important to avoid fake "duplicate" descriptor lines in input files and also to cross-check for common occurrences of the same compound in several of the data sources mentioned below.

GDB-17 is a database that was formed by enumerating organic molecules of up to 17 (inclusive) atoms of C, N, O, S, and halogens, based on first principles and by starting from mathematical graphs, irrespective of pre-existing building blocks to avoid historical bias in structure selection.[18] The complete GDB-17 database contains 166.4 billion organic molecules. GDB-17 reaches into molecular sizes compatible with many drugs (367 approved drugs comprise $\leq$ 17 atoms) and typical for "lead" compounds and molecules used in "fragment-based drug design" (100 < molecular weight < 350 Da).[26] FDB-17 is a "fragment-like" subset of GDB-17 containing 10 million molecules with a limited number of functional groups and spanning evenly across molecular size and stereochemical and functional group complexity.[27] FDB-17 has intentionally limited functional group diversity to focus the structural diversity on scaffolds. For example, halogen atoms and non-aromatic double bonds are omitted because they resemble methyl and ethyl groups, respectively, in the first approximation, and therefore, they are partly redundant because of their saturated carbon analogues. Furthermore, no more than one positive and one negative charge at neutral pH value is allowed in FDB-17 molecules because multiply charged molecules tend to dominate in the fully enumerated database GDB-17. This set was employed as the source of frame and coverage subsets at the map building stage and was also the subject of the subsequent detailed chemical space analysis.

FDB_bis-17 and GDB_rand-17 represent, as already hinted, additional subsets of the 166 billion total collection of feasible compounds, separate from the 10 million compounds above. Herein, they were used for map validation purposes. FDB_bis-17 was built by following the above-mentioned "flat" random picking protocol, in as far as possible: at low N values, all the existing compounds with N heavy atoms were already admitted into the reference FDB-17 database. By contrast, GDB_rand-17 is the result of unbiased random picking from the 166 billion compound set: it is therefore statistically composed of a majority of 17-atom compounds.

PubChem[28] and ChEMBL[17] are public databases of organic ligands and other organic compounds, with biological activity annotations. The Pubchem-17 and ChEMBL-17 size-limited subsets (as extracted from the web servers in March 2017) were used as representatives of real organic fragment-like and lead-like molecules, which are of relevance, or at least thought to be of relevance (interesting enough to be synthesized and tested), in various stages of drug discovery. This actual or assumed relevance is the main source of human knowledge-induced bias in characterizing these selections, relative to the plain, unbiased FDB-17 subsets. Pubchem-17 contains up to 1 million compounds (after removal of molecules also reported in ChEMBL; see paragraph about standardization above). There are some 0.1 million compounds in the smaller ChEMBL-17 set. All of these compounds have some kind of activity annotation, but only the compound series of sufficient size (>90 molecules), sharing the same activity measure, and containing a minimum of 30 active compounds could be used for cross-validated map coloring/property prediction.

Frame, coverage, and selection sets: Frame and coverage set compounds were randomly selected. The evolutionary map tuner supported various frame set options in order to pick the best suited one: a pool of frame set candidates was provided by a selection of 100 000 compounds, randomly taken form the entire pool of >40 million compounds above (FDB-17, FDB_bis-17, GDB_rand-17, PubChem-17, and ChEMBL-17). Out of this pool, the evolutionary map tuner was enabled to choose either one of the halves (50 000 each) or one of the tiers (33 000 compounds each) of this pool. Coverage sets included distinct randomly selected subsets of 50 000 compounds from each of the five above-mentioned libraries and were not included in the frameset pool. Coverage was thus monitored with respect to 0.3 million compounds.

Selection and external validation sets were structure–property series that were used, by cross-validated projection on a map, to create a property landscape able to predict the properties of left-out compounds and, hence, to quantitatively validate the map. Maps simultaneously supporting high-quality landscapes for a large number of selection sets associated with various properties were preferred by the evolutionary map tuner (herein, active versus inactive classification landscapes were used specifically). Selection sets were created by an automated data curation procedure applied to the PubChem database:

1) For each PubChem assay, the results were retrieved under csv file format. Only lines reporting half-maximal inhibitory concentration (IC_{50}) or inhibition constant (K_i) values were kept. If there were less than 200 such entries, the assay was no longer considered. No analysis of the actual IC_{50} or K_i values was undertaken, but the PUBCHEM_ACTIVITY_OUTCOME field was taken as such, in order to assign compounds into either the "active" or "inactive" category.

2) By using the reported PubChem compound identifications (CIDs) as a key, assay results were matched against the list of standardized PubChem compounds with less than 18 heavy atoms. Note that several distinct PubChem CIDs might correspond to the same standardized structure. Only PubChem-17 structures receiving an unambiguous flag as either "active" or "inactive" with respect to this assay were kept at this step.

3) Next, it was checked whether the assay-specific standardized structure–activity class table contained (a) more than 30 "active" entities and (b) at least twice as many "inactive" entities as "active" ones. For large sets with more than 5000 entries, a random subset of 5000 entries was picked. If the initial set contained less than 1000 "active" compounds, then the "inactive" ones were specifically discarded until a final set size of 5000 was obtained; otherwise, representatives of both classes were discarded on a pro rata basis.

All the hereby extracted data sets are provided (as SMILES-activity class label text files, PubChemID.smi_class) in the Supporting Information. Next, a modelability study of the resultant structure–activity class sets was undertaken: all these sets were encoded under the form of the various herein-considered ISIDA fragment descriptor sets (see below) and subjected to an evolutionary Support Vector Machine (SVM) classification model tuner.[29] If the latter succeeded in finding at least one combination of molecular descriptor scheme and SVM parameters supporting a model with threefold cross-validated balanced accuracy (BA) above 0.65, the set was declared "modelable" and, thus, eligible as either a map selection set or an external validation set. The 28 selection sets, randomly picked out of the pool of modelable sets, were mapped/cross-validated for every GTM manifold that the evolutionary tuner attempted to build, and their cross-validated classification proficiency entered the map fitness score calculation (map fitness is the mean of the selection-set-specific cross-validated BA values, empirically penalized by $0.5 \times$ the standard deviation of set-specific BA values). External sets were mapped and cross-validated on the final map only; they have no influence whatsoever on its construction. The Supporting Information also features the list of modelable sets, as well as the one of randomly picked selection sets, as a list of PubChem IDs versus the cross-validated balanced accuracy scored in the modelability study.

Molecular descriptors

Two software packages, ISIDA[20] and MOE,[30] were used as the source for the herein-employed molecular descriptors. The former served to generate candidate descriptor sets, out of which to pick the optimal choice for manifold construction, as detailed below. The latter served to calculate descriptors representing calculated properties (logP, log solubility) and other interpretable molecular descriptors that were of direct chemical interest, to allow these to be rendered as property landscapes on the map, for library analysis. MOE descriptors were computed by using the MOE v.2015.10 software.[30]

The ISIDA Fragmentor[20] software may generate a vast choice of fragment count descriptors, based on various fragmentation schemes (from "fine-grained" circular fragments, featuring bond order information, to "coarse" atom pair counts) and employing various atom coloring strategies as a means to enable capturing specific chemical information. Meaningful fragmentation schemes yielding descriptors that were proven useful in various drug-design-related endeavors may not be automatically appropriate for the current study focusing on the universe of significantly smaller lead- and fragment-like compounds. Therefore, a systematic scan of possible fragmentation schemes was undertaken as follows:

1) Given a randomly picked subset of 5000 compounds (from the five concerned collections), a systematic search was performed. The search looped over 1) all possible fragmentation types (sequences, circular fragments); 2) considered fragment coloring schemes[21] (by atom type, by pharmacophore feature, and by CVFF[31] force field type); 3) toggles of Formal Charge, Atom Pair, and All-Paths options (please refer to the Fragmentor manual[20] for details); and 4) minimal and maximal fragment sizes to be enumerated. For each combination of the above choices (more than 800 were scanned), the fragmentation scheme with maximal fragment size that resulted in descriptor vectors of dimensionality below 5000 was memorized.

2) The above-memorized fragmentation schemes were all systematically applied to the frame set pool of 100000 compounds. In the default "open-ended" fragmentation mode, the ISIDA Fragmentor added newly encountered substructures that are not present in the initial 5000 molecules as novel elements to the descriptor vector. If the dimensionality of the vector exceeded 7500, the fragmentation scheme was discarded. Otherwise, the fragmentation scheme was declared as a valid choice for the evolutionary map tuner.

All other relevant sets for map building (i.e., coverage and selection sets) were subjected to fragmentation according to the above-validated (72 per total) fragmentation schemes by using the -StrictFrg Fragmentor option and no longer admitting previously unseen substructures into the descriptor vector. The series of 72 considered candidate descriptor schemes is available (as an ISIDA Fragmentor command line file) upon request. For practical reasons, the complete five databases were only fragmented (again, in -StrictFrg mode) according to the scheme selected by the evolutionary map tuner as the most appropriate, in view of their projection on the optimal GTM manifold.

Technical enhancements of the evolutionary parameter choice strategy: The coverage criterion

In this work, the previously designed evolutionary map selection scheme was updated to include a "coverage control" step (the new block in the red frame of Figure 1).

After the construction of the manifold according to the parameters defined in the "chromosome" (parameter vector), frame set compounds fitted on the manifold were sorted by their log likelihood.[32] The threshold value topping the log likelihood values of the bottom 5% frame set molecules less well matched by the manifold (logL5) was recorded. Next, above-mentioned "coverage sets" were projected on the manifold and the percentage of "badly" mapped compounds (in the sense of log likelihood < logL5) was monitored. This fraction should never exceed an empirical threshold, herein set to 15%. Otherwise, map construction was stopped and the attempt counted as a failure of the evolutionary procedure: the current manifold did not extrapolate well into chemical space regions not covered by the frame set. If each of the (herein, ten distinct) coverage sets passed the test, the manifold was assessed in terms of predictive power, by means of the "classical" three-fold cross-validated cycle of projection–coloring–prediction of selection sets, that is, compound sets with known biological activities.

Map analysis tools

Tools used for a posteriori map analysis included state-of-the-art quantitative library comparison indices (overlap scores), visualization of density-modulated class and property landscapes, and a novel "zooming" procedure (see below), in order to highlight specific structural patterns associated with given map locations.

Quantitative library comparison

Quantitative indices that can be derived from a GTM model in order to characterize any compound collection all rely on the cumulated responsibility vector of the collection; this is a central concept in GTM theory and, as such, has already been discussed in the Introduction. The similarity of two libraries may thus be quantitatively measured by applying different similarity scores calculated with $CumR_k(C_i)$ vectors, for example, the Tanimoto coefficient, Bhattacharyya coefficient,[14] and Euclidian distance.

Landscapes

The algorithms employed to generate density-modulated landscapes have already been described in various previous publications,[13,16,32] and will not be revisited here. Color and density
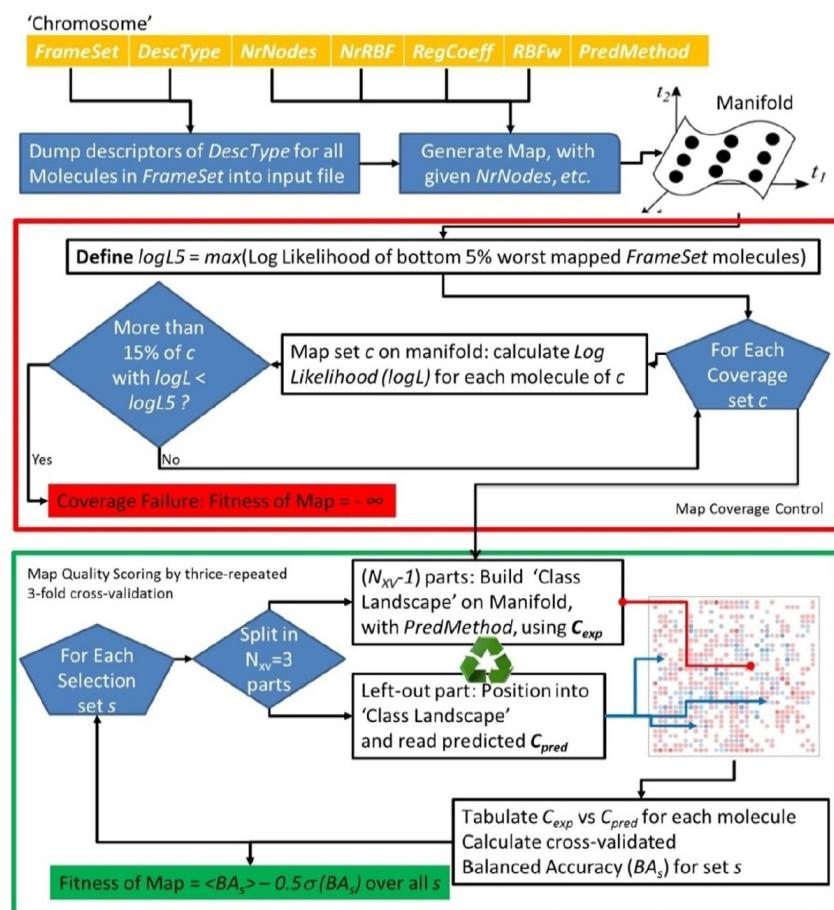
Figure 1. Evolutionary map selection scheme.

in the inter-nodal continuum are obtained by polynomial interpolation (on the basis of the four closest nodes). The color intensity was modulated by the total compound density at a node, so that color rendering can be tuned from completely transparent (if the total density is below a minimal threshold) to full saturation (if the density exceeds a maximal threshold), with range-wise interpolation in between. Practically, when comparing landscapes of collections of different sizes, it is more useful to modulate the intensity by the relative compound density, that is, the ratio between the actual $CumR_k(C_i)$ value in a node, and its default value of (library size)/(total number of nodes). Thus, nodes harboring identical *fractions* of a library will light up at identical color intensity.

GTM hierarchical zooming

Map resolution, that is, the number of nodes, is a tunable parameter that can be chosen such as to ensure meaningful, cross-validating property landscapes for selection sets. However, such sets of biologically tested compounds are intrinsically small. The evolutionary algorithms showed that a limited number of nodes is sufficient to ensure cross-validated separation of actives versus inactives. Yet, the number of items that can be mapped a posteriori with GTM is unlimited. For each node, the subset of compounds that is significantly associated

with it ($R_{kn} >$ threshold) may become too numerous for a simple visual inspection or common substructure analysis to be sufficient to highlight the common structural patterns characterizing it.[33] Tino et al.[34] suggested the use of several maps built on the same descriptor set, in which one of them is the main map obtained for the entire initial data set and the others are for the subsets extracted from zones of the first map that are too dense. This approach was called "hierarchical GTM" and was applied to some "toy" data sets that contained no more than 3000 items. Herein, we applied this technique to "zoom" into the projection of 21 million compounds from FDB-17, PubChem-17, and ChEMBL-17 in order to perform more exhaustive structural analysis.

The idea of hierarchical GTM, or zooming, is to extract the molecules from one node or a cluster of nodes (e.g., an area of nine nodes) in which the researcher is interested and to build a new GTM manifold just for this subset, with the same operational parameters and descriptor set. Moreover, this process is iterative, so, if necessary, it can be repeated on the built submap. Thus, multilayer zooming can be applied to big data, when we deal with millions of compounds, to avoid costly calculations. If the number of molecules does not exceed some constant number (e.g., 100 molecules), a usual structure analysis method, such as scaffold analysis,[35] can be applied.

Results and Discussion

Map generation by the evolutionary algorithm

Evolutionary optimization, after 741 generations, led to a best map that fulfilled the coverage criterion for all coverage and selection sets, with the following parameters: square root of the number grid nodes 29×29, number of RBF centers 18×18, width of RBF 0.4, regularization coefficient 3.236, size of frame set 100000 compounds, and descriptor space IA-FF-FC-2-3 (sequences of 2 and 3 atoms, colored by their CVFF force field types and including formal charge information;[20] this represented a sparse, 6142-dimensional vector). The propensity of the map to discriminate between actives and inactives in the 28 selection sets, calculated by means of a threefold, quite aggressive cross-validation procedure (see below), resulted in a mean balanced accuracy criterion of 0.645.

External validation and consistency check of winning map

Once the map manifold is generated with the above-mentioned so-far best parameter setup, it is ready to project any other chemical compound.

Manifold quality control

The first implicit external validation was to entirely project all five libraries on the map and to estimate the fractions of compounds not properly "covered" by the manifold, in the sense of the herein-defined coverage criterion. Figure 2 shows the distribution of log likelihood values in the five libraries, in parallel with that of the frame set molecules. None of the libraries display any extreme increase in the fraction of compounds with low log likelihood values. All collections easily pass the imposed threshold maximum 15% of compounds not properly covered by the manifold. In fact, their distributions are strikingly identical to that of the frame set compounds, and the fraction of compounds not properly covered (with log likelihood $<$ logL5) is always much closer to 5%, which is the fraction of

Figure 2. Log likelihood distributions within considered databases and the frame set.

frame set compounds at which logL5 was defined. Thus, 50000 randomly picked representatives serving as the frame set successfully spanned the relevant chemical space containing more than 40 million lead- and fragment-like compounds.

Validation of the map propensity to support predictive classification models for external structure–activity sets not seen at the map-tuning stage

For each of the 24 external structure/activity class data sets, the propensity of the map to separate actives from inactives was assessed by following the same threefold cross-validation scheme that was used, with selection sets, during evolutionary optimization. Two thirds of the external set members were projected on the map, in order to "color" it by class; nodes in which the number of "residing" active compounds exceeds the statistical expectation were colored as "active" nodes, whereas the others were labeled as "inactive". Eventually, the remaining unmapped third of the current data set was mapped onto the colored manifold, and each of its compounds was associated with a class, depending on the class(es) of the neighboring node(s). The herewith predicted classes were compared to the actual experimental labels of the compounds, which allowed, after cycling over the data set tiers that were kept out, cross-validated BA scores to be estimated for each of the 28 and 24 targets associated with the selection and validation data sets, respectively. As can be seen from Figure 3, 90% of the targets achieved balanced accuracies above 0.6, which is a robust result. Notably, in the context of the relatively small data sets (which is a consequence of limiting size to 17 atoms and less), the aggressive threefold cross-validation was perhaps too challenging a choice; five- or tenfold schemes would yield better statistics. There was no significant difference in distribution of BA values between selection sets and external sets, respectively; the latter were as accurately accommodated on the map exclusively selected on the basis of the former.

Map consistency check: Comparison of FDB-17 to the alternative samples FDB_bis-17 and GDB_rand-17

Even though these processed chemical libraries represent, to our knowledge, the largest to be successfully mapped so-far by using GTM technology, it cannot be denied that tens of millions of compounds represent a vanishingly small part of the 166 billion GDB-17 compounds. Furthermore, FDB-17 is a specially designed subset of fragment-like compounds: How representative is it? To what extent would alternative samples of 10 million compounds from GDB-17 cover the same chemical space as FDB-17? Is a sample of 10 million compounds meaningful, or does it completely ignore entire chemical space zones? If so, then alternative samples, stochastically visiting different subspaces, should be weakly overlapping. Is the degree of overlap sensitive to the library sampling protocol? The alternative samples FDB_bis-17 and GDB_rand-17 were introduced to address these questions, with the former being produced by following the same protocol[27] as FDB-17 and the latter being a plain random subset of GDB-17.
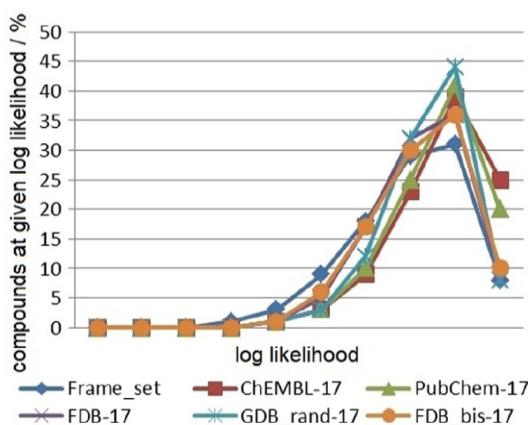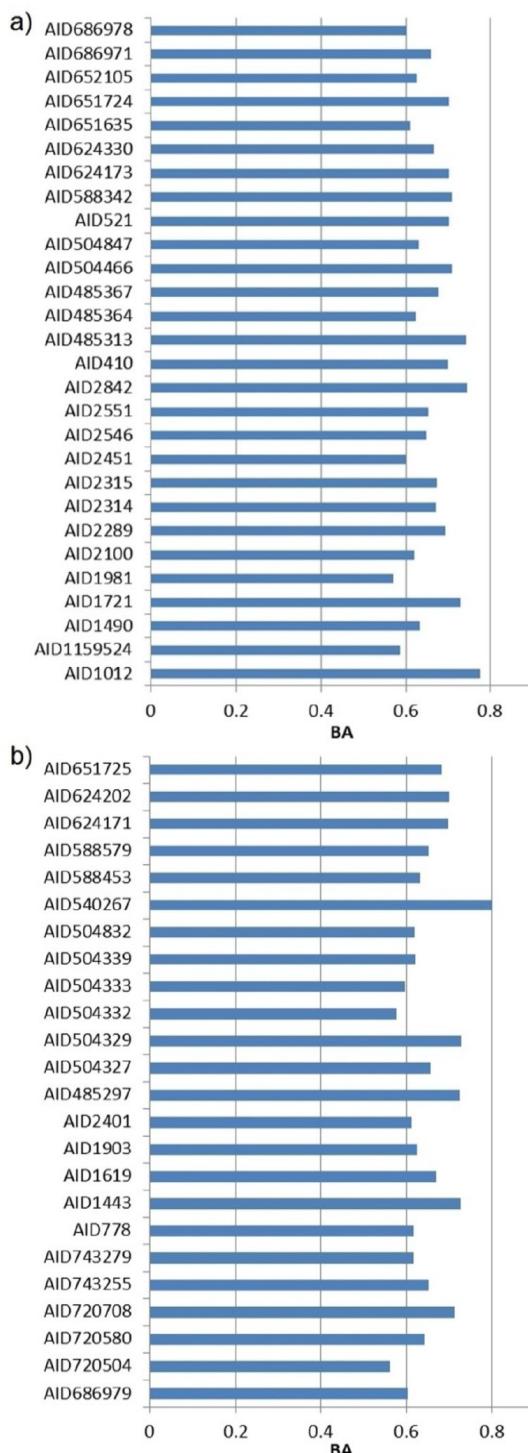
Figure 3. Cross-validated balanced accuracy (BA) proving the effectiveness of active versus inactive separation on the GTM manifold for a) 28 selection targets and b) 24 validation targets. For an explanation of the target identities, see Tables S1 and S2 in the Supporting Information.

Figure 4. Fuzzy classification landscapes showing GDB subset overlaps: a) FDB-17 (blue) versus FDB_bis-17 (red); b) FDB-17 (blue) versus GDB_rand-17 (red); c) density-modulated color samples.

Fuzzy class landscapes, with FDB-17 compounds arbitrarily assigned into a "blue" class and the alternative sample labeled as "red", are shown in Figure 4. Herein, the zones predominantly populated by the members of either library are colored by the extreme colors of the spectrum. Residents of dark blue and red areas of the landscape are FDB-17 members to an extent of $>90\%$ and $<10\%$, respectively (extreme colors should not be interpreted as a complete absence of members of the other library). In areas that are equally well populated by both, the color is in the intermediate spectral range of orange/yellow/green. The latter tones completely dominate the FDB-17 versus FDB_bis-17 landscape, which proves that a subset of 10 million compounds sampled according to the same protocol does indeed reproducibly cover the relevant chemical space. FDB-17 and FDB_bis-17 are, grossly, redundant sublibraries. The subtle differences do not impact on this conclusion because, as already mentioned, the extremely small compounds were all co-opted into FDB-17 and are therefore absent from FBD_bis-17. FDB-17 and GDB_rand-17, however, are not the same thing; as expected, the latter is dominated by the most numerous 17-atom species. Furthermore,[27] highly complex molecules in GDB-17 (rich in stereocenters and heteroatoms) were excluded from FDB-17. The library comparison by mapping perfectly matched our expectations; it confirmed the representativity of FBD-17 for the fragment-like universe of compounds and displayed the expected dependence on sublibrary selection protocols.

Quantitative assessment of library similarity

In terms of quantitative the inter-library distance scores based on the respective cumulated responsibility vectors, ChEMBL-17 clearly stood out because of its much smaller size (by nearly two orders of magnitude). Its raw, not normed, cumulated responsibility values are significantly smaller, which implicitly triggered important Euclidean distances in comparison with FDB-17 and PubChem-17. The dissimilarity metrics reflecting the covariance scores of two vectors ($1-S_{\text{Tanimoto}}$ or $1-S_{\text{Bhattacharyya}}$) were, however, less affected by the amplitude differences. Nevertheless, the conclusion that can be drawn from Figure 5 is the same, irrespective of the used metric: Pub-Chem-17 has a significantly distinct density pattern with respect to the "unbiased" FDB-17, in spite of their comparable sizes, whereas ChEMBL-17 displays an even more marked discrepancy, as expected on the basis of its small size. ChEMBL and PubChem are rather strongly correlated, which is not sur-prising because they actually share many compounds. Yet, at only 70% of Tanimoto similarity, they are not redundant collections; PubChem harbors much more primary high-throughput screening data of large (sometimes combinatorial) compound collections, whereas ChEMBL prioritizes publications reporting small homogeneous compound series.

In-depth library comparison

Comparative chemical space overlap

The three libraries were compared by means of fuzzy class landscapes (Figure 6), by following the principle already mentioned in the map consistency check section above. The "predominance of one collection over the other needs to be understood in the context of the relative library sizes; for example, a map zone might be occupied by a significant fraction of ChEMBL-17 compounds and be very sparsely populated by PubChem-17 molecules.[15] Yet, because the latter are 100 times more numerous, it may be that, in absolute numbers, there will be more PubChem-17 than ChEMBL-17 residents in that area (see Figure 6a). Predominance is thus defined after normalization with respect to total library sizes, which is relevant only for ChEMBL-17, because PubChem-17 and FDB-17 are equally large.

Overlap landscapes are perfectly suited to highlight FDB-17 diversity holes that are nevertheless populated by PubChem
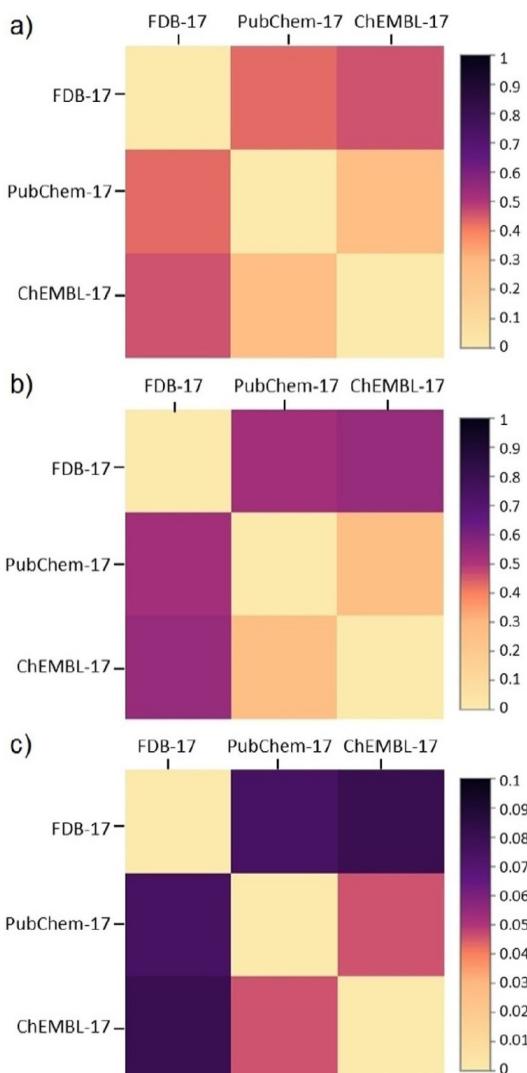
Figure 5. Heat maps representing similarities between three libraries on the two-dimensional GTM map by using the GTM-based: a) Bhattacharyya coefficient ($1-S_{\text{Bhattacharyya}}$);[14] b) Tanimoto coefficient ($1-S_{\text{Tanimoto}}$); and c) Euclidean distance as a metric.
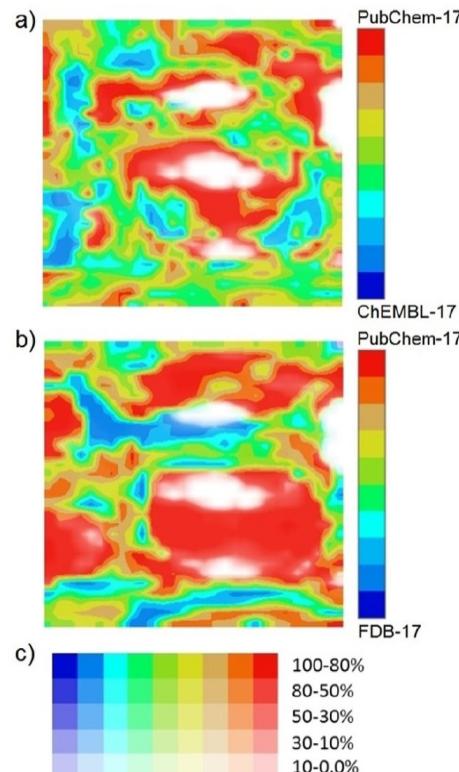
Figure 6. Fuzzy classification landscapes monitoring library overlap: a) ChEMBL-17 (blue) versus PubChem-17 (red); b) FDB-17 (blue) versus Pub-Chem-17 (red); and c) density-modulation of color intensity; white spots are devoid of compounds.

compounds (Figure 6 b). These holes are clearly visible in the combined perspective of an overlap map (FDB-17 versus Pub-Chem-17). Regions in red are well populated in PubChem17 but correspond to underpopulated or outright empty zones in FDB-17. At the same time, there are some regions populated principally by FDB-17 (in blue). In Figure 7, some examples are shown for the particular red and blue areas. Therein, the FDB-17 dominant species are primary, secondary, and tertiary amines and nitriles. Red zones, which correspond to PubChem-17 compounds, mostly contain N-substituted amino acids, esters containing aromatic nitro derivatives, and triazoles connected to thiophenes by an aliphatic linker (L). Note that the structures associated to each zone have a clear structural signature that defines them as chemical families, even though this signature is not necessarily a single common scaffold, in the strict sense employed by scaffold-analysis software. Similarly to observations in previous works,[33] the underlying structural patterns were not predefined but emerged from the information in the molecular descriptors as a result of the mapping process.

Empty regions in FDB-17 might be explained by two main reasons:
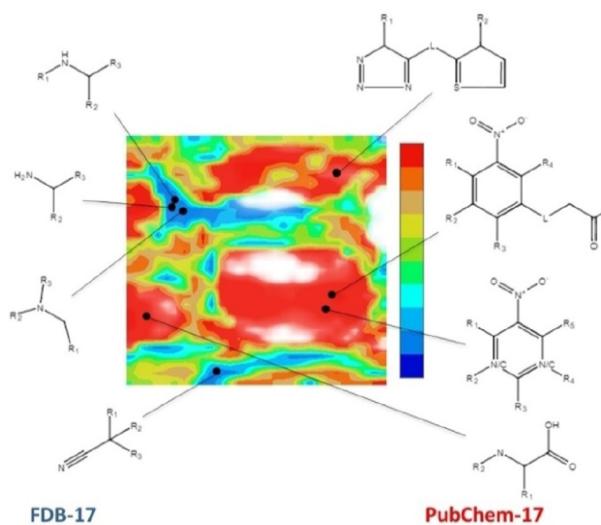
1) FDB-17 represents $<0.1\%$ of the entire collection of feasible compounds. Even though it was shown to be representative of the fragment-like universe in terms of chemical space overage, this does not imply that it will contain every single possible chemotype.

2) The FDB-17 selection process[27] filtered out several chemical elements (Cl, Br, I, and F) present in the entire GDB and various combinations of functional groups that would systematically violate the fragment-likeness "rules of three". GDB-17 itself is a "complete" enumeration only within the frame established by the rules of its enumeration protocol. It does not contain P or Si atoms, in contrast to PubChem and ChEMBL, or "black-listed" reactive groups that may, in certain chemical contexts, be nevertheless stable enough for medicinal chemistry purposes.

Hierarchical GTM and chemotype detection

Direct analysis of compound subsets associated to specific map zones by displaying associated structures and identifying their common patterns, as exemplified above, is no longer feasible if the subset size exceeds the magnitude order of thousands. Zooming into these subsets by remapping them on a detailed GTM model, dedicated to that chemical subspace zone, may help the user to browse through the otherwise overwhelming wealth of chemical data. This concept was exemplified for one mixed area, populated by both FDB-17 and PubChem-17 compounds (Figure 8). It contains 9 nodes, and 650 000 compounds were extracted by the following rule: the molecule is taken into account only if the sum of its responsibilities in the chosen area is higher than 0.8.

Based on a "local" frame set randomly chosen from the 650 000 selected residents of the zone, a zone-specific GTM manifold was built with the same operational parameters as the initial one (see above). This means that the 29×29 square mesh of nodes initially used to cover the entire fragment-like chemical space is now "refocused" on the above-selected zone.

On the much finer scale of the obtained local map, the new class landscape (Figure 8, middle map) displays a much more effective separation of the PubChem-17 and FDB-17 residents.

Figure 7. Some examples of PubChem-17 and FDB-17 molecules extracted for the pure PubChem-17 (red) and FDB-17 (blue) zones.

FDB-17   PubChem-17

21 million compounds   650 000 compounds   2 500 compounds
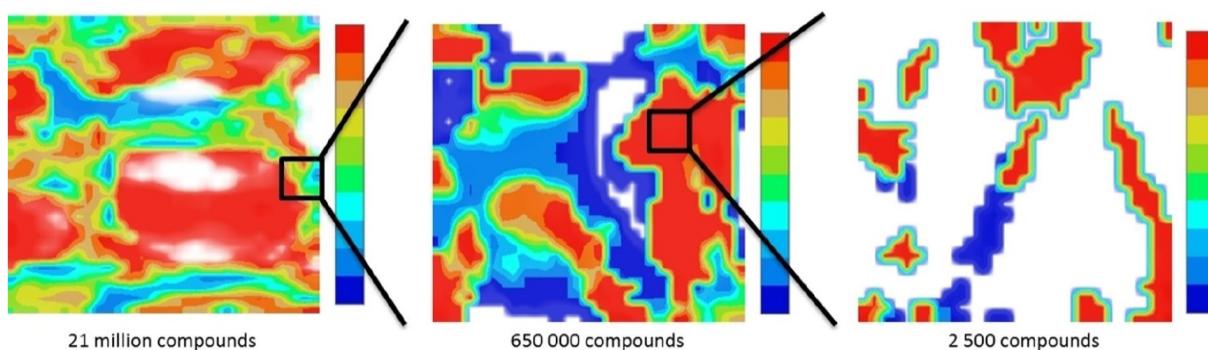
Figure 8. Hierarchical GTM zooming of the chemical space occupied by FDB-17 (blue) versus PubChem-17 (red, on the fuzzy classification maps). For each handpicked zone on a map, a local GTM model with identical parameters is refitted for the local residents only.

110

A second zooming iteration on a still heavily populated area dominated by PubChem-17 (2500 compounds) led to a clear separation of the compounds (right-hand map).

The minority of FDB-17 compounds that required two successive zooming stages in order to be separated from PubChem-17 co-residents form the central blue spot on the right-hand map. They were parsed by the Scaffold Hunter tool[35] and shown to be various furan derivatives (Figure 9). In addition, two structures were extracted from the center of this area (Figure 10), and a similarity search in PubChem (as implement-
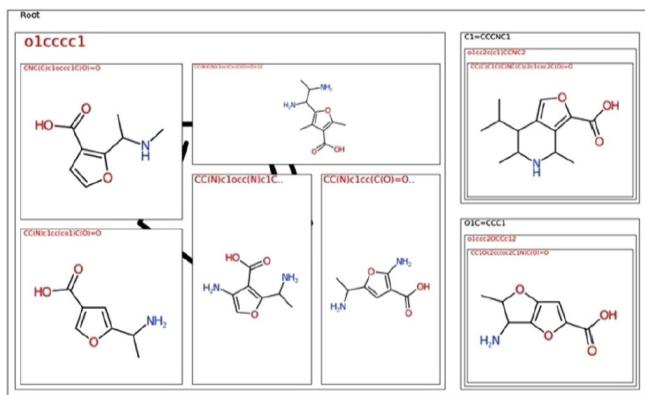

Figure 9. Scaffold Hunter view of the compounds extracted from the FDB-17/PubChem-17 map as a result of hierarchical GTM zooming. These compounds are unique for the FDB-17 data set and are not presented in PubChem-17.
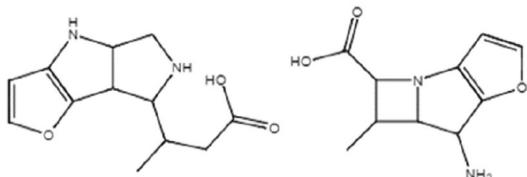

Figure 10. Examples of compounds extracted from the pure FDB-17 area on the zoomed FDB-17/PubChem-17 GTM map. These compounds have no similar neighbors in the PubChem database.

ed on the PubChem server) retrieved no results for them. In other words, the PubChem-17 co-residents were not similar to these compounds, in terms of the PubChem web server definition of molecular similarity. The compounds are, in this sense, original, which does not preclude the fact that, according to different similarity measures, they may have many near neighbors in PubChem-17. In terms of pharmacophore patterns, for example, they are zwitterionic amino acids, a quite ubiquitous motif that explains why two zooming iterations were needed to eventually separate this class. This shows how hierarchical GTM may help for in-depth big data analysis, because there is no single map offering both complete chemical space coverage and detailed separation of relevant chemotypes.

Property landscapes

Property landscapes are a synthetic way to highlight similarities or differences in terms of the distribution of various chemical properties in libraries.

The landscapes of the number of heavy atoms for ChEMBL-17 and PubChem-17 (Supporting Information Figure 11) are similar. However, PubChem-17 possesses a more distributed excess of the top size molecules (strong red) than FDB-17. This is the result of two sources of bias. On one hand, PubChem is subject to a bioactivity-driven selection bias: very small compounds, which cannot possibly be strong ligands (in spite of putatively high ligand efficiency, per heavy atom), are rare in PubChem. On the other, the current selection of FDB-17 was specifically sampled with the goal of achieving a balanced number of participants for all compound sizes: it has voluntarily oversampled compounds in the middle of the heavy-atom number range, whereas the entire collection of feasible compounds is, for obvious combinatorial reasons, completely dominated by 17-atom structures.

The landscapes for the entropy of the element distribution (the a_ICM index of MOE) in a molecule are similar for ChEMBL-17 and PubChem-17, whereas FDB-17 contains less diverse structures, in the sense of a bias in favor of carbon chains over functional groups. Elementary chemical stability rules prevent arbitrary concatenations of heteroatoms to be enumerated in GDB-17, and carbonated chains are statistically predominant. By contrast, in libraries of existing compounds, reactivity- and property-yielding functional groups were voluntarily introduced, at an important energetic cost. This subtle bias is clearly highlighted by the maps.

In contrast to PubChem-17, FDB-17 is characterized by a large number of molecules with two or more chiral centers (Figure 11). It is statistically easy to obtain branched chiral chains by theoretical enumeration but difficult to separate diastereomers after synthesis. Human selection of compounds is clearly biased against chirality, and this can be clearly read from the maps as well.

The distribution of the log of the octanol/water partition coefficient (as estimated by the associated MOE descriptor serving as the "property" for map coloring) is almost the same, except for two small regions of very low log$P$ in PubChem-17. The property landscapes of the log of the aqueous solubility (log$S$, again, as predicted by MOE, see Figure S1 in Supporting Information) are also rather similar, with some increase of alleged insoluble compounds in PubChem. This is intriguing: Do chemists prefer to synthesize insoluble molecules amongst the feasible ones? This could be tentatively explained by comparison of the property landscapes for the number of aromatic atoms and chiral centers.

The spread of aromatic compounds over the chemical space shows that PubChem-17 is clearly richer in this respect than FDB-17. By contrast, the latter is significantly richer in chiral compounds, as already mentioned. This clearly highlights the preference of chemists for aromatic, "flat" achiral compounds, simply for reasons of synthetic facility. However, the former are also reputed to be more insoluble (or, at least, the MOE de-
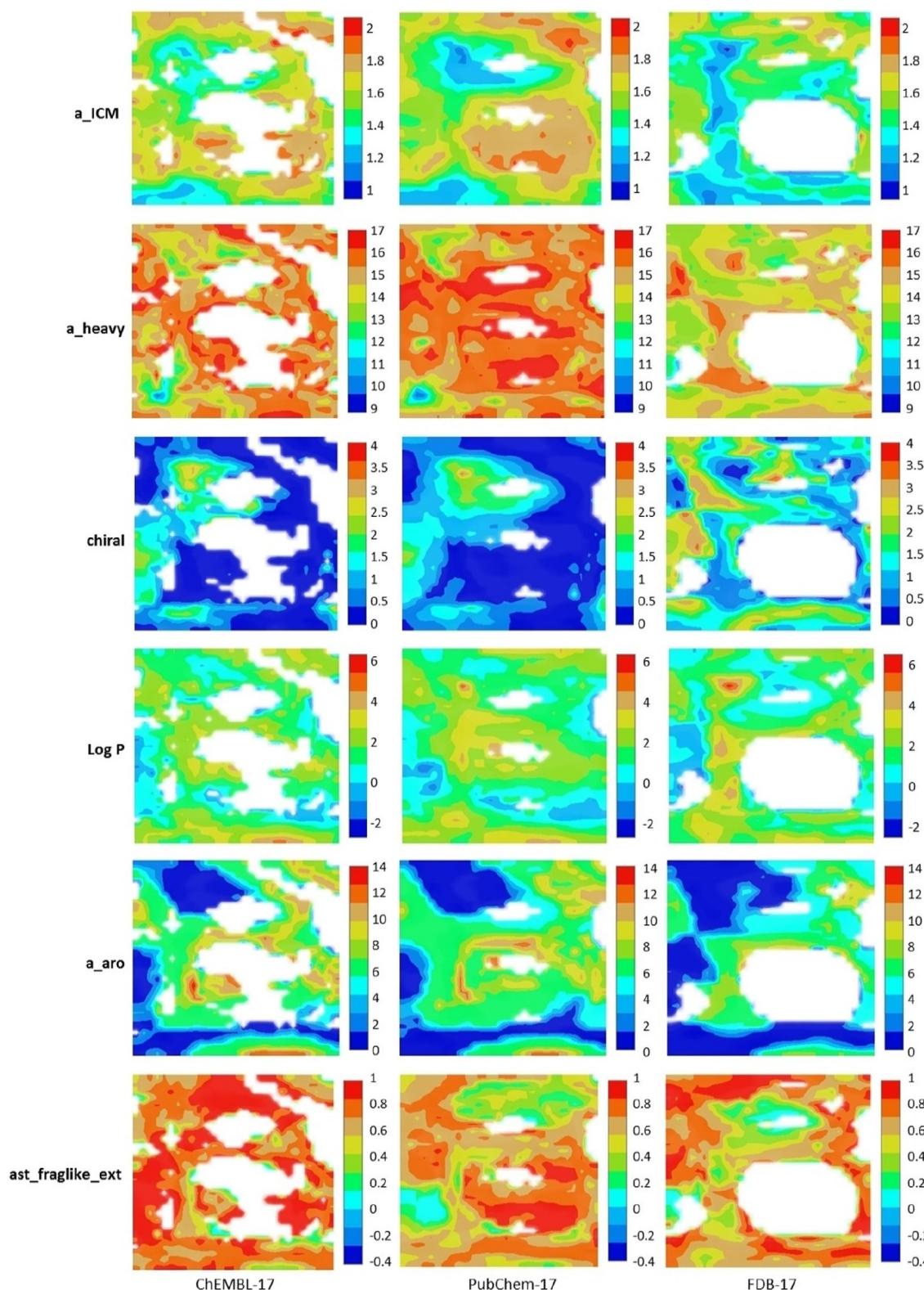
Figure 11. GTM property landscapes for the a_ICM (entropy of the element distribution in a molecule), a_heavy (number of heavy atoms), chiral (chirality), logP, a_aro (number of aromatic atoms), and ast_fraglike_ext (ASTEX Fragment-like Status) distributions.

scriptor estimating the log of solubility used here as a mapped property was probably trained to account for such a correlation).[36] Thus, the observed increase in aromaticity and (predicted) decrease in solubility are probably related.

With regard to the ASTEX Fragment-like Status as calculated by the MOE package,[30] the ChEMBL-17 database contains more fragment-like compounds than PubChem-17, which is not a surprise, considering that ChEMBL-17 was created in

order to accumulate such compounds. The fragment-likeness of the FDB-17 compounds is also quite high (this independent assessment indirectly confirms the effort to orient the FDB-17 library toward fragment-like molecules), and this library also covers some high lead-likeness zones that are not well represented in ChEMBL-17 or PubChem-17.

Thus, the findings with respect to these directly available global structural features or easily predictable physicochemical properties, as they can be "read" from the maps, perfectly matched the already reported trends by Reymond's team[18,27] from distribution histograms, but they are more intuitively rendered and, furthermore, directly searchable. Chemotypes associated with zones of specific property values, or property value differences between libraries, can be easily highlighted as such or after hierarchical zooming. Such in-depth analysis is, however, beyond the scope of this paper.

Conclusions

This work concerns the visualization and analysis of a large data set of more than 40 million compounds from GDB-17, PubChem-17, and ChEMBL-17 and represents, to our knowledge, the most complete GTM-based analysis of fragment-like chemical space to-date. The goals of the study were multiple:

1) To prove that a significant and robust map of fragment-like chemical space that can successfully accommodate tens of millions of species can actually be built, in spite of the technical constraints on the maximal frame set size and relative sparseness of structure–activity data (providing selection sets for the evolutionary map optimization). To this purpose, the evolutionary map building strategy has been updated to include a "coverage check" step, to ensure that the size-limited frame sets used to build the map were actually representative for the entire chemical space populated by compounds that were ignored at the map-building stage. Eventually, the algorithm supported the generation of a useful map, based on force-field-type-colored ISIDA atom pair counts. The map has a robust propensity to separate actives from inactives, in as far as the rather sparse affinity data associated with lead-like compounds of this size can tell us. Unlike previous work that benefitted from the full information richness of entire drug-like compound collections, this exploration of the universe of lead-like compounds was less well supported by criteria based on structure–affinity relationship success. Nevertheless, map-supported classification models for external structure–activity sets, associated with completely new biological properties, were shown to cross-validate as well as the sets on the basis of which the map was selected. The map was then shown to properly accommodate the entire collection of 40 million compounds, theoretically enumerated and real compounds alike. All of these compounds returned log likelihood values (an intrinsic measure of their distances to the manifold on which they are projected) comparable to the ones of the frame set items used to fit the manifold. It was thus proven that a frame set of the order of 10^4 com-

pounds suffices to span a manifold that properly covers the entire fragment-like chemical space.

2) To assess how far the targeted selection of 10 million compounds in the fragment- and lead-like FDB-17 subset of the 166 billion compounds in the theoretical GDB-17 represents a significant sample of fragment-like molecules. Map-based comparison of FDB-17 to the similarly sampled but distinct FDB_bis-17 showed that the extraction protocol of FDB-17 is reproducible: if repeated, it leads to an equivalent library of similar chemical space coverage. By contrast, a plain random sample of 10 million GDB-17 compounds showed a slightly different signature in terms of coverage, which validated the fact that coverage analysis is sensitive to the nature of the employed library subsetting protocol. Therefore, coverage analysis based on the present map is validated as a library comparison tool.

3) To actively use the map as a library comparison tool in order to highlight systematic differences and biases of real-world molecules in PubChem-17 and ChEMBL-17 versus the universe of all possible fragment-like species, as represented by FDB-17. The findings here were in perfect agreement with the results obtained from classical property distribution histogram comparisons, but their rendering was much more intuitive and could be directly linked to examples of the chemotypes underlying every zone of interesting property values. The comparisons clearly evidenced that current synthetic chemistry is biased in favor of aromatic compounds, relative to the background of putative feasible species. However, aromaticity per se is not necessarily wanted (and is potentially associated with low solubility) but is rather a "refuge" from the much higher synthetic effort needed to tackle the production of polycyclic chiral molecules.

4) To identify some specific structural patterns, proper to some libraries and absent from others (diversity holes). A hierarchical zooming technique was applied, which actually builds a new map for items residing in the interesting area of the main (initial) map. This allowed, for example, the highlighting of FDB-17-specific structures that are not yet present (as such or in the form of structurally close analogues) in public databases. Conversely, the study also revealed diversity holes in FDB-17, accounted for by the systematic exclusion of certain atom types/functional groups that are present in PubChem-17.

Conflict of interest

The authors declare no conflict of interest.

[1] D. Horvath, *Methods Mol. Biol.* **2011**, *672*, 261–298.

[2] S. J. Lusher, R. McGuire, R. C. van Schaik, C. D. Nicholson, J. de Vlieg, *Drug Discovery Today* **2014**, *19*, 859–868.

[3] I. V. Tetko, BIGCHEM—Big Data in Chemistry, **2016**, http://bigchem.eu.

[4] M. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York, **1990**.

[5] A. Schuffenhauer, P. Ertl, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

[6] T. Varin, A. Schuffenhauer, P. Ertl, S. Renner, *J. Chem. Inf. Model.* **2011**, *51*, 1528–1538.

[7] B. Zhang, M. Vogt, G. M. Maggiora, J. Bajorath, *J. Comput.-Aided Drug Des.* **2015**, *29*, 937–950.

[8] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, **2002**; c) G. H. Dunteman, *Principal Components Analysis*, Sage, Newbury Park, CA, **1989**.

[9] D. Horvath, M. Lisurek, B. Rupp, R. Kühne, E. Specker, J. von Kries, D. Rognan, C. D. Andersson, F. Almqvist, P.-A. Enqvist, A.-L. Gustavsson, N. Remez, J. Mestres, G. Marcou, A. Varnek, M. Hibert, J. Quintana, R. Frank, *ChemMedChem* **2014**, *9*, 2309–2326.

[10] T. Kohonen, *Self-Organizing Maps*, Springer, Heidelberg, **2001**.

[11] C. M. Bishop, M. Svensén, C. K. I. Williams, *Neurocomputing* **1998**, *21*, 203–224.

[12] C. M. Bishop, M. Svensén, C. K. Williams, *Neural Comput.* **1998**, *10*, 215–234.

[13] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2014**, *55*, 84–94.

[14] H. Gaspar, G. Marcou, D. Horvath, A. Arault, S. Lozano, P. Vayer, A. Varnek, *J. Chem. Inf. Model.* **2013**, *53*, 3318–3325.

[15] N. Fechner, G. Papadatos, D. Evans, J. R. Morphy, S. C. Brewerton, D. Thorner, M. Bodkin, *Bioinformatics* **2013**, *29*, 523–524.

[16] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087–1108.

[17] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

[18] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[19] PubChem Ver. 2010, US National Institutes of Health, https://pubchem.ncbi.nlm.nih.gov/.

[20] ISIDA, Laboratoire de Chemoinformatique, Strasbourg, France, **2012**, http://infochim.u-strasbg.fr/spip.php?rubrique41.

[21] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855–868.

[22] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191–198.

[23] ChemAxon Ver. 2008, ChemAxon, Budapest (Hungary), **2008**, https://chemaxon.com/products/chemical-structure-representation-toolkit.

[24] ChemAxon Ver. 2007, ChemAxon, Budapest (Hungary), **2007**, https://docs.chemaxon.com/display/docs/Tautomer+Generation+Plugin.

[25] ChemAxon Ver. 2013, ChemAxon, Budapest (Hungary), **2013**, https://docs.chemaxon.com/display/docs/pKa+Plugin.

[26] K. Heikamp, J. Bajorath, *J. Chem. Inf. Model.* **2013**, *53*, 791–801.

[27] R. Visini, M. Awale, J.-L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 700–709.

[28] W. A. Warr, *Mol. Inf.* **2014**, *33*, 469–476.

[29] D. Horvath, J. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.

[30] MOE, Ver. 2015.10, Chemical Computing Group, Montreal, QC (Canada), **2015**.

[31] A. T. Hagler, E. Huler, S. Lifson, *J. Am. Chem. Soc.* **1974**, *96*, 5319–5327.

[32] a) H. A. Gaspar, P. Sidorov, D. Horvath, I. I. Baskin, G. Marcou, A. Varnek in *Frontiers in Molecular Design and Chemical Information Science—Herman Skolnik Award Symposium 2015: Jürgen Bajorath, Vol. 1222* (Eds.: R. J. Bienstock, V. Shanmugasundaram, J. Bajorath), American Chemical Society, Washington, DC, **2016**, pp. 211–241; b) H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *Mol. Inf.* **2015**, *34*, 348–356.

[33] K. Klimenko, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2016**, *56*, 1438–1454.

[34] P. Tino, I. Nabney, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *24*, 639–656.

[35] Y. Hu, D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2016**, *59*, 4062–4076.

[36] F. Lovering, J. Bikker, C. Humblet, *J. Med. Chem.* **2009**, *52*, 6752–6756.

6.2   Conclusion

The Generative Topographic Mapping (GTM) method was trained and applied to analyze and compare large public chemical databases. It was shown that ChEMBL-17 is very similar to PubChem-17 since the first one is a part of the PubChem database. At the same time, virtually generated FDB-17 differs significantly (Soergel distance to PubChem-17 is about 0.55). The GTM class landscape demonstrated that there are some areas on the map populated only by PubChem-17 compounds. Scaffold analysis showed that the chemotypes allocated in these areas were discarded by the authors of the FDB-17 collection due to the rules used to gather the last one.

An example of the application of hierarchical GTM zooming was also demonstrated to increase the map resolution. With the help of this technique, a mixed zone populated equally by PubChem-17 and FDB-17 compounds was zoomed. The multilevel zooming discovered some chemotypes presented in FDB-17 but missed by the PubChem database. Thus, GTM becomes an attractive tool that can be efficiently applied for novelty analysis.

Finally, the data sets were compared in terms of molecular properties (LogP, chirality, number of aromatic atoms, etc.). It was shown that FDB-17 is richer in terms of chirality and it is more homogenous in terms of heavy atoms' types in a molecule (more or less the same atom types are used in the virtual structures).

7 Chemical Library Enrichment

7.1 Introduction

Structural library enrichment is an important task for the pharmaceutical industry. The number of hits in screening campaigns depends on drug-likeness and diversity of the underlying screening set. To be efficient in drug-discovery, the existing screening pool needs to be regularly updated to include new chemotypes.

One can suggest two different scenarios of the screening pool enrichment with new chemical matter: computer-aided enumeration of virtual structures under some constraints (e.g. molecular weight, LogP, etc.), or selection of existing structures from an external database. Recently, several attempts were made to create a workflow for an efficient molecular *de novo* design [2, 78, 99–101]. However, synthetic feasibility of virtual structures including synthetic routes and optimization of reaction conditions still needs to be assessed. The second scenario is more practical because new structures selected as a result of a comparison of two data sets (a reference set and an external set) do exist and can be purchased or synthesized following the reported in the literature procedure.

Different approaches to chemical database comparison were reported so far: cell-based clustering [102], pairwise distance analysis [103], and some dimensionality reduction methods (Principle Component Analysis or PCA [27], Self-Organizing Maps or SOM [104], Generative Topographic Mapping or GTM [45]) providing with the visualization support.

GTM is a method of choice in this study because of its clear advantage over PCA and SOM approaches.

Recently we demonstrated that GTM represents an efficient tool for comparison of large chemical libraries FDB-17 and PubChem-17 [50]. The hierarchical GTM zooming technique [11] was successfully applied in [50] in order to analyze the chemotypes of molecules populated selected zones and to highlight the scaffolds present exclusively in FDB-17.

In this study, the zooming technique was automatized and coupled to a Maximum Common Substructure (MCS) extraction protocol ("AutoZoom" tool). The developed tool was used for the enrichment of the in-house collection of Boehringer Ingelheim (further on referred to as the "BI Pool") by the compounds from the commercial Aldrich-Market Select (AMS) database. A drug-likeness and an activity profile of selected AMS compounds against 749 biological targets were assessed using the ChEMBL data-driven predictor based on Universal GTMs [10, 58]. The paper reporting these results has been recently accepted in J. Computer-Aided Molecular Design.

7.2   Data

Boehringer Ingelheim (BI) is steadily committed to innovation in medicinal chemistry and is hence interested in new compounds featuring new scaffolds. At the same time, new structures have to be synthesizable and should have the potential to be active.

As a basis in this work, we used the in-house collection of drug-like compounds provided by BI (BI Pool) which contained more than 1.7M structures. The source for novel compounds was the publicly available Aldrich-Market Select (AMS) collection of purchasable   compounds   containing   more   than   8.2M   items (http://www.aldrichmarketselect.com). The data was standardized by ChemAxon's standardizer tool using a list of rules, such as aromatization, removing stereo labels, the standard representation of N-oxides including nitro group, etc.[105]

7.3 Method

The computational workflow consists of three parts. First, the mapping of AMS chemical space was undertaken by calibrating a pertinent GTM manifold, followed by projection of entire AMS and BI Pool collections. Then, the hierarchical GTM zooming was performed for selected areas of the map followed by MCSs extraction. The most of interest represented some zones exclusively populated by AMS compounds. The latter was extracted and profiled using universal GTMs described in our previous papers [10, 58]. To this purpose, the publicly available virtual screening webserver of the Laboratory of Chemoinformatics (http://infochim.u-strasbg.fr/webserv/VSEngine.html) was employed. In addition, simple molecular properties, like LogP, number of H-bond donors and acceptors, molecular weight, and TPSA, were computed using ChemAxon's JChem engine [81].

7.3.1 GTM training

The Generative Topographic Mapping (GTM) method relates the data points positions in the initial N-dimensional space and in the latent 2D space. The GTM algorithm is described in a range of publications [4, 6, 45, 50]. Briefly speaking, GTM injects a 2D hypersurface (*manifold*) into a multidimensional data space populated by a set of representative items (the Frame Set, FS). The algorithm fits the manifold to the FS data distribution by changing the positions of Radial Basis Function centers and, hence, maximizing the data log-likelihood (*LLh*). At the next stage, the data points are projected on the manifold followed by the manifold unbending. Each compound in the latent space is represented by a vector of normalized probabilities (*responsibilities*) computed in the nodes of a square grid superposed with the manifold. In turn, the entire data set can be characterized by a vector of cumulative responsibilities. This enables the user to perform an efficient data sets comparison as well as QSAR/QSPR studies [6, 45, 49].

In our early study [50], the frame set compounds were randomly selected from large chemical libraries. Here, a FS containing 25K AMS compounds of controlled diversity

(featuring no two compounds more similar than a given threshold) was prepared. To measure the dissimilarity, Soergel distance [106] basing on Morgan fingerprints [107, 108] of radius 4 was computed. FS compounds are expected to represent a non-redundant, representative "core" of spanned chemical space. They are not subjected to any other specific constraints, meaning that any state-of-art molecular descriptor/dissimilarity metric can be equally well used for selection.

The GTM manifold was trained using an incremental algorithm described by H. Gaspar et al.[5] The parameters were taken from the previous study [50]. The experience of previous projects [9, 50, 109] showed that the usage of ISIDA descriptors is a good choice for GTM training. The initial descriptor space features ISIDA counts of sequences of 2 and 3 atoms, colored by their CVFF [94] force field types and including formal charge information (IA-FF-FC-2-3) [80, 95]. Fragmentation of the FS compounds produced 6142 distinct fragments. However, the vast majority thereof is sparsely populated: only 798 terms were considered for actual manifold construction (the descriptors for which standard deviation over the FS compounds exceeds 2% of their value range width). This (or closely related) fragmentation schemes were often selected by evolutionary [48] map tuning procedures [50, 58]. Other adopted map parameters include resolution (841 nodes), the number of RBFs (324), the regularization coefficient (3.236), RBF width (0.4), and incremental block size (10K compounds).

When the Expectation-Maximization algorithm used to train the manifold has achieved a certain level of convergence ($LLh_{new} - LLh_{prev} \leq 0.001$), the entire data was projected, and the compounds considered as out of Applicability Domain (the structures positioned far away from the manifold) were removed. To do so, a new strategy for GTM Applicability Domain (*AD*) identification was suggested where a Gaussian is fitted to the FS compounds distribution minimizing the root mean square error. Once the fitting is done, the LLh threshold is determined as the LLh value with the highest population (peak) minus three Gaussian widths ("3σ" rule, Figure 31).

For visualization and analysis purposes, property and fuzzy class landscapes are used to "color" the map. To this goal, the mean class/property value in each node is taken as responsibility-weighted means of class labels/property values of resident items [6]. In consequence, areas of interest (for example, clusters of nodes exclusively populated by AMS compounds) can be easily highlighted.
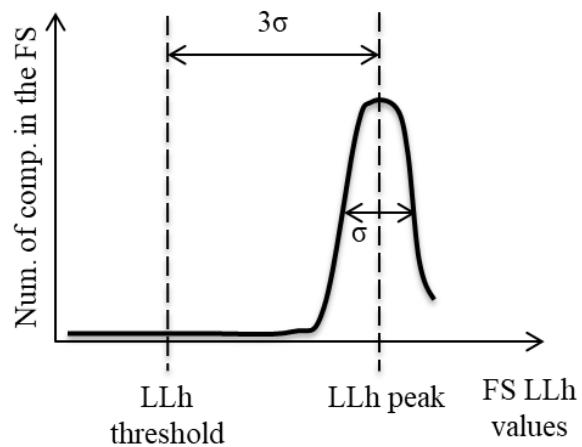

Figure 31. GTM Applicability Domain is identified by log-likelihood threshold $LLh_0 = LLh_{peak} - 3\sigma$. Here, LLh_{peak} and σ are, respectively, a position and with of a Gaussian function which fits the LLh distribution.

7.3.2 Zooming

GTM landscape analysis is the following step in the library comparison process. The goal is to bind a certain chemotype to a particular area on the map. In simple cases, map zones (square clusters of nine nodes) do indeed contain structurally quite homogeneous populations of residents. If so, it is straightforward to search for common scaffolds or maximum common substructures (MCSs). However, if too many compounds (e.g. more than 1000 items) reside in one zone, searching for common scaffolds or MCSs is not efficient. Therefore, when the algorithm detects highly populated zones, zooming is automatically applied. For this purpose, the compounds for which the sum of its responsibilities within the zone is higher than 0.95 are selected and used as frame set source for the fitting of a new GTM manifold (using the same setups as those of the global map).

For this purpose, the FS - of minimal 1000, but maximal 10% of the local compound pool size - is randomly selected. The "submap" is likewise checked for the zones with a population exceeding 1000 items. If necessary, the procedure is repeated (multi-level zooming). If a zone contains less than 1000 compounds, it will be analyzed as such, without further zooming.

7.3.3 Maximum Common Substructure (MCS) searching

The responsibility patterns (RP) method has been used to identify the shared underlying features (scaffolds, substructures, pharmacophore patterns) for a chosen area on the map [49, 65]. Compounds sharing the same RP will typically share some common structural features that are further manually processed to annotate the map. This is a tedious and error-prone task. As an alternative, it is proposed here to exploit the MCS search to automatically highlight shared features. Our solution is based on ChemAxon's JChem engine [81].

The problem of MCS searching for a set of compounds was already discussed earlier by Hariharan et al.[110]. The authors showed that in some situations, the intersection of pairwise MCS search is empty or results in small, non-specific substructure, while the molecules in a given set share large and complex substructures. The problem is that such a common substructure of a compound set is not the maximum common substructure of any compounds pair. As a solution, Hariharan et al enumerated all maximal cliques for each pair of molecules, and then intersected the generated lists. The so-called multi-MCS is the largest of the identified substructure that is common to all compounds in the set.

However, when the molecule set is very large, the idea to return a single multi-MCS does not work anymore. In this case, we aimed at identifying lists of frequent substructures. In our approach, an arbitrarily selected structure in the list of N items is compared to the other N-1, resulting in N-1 connected MCS (Figure 32). Since we are working with large sets, this already results in a large list of chemically relevant substructures, although the list

might not be exhaustive. Additionally, a size filter keeps only the MCS covering at least 30% of the heavy atoms in both structures of a pair. Then, duplicate MCSs are removed from the list and sorted according to their occurrence in the list. The most frequent MCS is selected. Structures featuring the selected MCS are removed from the list, and a new iteration is started. In contrast with the previous scenarios, the new strategy returns a list of MCSs which is more relevant in the context of Big Data.
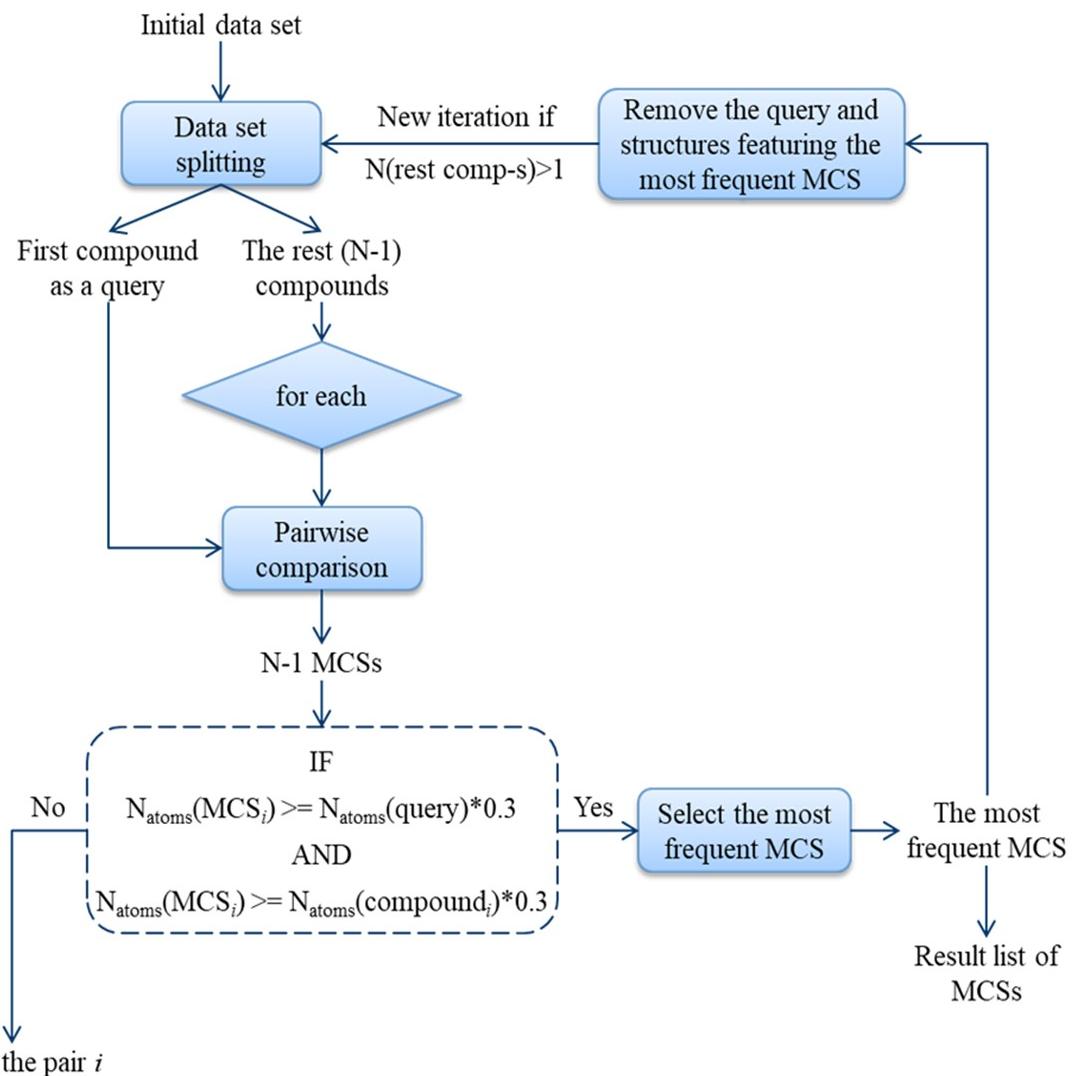

Figure 32. MCS extraction protocol.

The entire workflow is implemented in Python3 language using NumPy [111, 112] and Plotly [113] libraries. When the MCSs absent in the BI pool were found, the structures

containing these MCSs were retrieved from the AMS collection, and their biological profile was predicted using previously developed universal GTMs [13].

7.3.4 Virtual Profiling of Novel Compound Candidates

The approach supported on the public property prediction server (http://infochim.u-strasbg.fr/webserv/VSEngine.html) utilizes consensus prediction of the activity class (active or not) of a compound with respect to 749 biological targets for which structure-activity records found in ChEMBL v.24 were considered to be sufficiently robust to provide for meaningful activity class landscapes on the seven distinct "universal" GTMs of drug-like space. Each candidate is iteratively projected onto each of the seven universal maps [58], and its projection is then placed in the context of the map-specific activity landscapes of each of the 749 targets. For each target, the compound is assigned a probability to belong to the "active" class, which corresponds to the relative excess of "active" population in its residence zone (or zero if the target-specific data from ChEMBL do not occupy at all this residence area). Herewith, a consensus probability $\bar{P}$ to be active on a target is taken as the mean of the seven predictions of the complementary universal maps. This mean is penalized by the standard deviation of the seven estimations, to signal that mutual agreement of predictions enhances the trustworthiness of consensus:

$$P_{corrected} = \bar{P} - \sqrt{\frac{1}{6}\sum_{i=1}^{7}(P_i - \bar{P})^2} \qquad (7.1).$$

where $\bar{P}$ – the mean probability over the 7 universal maps; P_i – the probability to be active on a map i; $P_{corrected}$ – the corrected consensus probability.

The tool supports processing of up to a few million compounds, operating on the HPC cluster of the University of Strasbourg, in order to return a virtual profile matrix of input compounds × 749 predicted consensus probabilities.

7.4 Results and Discussion

To train the GTM manifold, a Frame set (FS) of 25K compounds needed for the manifold construction was diversity-picked from the AMS library with the dissimilarity threshold equal to 0.4. At the next stage, the log-likelihood threshold LLh = -2501.52 was determined as described in Figure 31 in order to delineate the GTM Applicability Domain (AD). With this threshold, 95.5% of the FS items passed the AD criteria (23.9K compounds out of 25K). Figure 33 visualizes the distribution of the FS compounds over the map. The density landscape shows that the FS covers most parts of the map, and the maximal population of compounds in each node doesn't exceed 5% of the entire FS.
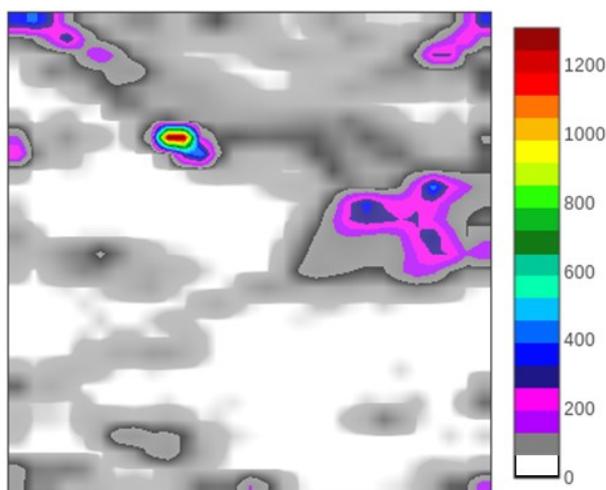

Figure 33. Frame set density landscape. Here, the white space means non-populated areas. Both color intensity (transparency) and color choice are associated to local density values (red areas have no transparency).

To understand how the two chemical collections relate to each other, they were projected on the map and rendered as individual density landscapes and a fuzzy classification landscape, respectively (Figure 34). Some 94.1% of the BI Pool and 95.8% of the AMS collections passed the LLh threshold which means that the frame set extracted from AMS is diverse enough to describe both databases. We assume that as far as the frame set is diverse enough to span the relevant chemical space zone, its explicit composition is of rather little importance – a recurrent conclusion in all our GTM studies, notably the creation

of "universal" maps [9] where a frame set of the order of 10K random compounds was shown to suffice for the coverage of ChEMBL chemical space and supporting robust predictive activity models for hundreds of independent targets. The density landscapes in Figure 34a-b show that the libraries are globally similar since they both mostly reside in the same areas. However, there are some areas where the AMS library has a strong presence and even fills some "holes" of the BI Pool. In the fuzzy class landscape, AMS-dominated areas are dark red (Figure 34c).
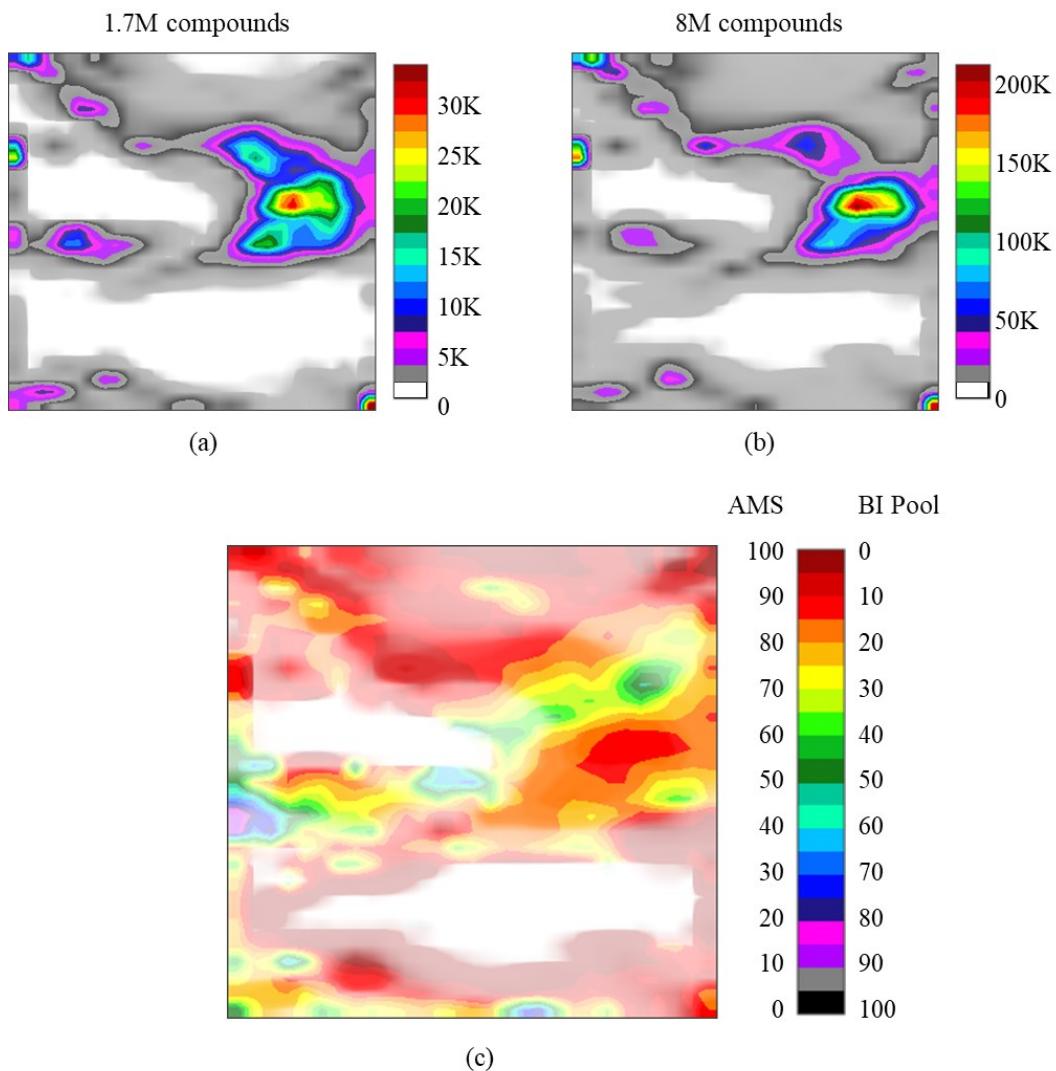

Figure 34. BI Pool vs AMS comparison: (a) BI Pool density landscape, (b) AMS density landscape, and (c) fuzzy class landscape. Here, the white space means non-populated areas, and the transparency corresponds to the density.

The dark-red areas can serve as a source of new chemotypes for the BI collection. However, even mixed zones might also contain some structural patterns not shared by both libraries [50]. To investigate this possibility, 187 zones were checked whereby 151 zones were zoomed (the maximal level of zooming was up to 4). The procedure took approximately 7 days using 48 CPUs. An example of multi-level zooming is given in Figure 35.
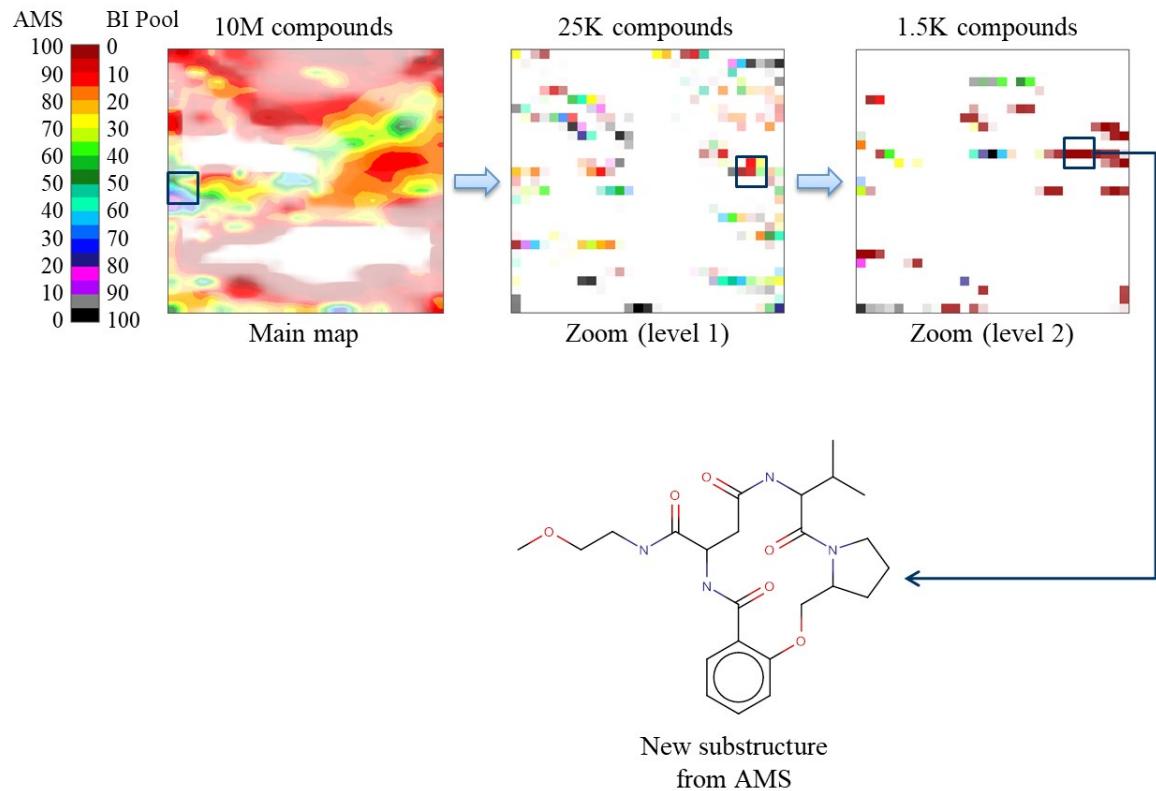
Figure 35. An example of zooming analysis. Here, a new substructure from AMS collection was discovered using 2-levels zooming. The white space means non-populated areas, and the transparency corresponds to the density of population.

In total, more than 222K substructures were processed. This set included some 45.5K MCS present only in AMS collection. More than 401K structures containing these MCSs were extracted from the AMS collection and projected onto the map. The density landscape with some examples of the most popular new AMS substructures is given in Figure 36.

Comparing the density landscape from Figure 36 and the fuzzy class landscape from Figure 34, we see that most of the compounds came from the areas where AMS dominated. At the same time, several thousands of structures also came from mixed areas (green and yellow). This was achieved by the application of zooming.
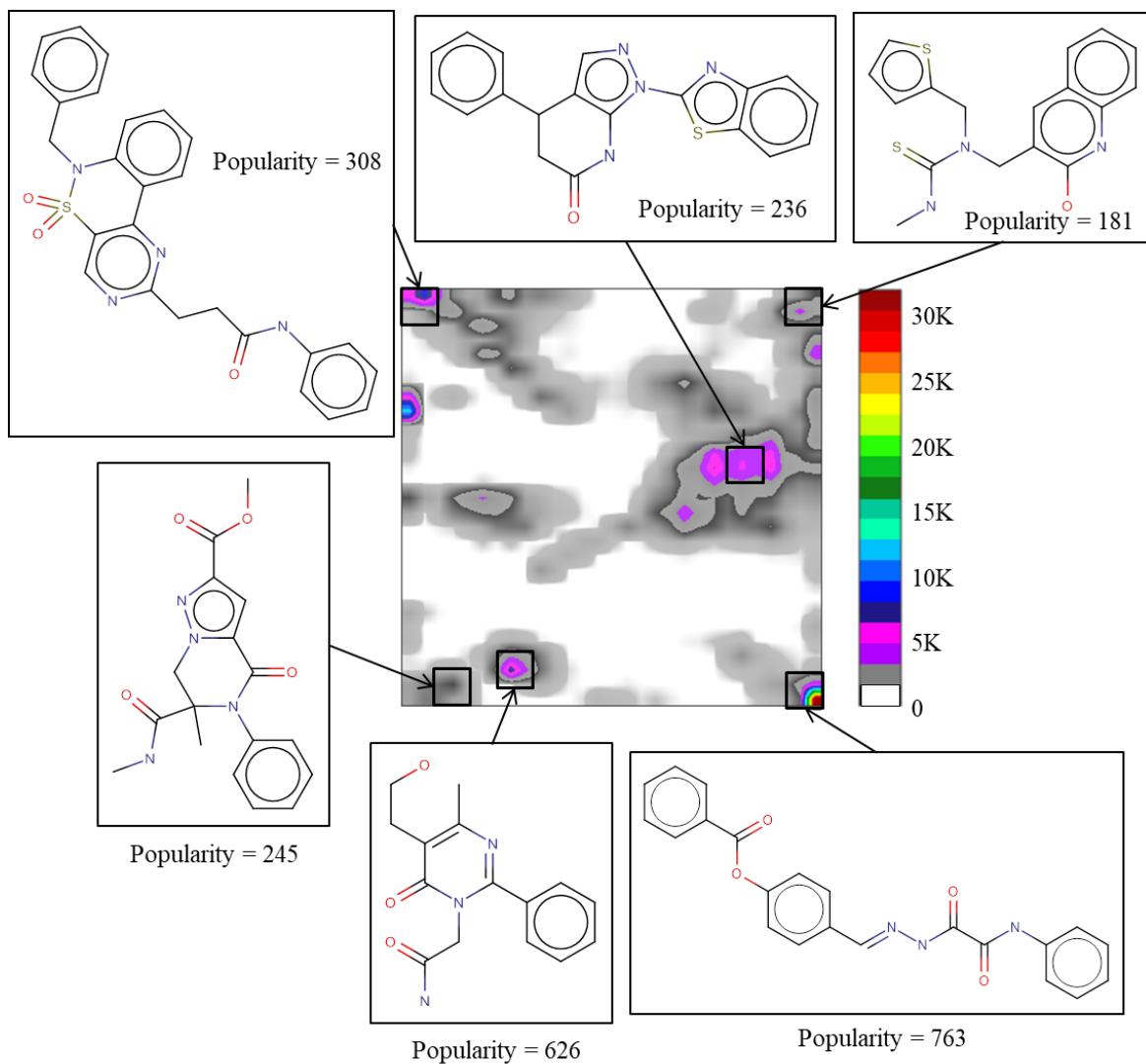
Figure 36. Density landscape for the new 401K structures. Here, several most popular (within the particular zone) new substructures are shown. The number of corresponding compounds is presented here as a popularity score.

To check the drug-likeness of the extracted structures, simple molecular properties, namely the number of H-bond donors and acceptors, LogP, molecular weight, and TPSA were computed (Figure 37).
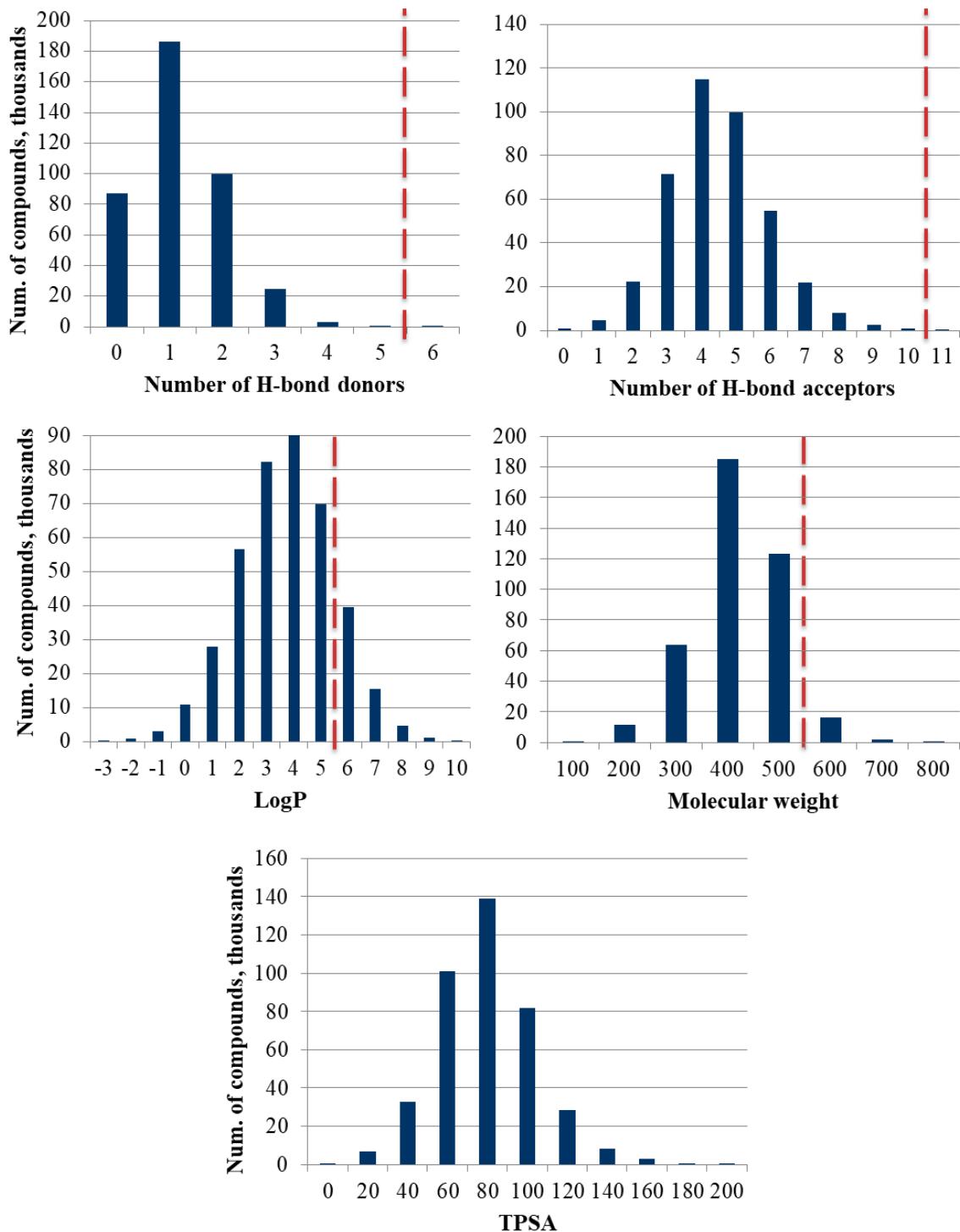
Figure 37. Histograms represent the number of H-bond donors and acceptors, LogP, molecular weight, and Topological Polar Surface Area (TPSA) computed for the extracted 401K AMS compounds. Here, the red dashed line represents Lipinski's thresholds [13].

Accordingly to Lipinski's rule of five [13], most of the extracted compounds can be classified as drug-like. These structures were also virtually profiled against 749 ChEMBL targets. 109.5K compounds were predicted as active against at least one out of 749 ChEMBL targets with a probability score $P_{corrected}$>0.5.

About 1.2K compounds out of it were predicted according to equation (7.1) as active with $P_{corrected}$>0.8 and passed BRENK [114], PAINS [115] and NIH [116, 117] filters. The four examples with the highest corrected consensus probability to be active in one of the CHEMBL targets are shown in Figure 38, where the compounds are predicted as active against Photoreceptor-specific nuclear receptor (CHEMBL4374), Cholecystokinin B receptor (CHEMBL3508), Muscarinic acetylcholine receptor M4 (CHEMBL317), and Pyruvate dehydrogenase kinase isoform 1 (CHEMBL4766) [93].

The type of the source of the structures (a chemical online store) allows us to say that these compounds are potentially synthesizable or even purchasable (the real synthesizability depends on a supplier since some suppliers just claim that it can be synthesized if a client asks). This and the number of predicted actives demonstrate that the revealed substructures are new and useful for the pharma company. Also, it supports the statement that GTM is a powerful method for the efficient library comparison and enrichment (in terms of structural diversity).
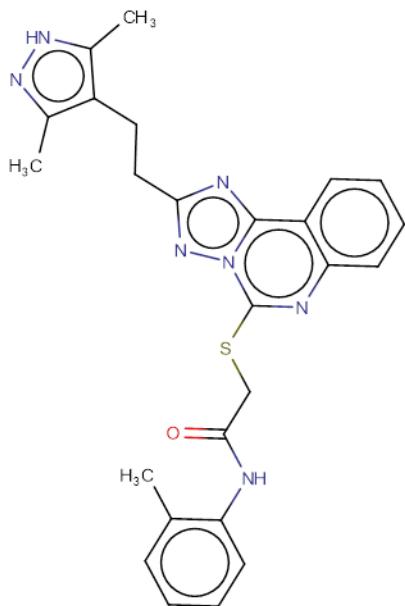
AMS structure ID 41419963
Target: Photoreceptor-specific
nuclear receptor
Probability to be active is 93%
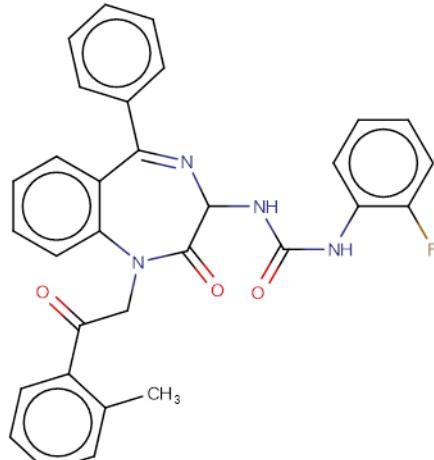
AMS structure ID 414778192
Target: Cholecystokinin B receptor
Probability to be active is 92%
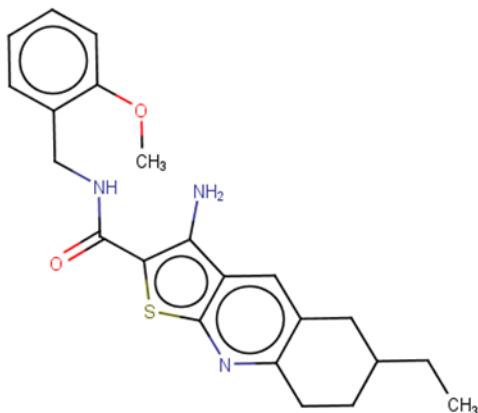
AMS structure ID 29149085
Target: Muscarinic acetylcholine
receptor M4
Probability to be active is 91%
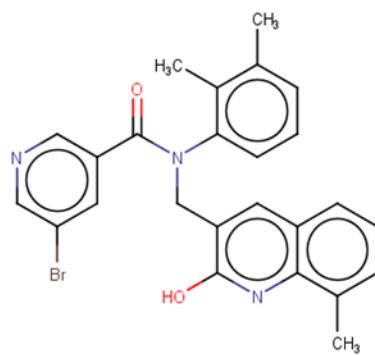
AMS structure ID 48316039
Target: Pyruvate dehydrogenase
kinase isoform 1
Probability to be active is 90%

Figure 38. Examples of structures predicted as actives and taken from the extracted 401K AMS compounds. Here, the probability to be active returned by the web server is computed according to equation (7.1).

7.5   Conclusion

Generative Topographic Mapping was enabled to provide automated hierarchical analysis of large libraries, by means of the herein described "AutoZoom" tool. This integrates automated zooming and a new MCS extraction protocol and was successfully applied to diversify the in-house collection of Boehringer Ingelheim (BI). Some 45.5K substructures were found to be absent in the BI collection. The corresponding structures (401K items) were checked for Lipinski's rule compliance and classified as drug-like. In addition, they were virtually profiled against 749 ChEMBL targets. More than 1.2K compounds were predicted active against different targets with a corrected consensus probability (removing a standard deviation) higher than 80%. The discovered structures were recommended to the company to be imported as novel chemical matter that would be useful in diversifying the in-house collection.

8 Software Development

Several tools were developed during this PhD project. These tools are used to preprocess the descriptors, to assign the labels, to visualize the GTM landscapes, etc. They are written in Python3 and Java languages and available by a request to the Laboratory of Chemoinformatics.

8.1   GTM Preprocessing

8.1.1 Descriptor Standardization

As it was described in chapter 4.1, GTM is sensitive to preprocessing. Therefore, the standardization scheme was implemented using Java programming language (*standardizeDescriptors.jar*). The incremental algorithm to compute the mean values and variances is used in the program:

$$\bar{x}_i = \bar{x}_{i-1} + \frac{x_i - \bar{x}_{i-1}}{i} \tag{8.1},$$

$$var_i = var_{i-1} + i * (i-1) * (\bar{x}_i - \bar{x}_{i-1})^2 \tag{8.2},$$

where $\bar{x}_i$ and var_i are the mean value and the variance of a descriptor after passing the ith molecule, respectively. Next, the standard deviation is computed as a square root out of the variance, and the settings file containing the number of descriptors, mean values, variances and standard deviations is created. This settings file can be used later to transform other data sets which should be projected to the map.

8.1.2 Descriptors Filtering

Dimensionality reduction is a hot topic since large chemical data sets are complicated objects, and the molecules in these data sets cannot be well described only by few descriptors. At the same time, even the most effective techniques such as PCA, SOM or GTM cannot handle millions of descriptors which might happen in the case of Big Data. Therefore, dimensionality reduction should be split into at least two steps: (a) conditional descriptors selection, and (b) exhaustive dimensionality reduction. The last one can be done by PCA, SOM or GTM, whereas the first step should be simple and straightforward. As one of the possible solutions, descriptor filtering accordingly to its standard deviation was proposed.

First, the settings file containing mean values and standard deviations for the given data set should be generated by *standardizeDescriptors.jar* (chapter 8.1.1). Next, the initial SVM file, as well as the header file (in case of ISIDA fragment descriptors generated by ISIDAFragmentor2017 tool [95]) are filtered accordingly to the threshold on standard deviation set by the user. This threshold is a percentage out of the maximal standard deviation detected across the file (2% by default). So, if a descriptor possesses the deviation which is less than the threshold, such descriptor will be removed from the SVM file.

Since the standardization process of a large number of descriptors (>100K) is a computationally heavy task, it is recommended first to generate the settings file using *standardizeDescriptors.jar*, then to filter the descriptors using *filterISIDAdescriptors.jar*, and after to standardize the filtered SVM file using the filtered settings file.

8.2   Likelihood-Based GTM Applicability Domain Implementation

The likelihood-based GTM Applicability Domain (AD) is already described in chapter 4.2 and its basic idea is to discard the items which log-likelihood (LLh) is lower than a certain threshold. As was mentioned, in this project we propose to generate the

threshold fitting a Gaussian minimizing a root mean square error (RMSE). The workflow consists of four steps:

1) Sorting and clustering the data set accordingly to its LLh with step=1;

2) Initialize the parameters of the Gaussian function (the width ω^{init}, the amplitude A^{init}, and the peak position μ^{init});

3) Fit the Gaussian minimizing the RMSE;

4) Compute the LLh threshold.

The Gaussian function is determined as:

$$D_i' = A * \exp\left(-\frac{LLh_i - \mu}{2\omega^2}\right) \tag{8.3},$$

where D_i' is the predicted number of items at the LLh_i. Here, A is initialized as the largest number of items possessing the same LLh, and μ is initialized as:

$$\mu^{init} = \frac{\sum_{i=1}^{n} LLh_i * N_i}{n} \tag{8.4},$$

where n is the number of items in the data set, and N_i is the number of items corresponding to the LLh_i. Thus, ω is initialized as:

$$\omega^{init} = \frac{stdv}{2} \tag{8.5},$$

$$stdv = \sqrt{\frac{\sum_{i=1}^{n}(LLh_i - \mu^{init})^2}{n - 1}} \tag{8.6}.$$

To optimize the Gaussian parameters, brute force is used. For each combination μ-A-ω rmse is computed using the equation (8.3), where $\mu \subset [\mu^{init}; LLh(A^{init})*0.95]$, $A \subset [A^{init} * 0.9; A^{init} * 1.1]$, and $\omega \subset [\omega^{init}; \omega^{init} * 3]$. In order to boost the calculations, the algorithm checks the ω values until $RMSE_{new} - RMSE_{old} \leq 0.001$. For A and μ, all values are checked.

Once the grid search is finished, the attempt with the minimal RMSE is selected, and the LLh threshold ($LLh_{threshold}$) is computed as:

$$LLh_{threshold} = \mu - 3 * \omega \qquad (8.7).$$

The described algorithm is implemented in Python3 and can be easily used as a Python library. As input, it needs only the file with the responsibilities generated by the GTMapTool.

8.3   GTM Landscape Building and Visualization

The concept of GTM landscapes is already discussed in chapter 3.1.3. Here, we describe the tool which is used to build and to visualize the landscape, to make the QSAR/QSPR predictions, and to validate the model. The tool named *GTM2018.py* is written in Python3 and it has two dependencies: Plotly [113] and SciKit-Learn [118].

The tool is mainly used to build classification, regression and density landscapes. The output landscape is saved as an XML file which can be used later to make the predictions for the new compounds. The landscape can be also visualized in an interactive way. For this purpose, an HTML page is generated which can be customized by the user adding smooth and transparency which, in turn, corresponds to density, changing the map size (width and height), setting the minimal and maximal property values used to compute the color scale, etc. Note that the tool uses dynamic transparency thresholds to display density using the *minimal Density* threshold.

In addition, the tool is able to compute basic statistics used in QSAR studies, namely determination coefficient (R^2), Balanced Accuracy (BA) and Area Under the Receiver Operating Characteristics Curve (ROC AUC). For this purpose, a test file with its responsibilities and known labels/property values are specified.

8.4   AutoZoom

To analyze and to compare large chemical collections, the *AutoZoom* tool was developed. This tool takes the manifold and GTM class landscape (chapter 3.1.3) built for

the libraries as input matter. Also, it requires the responsibilities, the list of smiles and the descriptors for each library separately.

The algorithm implemented in the AutoZoom tool first scans the landscape in order to find the zones which are needed to be zoomed (chapter 4.3). If such are found, it runs recursive (multilevel) zooming until the density in the cluster satisfies the required threshold. Next, the algorithm runs Maximum Common Substructure (MCS) search described in chapter 4.4. The discovered MCSs are then collected and stored as a pickle archive (Python package to work with binary files). Besides that, the tool collects the information on the *parent* nodes (the full path to the node where the MCSs were extracted from) and smiles returned these MCSs.

The program has several dependencies, such as NumPy, Plotly, GTMapTool, and ChemAxon's JChem cartridge.

8.5   GTM Constrained Screening

The tool developed for Constrained Screening (CS) is web-based. The backend part is written in Python where the *GTM2018.py* tool is used as a library (see chapter 8.3). The server is run by Django software [119]. The frontend part is done in JavaScript, HTML5, and JQuery. The new page is shown in Figure 39.

Figure 39. The client side of the Constrained Screening web tool.

To use the tool, the manifold file, and the classes/properties landscapes must be specified. To add more landscapes, the user should use the "+" button. To remove a landscape, the user should use the "−" button. Once the files are given, the X range (the desirable range for the given activity/property) for each landscape is specified. The query landscape can be built by pressing the "Build" button (Figure 40). The user can then continue the analysis of the query landscape in the Plotly's cloud or he/she can download it using the "Download" button. The numbers on the right side of the color bar represent the satisfaction score. This score means how much the nodes match the given query and it ranges from 0 to the number of conditions in the user's query. Thus, the score equal to 2 means that only 2 conditions are satisfied.
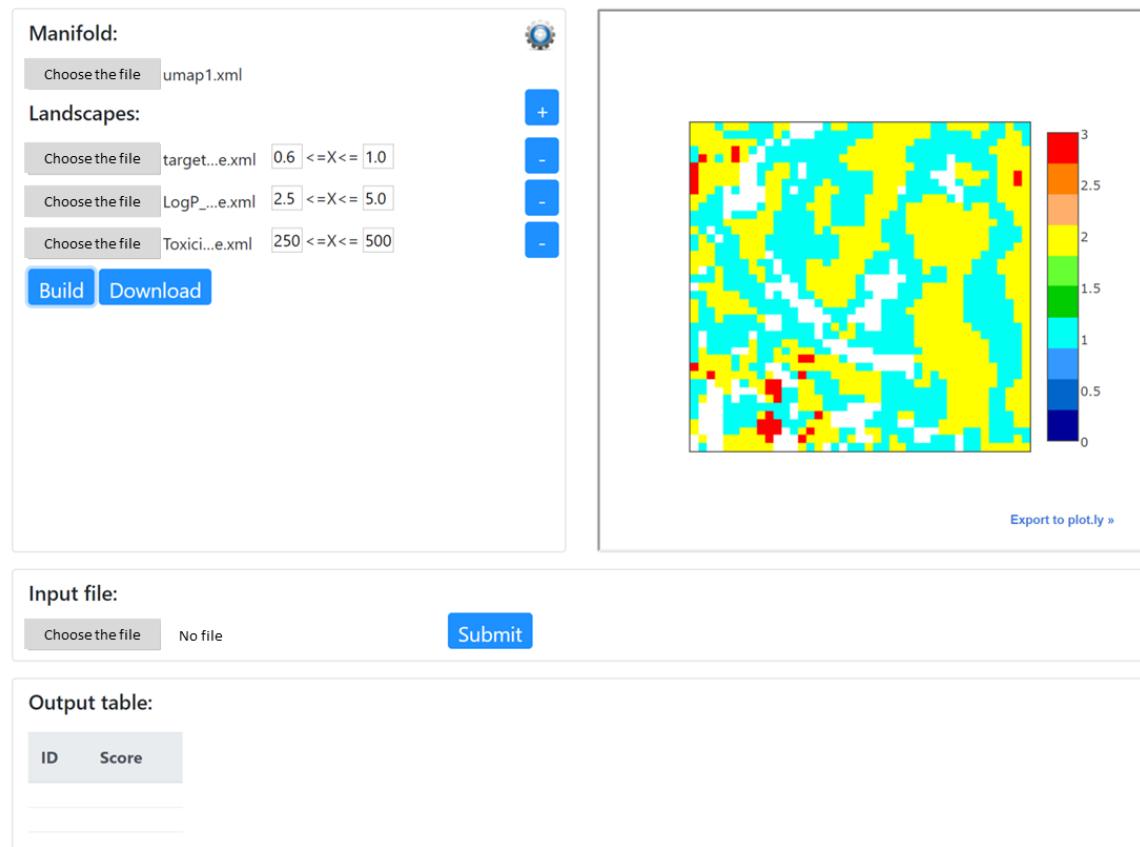
Figure 40. Training of the query landscape.

In case if the user wants to predict new compounds, he/she chooses the SVM file with the corresponding descriptors in the "Input file" field and pushes the "Submit" button. The tool will show the top-10 compounds with their order number and satisfaction score (Figure 41). The rest can be downloaded by the user using the "Download" button in the "Output table" section.
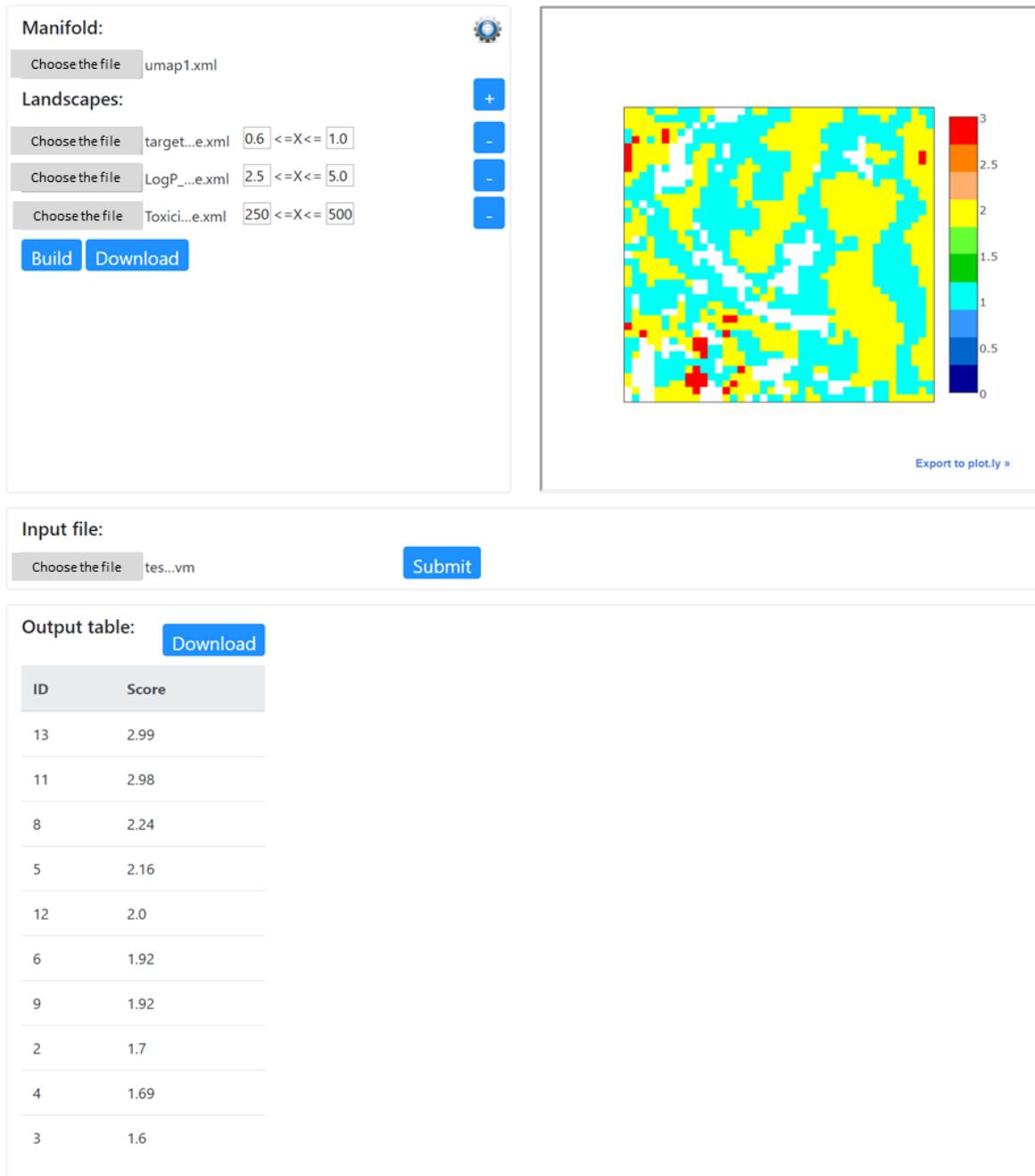
Figure 41. Predicting new compounds using the query landscape.

9 Conclusion and Perspectives

In this work, we dealt with: (*i*) methodological developments, (*ii*) design of algorithms for automatized maps analysis, (*iii*) GTM application to different chemoinformatics tasks (libraries comparison, library enrichment, and virtual screening) and, (*iv*) software development.

Methodological developments. Treatment of Big Data in chemistry is a challenge for any machine learning method, in particular, for GTM, which may need to use large frame sets (FS) in combination with large dimensionality of the initial data space. Since the capacity of earlier reported algorithms for manifold construction (classical and incremental) was limited, we designed the "Parallel GTM" algorithm based on simultaneous training of several manifolds on different FSs followed by their merging into one sole manifold. The developed algorithm allowed us to build a GTM for the ChEMBL-23 database (1.7 M compounds) using the entire database as a FS. Benchmarking of predictive performance of classification models, which were built on the manifolds obtained with different algorithms and FS sizes varying from 1K to 1.7M molecules, demonstrated that (*i*) the parallel algorithm performs similarly to classical and incremental ones, and (*ii*) a small frame set of 5000 molecules (*i.e.*, 0.003% of ChEMBL) is sufficient for obtaining well-performing manifold.

The log-likelihood (LLh) threshold is often used to delimit an applicability domain of GTM-based classification and regression models. In order to calculate the "optimal" the

LLh threshold, we proposed to use the width of the Gaussian function which fits the LLh distribution.

Using the existing pairwise Maximum Common Substructure (MCS) algorithm, we suggested a new protocol of MCS extraction from the ensemble of structures. Its efficiency was tested on different sets up to 1000 molecules.

Automatized maps analysis. Two new algorithms performing automatized maps analysis were developed: (*i*) selection of zones of interest [5] and, (*ii*) hierarchical GTM zooming. The zones of interest on GTM represent selected areas populated by molecules possessing a given activity (property) profile. They result from the superposition of a certain number of class and/or activity (property) landscapes. The developed algorithm automatically selects the zones, which entirely or partially correspond to the desired profile. Notice that the ensemble of these zones over the map form *Query Landscape*, which can be used in virtual screening by selecting hits dropping in the zones of interest.

The hierarchical GTM zooming approach proposed by Nabney et al. [11] in view of improving map's resolution, becomes desirable, in some cases strictly required for GTMs accommodating large volumes of data. The developed algorithm first screens the map in order to select rectangular zones susceptible to zooming procedure according to the data density threshold. Two scenarios were considered: overlapping and non-overlapping zones. The former allows increasing the overall size of zoomed areas because of the possibility to overcome the density threshold.

Applications. Developed tools were used in three projects: (*i*) application of GTM to virtual screening (VS), (*ii*) comparison of large databases, and (*iii*) enrichment of proprietary library.

In the VS project, two types of GTMs for the ChEMBL23 database were used: "universal" and "local". The formers were trained in a multitask manner to obtain simultaneously classification models for 236 activities, whereas the latter were trained

individually for each activity. The developed maps and class landscapes were benchmarked with several machine-learning techniques (similarity search with data fusion, neural network, and random forest) in virtual screening of the DUD database. It has been demonstrated that local GTMs perform similarly or even better than popular machine-learning approaches. In terms of predictive performance, "universal" GTMs were less efficient, but still acceptable. On the other hand, the models derived from the "universal" map have a larger applicability domain.

In another project, GTM was challenged to analyze large chemical data set of more than 21M compounds resulted from merging of 3 databases: ChEMBL-17 (100K compounds), PubChem-17 (11M compounds) and FDB-17 (10M compounds). Two former databases contained only existing molecules, whereas the latter contained virtual structures containing no more than 17 heavy atoms. The databases were compared using (*i*) Bhattacharyya, Soergel and Euclidean distances, (*ii*) GTM class and (*iii*) GTM property landscapes. The data analysis with the help of GTM allowed us to identify structural motifs exclusively present only in one of the considered databases.

In the 3rd project, the proprietary collection of Boehringer Ingelheim (1.7 M molecules) was superposed on GTM with commercial Aldrich-Market Select database (8.2 M). Analysis of non-overlapping zones revealed 1.2K commercial structures containing fully new cores, passed drug-like filters and predicted as active against at least one ChEMBL target. The corresponding molecules were recommended to BI to be synthesized or purchased.

Software development. New methodology and algorithms developed in this work were implemented as a command line and web-based software tools. Thus, the hierarchical GTM zooming technique was coupled with the MCS extraction protocol and presented as the "AutoZoom" tool written in Python3 language. The algorithm helping to delineate zones of interest was implemented as a web-based tool within the Django framework. The tools for the construction of GTM-based classification and regression models were prepared

using FreePascal and Python3 programming languages. These tools are accessible from the Laboratory of Chemoinformatics by a request.

Perspectives. Some projects initiated in this work have not been completed. Still, the *Query Landscapes* technique needs to be validated in virtual screening experiments. Another project may concern an application of the hierarchical GTM zooming to GTM-based classification and regression tasks.

In its current state, the MCS extraction protocol operates only with connected graphs. However, common structural motifs may range from specifically substituted scaffolds to fuzzier 'pharmacophore-like' patterns [65]. Therefore, the extension of our algorithm on disconnected MCS could improve the structural data analysis.

The manifold "fusion" protocol in Parallel GTM needs to be optimized. Thus, in the current version of the program, the manifold merging strategy simply computes the average positions of the RBFs. Weighted by likelihood merging could, in principle, be used as an alternative.

Studied in this work datasets of some 20 M molecules represent a small portion of all existing molecules (some 200 M). An application of GTM to larger datasets is an obvious extension of this work.

10 References

1. Tetko I V., Engkvist O, Koch U, Reymond JL, and Chen H (2016) BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. Mol Inform 35:615–621. https://doi.org/10.1002/minf.201600073

2. Ruddigkeit L, Van Deursen R, Blum LC, and Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 52:2864–2875. https://doi.org/10.1021/ci300415d

3. Polishchuk PG, Madzhidov TI, and Varnek A (2013) Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des 27:675–679. https://doi.org/10.1007/s10822-013-9672-4

4. Bishop CM, Svensén M, and Williams CKI (1998) GTM: The Generative Topographic Mapping. Neural Comput 10:215–234. https://doi.org/10.1162/089976698300017953

5. Gaspar HA, Baskin II, Marcou G, Horvath D, and Varnek A (2015) Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge. J Chem Inf Model 55:84–94. https://doi.org/10.1021/ci500575y

6. Gaspar HA, Baskin II, Marcou G, Horvath D, and Varnek A (2015) GTM-Based QSAR Models and Their Applicability Domains. Mol Inform 34:348–356. https://doi.org/10.1002/minf.201400153

7. Visini R, Awale M, and Reymond JL (2017) Fragment Database FDB-17. J Chem Inf Model 57:700–709. https://doi.org/10.1021/acs.jcim.7b00020

8. Huang N, Shoichet BK, and Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801. https://doi.org/10.1021/jm0608356

9. Sidorov P, Gaspar H, Marcou G, Varnek A, and Horvath D (2015) Mappability of drug-like space: Towards a polypharmacologically competent map of drug-relevant compounds. J Comput Aided Mol Des 29:1087–1108. https://doi.org/10.1007/s10822-015-9882-z

10. Lin A, Horvath D, Marcou G, Beck B, and Varnek A (2019) Multi-task generative topographic mapping in virtual screening. J Comput Aided Mol Des 33:331–343. https://doi.org/10.1007/s10822-019-00188-x

11. Tino P, and Nabney I (2002) Hierarchical GTM: Constructing localized nonlinear

projection manifolds in a principled way. IEEE Trans Pattern Anal Mach Intell 24:639–656. https://doi.org/10.1109/34.1000238

12. van der Pijl R, Bender A, de Vries H, van den Hoven OO, van Westen GJP, Mulder-Krieger T, van Vlijmen HWT, Wegner JK, and IJzerman AP (2012) Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. J Med Chem 55:7010–7020. https://doi.org/10.1021/jm3003069

13. Lipinski CA, Lombardo F, Dominy BW, and Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 64:4–17. https://doi.org/10.1016/j.addr.2012.09.019

14. (2019) Chemical Abstract Service. https://www.cas.org/about/cas-content. Accessed 8 Jul 2019

15. De Mauro A, Greco M, and Grimaldi M (2016) A formal definition of Big Data based on its essential features. Libr Rev 65:122–135

16. Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, and de Vlieg J (2014) Data-driven medicinal chemistry in the era of big data. Drug Discov Today 19:859–868. https://doi.org/https://doi.org/10.1016/j.drudis.2013.12.004

17. Hu Y, and Bajorath J (2017) Entering the 'big data'era in medicinal chemistry: molecular promiscuity analysis revisited. Futur Sci OA 3:FSO179

18. Bemis GW, and Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893

19. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, and Waldmann H (2007) The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J Chem Inf Model 47:47–58. https://doi.org/10.1021/ci600338x

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

21. Kenny PW, and Sadowski J (2005) Structure modification in chemical databases. Chemoinformatics drug Discov 23:271–285

22. Sorzano COS, Vargas J, and Montano AP (2014) A survey of dimensionality reduction techniques. arXiv Prepr arXiv14032877

23. Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, and Chen L (2008) Data visualization with multidimensional scaling. J Comput Graph Stat 17:444–472. https://doi.org/10.1198/106186008X318440

24. Sammon, J. W. J (1969) (Sammon Mapping)A Nonlinear Mapping for Data Structure Analysis. IEEE Trans Comput 18:401–409. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/T-C.1969.222678

25. Pearson K (1901) On lines and planes of closest fit to systems of points. Philos Mag 2:559–572

26. Hotelling H (1933) Analysis of a complex of statistical variables into Principal Components. Jour. Educ. Psych., 24, 417-441, 498-520. J Educ Psychol 24:417

27. Akella LB, and DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330

28. Kohonen T (1990) The self-organizing map. Proc IEEE 78:1464–1480

29. Belkin M, and Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373–1396

30. Foster DP, Kakade SM, and Zhang T (2008) Multi-view dimensionality reduction via canonical correlation analysis

31. Hyvärinen A, and Oja E (2000) Independent component analysis: algorithms and applications. Neural networks 13:411–430

32. Osborne JW, Costello AB, and Kellow JT (2008) Best practices in exploratory factor analysis. Best Pract Quant methods 86–99

33. Balasubramanian M, and Schwartz EL (2002) The isomap algorithm and topological stability. Science (80-) 295:7

34. Roweis ST, and Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science (80-) 290:2323–2326

35. Wang Y, Yao H, and Zhao S (2016) Auto-encoder based dimensionality reduction. Neurocomputing 184:232–242

36. Balakin K V. (2009) Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery. John Wiley & Sons

37. Maniyar DM, Nabney IT, Williams BS, and and Andreas Sewing (2006) Data Visualization during the Early Stages of Drug Discovery. J Chem Inf Model 46:1806–1818. https://doi.org/10.1021/ci050471a

38. Neal RM (2007) Pattern Recognition and Machine Learning. springer

39. Hutchison D, and Mitchell JC (1973) Intelligent Data Engineering and Automated Learning. Springer

40. Bishop CM, Svensén M, and Williams CKI (1998) Developments of the generative topographic mapping. Neurocomputing 21:203–224. https://doi.org/10.1016/S0925-2312(98)00043-5

41. Erwin E, Obermayer K, and Schulten K (1992) Self-organizing maps: Ordering, convergence properties and energy functions. Biol Cybern 67:47–55

42. Maggiora GM, and Bajorath J (2014) Chemical space networks: a powerful new paradigm for the description of chemical space. J Comput Aided Mol Des 28:795–802

43. Schneider G, Neidhart W, Giller T, and Schmid G (1999) "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. Angew Chemie Int Ed 38:2894–2896

44. Iyer P, Stumpfe D, Vogt M, Bajorath J, and Maggiora GM (2013) Activity landscapes, information theory, and structure–activity relationships. Mol Inform 32:421–430

45. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, and Varnek A (2012)

Generative Topographic Mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison. Mol Inform 31:301–312. https://doi.org/10.1002/minf.201100163

46. Gaspar HA, Marcou G, Horvath D, Arault A, Lozano S, Vayer P, and Varnek A (2013) Generative topographic mapping-based classification models and their applicability domain: Application to the biopharmaceutics drug disposition classification system (BDDCS). J Chem Inf Model 53:3318–3325. https://doi.org/10.1021/ci400423c

47. Davis LD, and Mitchell M (1991) Handbook of Genetic Algorithms. Computer (Long Beach Calif) 1–6

48. Horvath D, Brown J, Marcou G, and Varnek A (2014) An Evolutionary Optimizer of libsvm Models. Challenges 5:450–472

49. Sidorov P, Viira B, Davioud-Charvet E, Maran U, Marcou G, Horvath D, and Varnek A (2017) QSAR modeling and chemical space analysis of antimalarial compounds. J Comput Aided Mol Des 31:441–451. https://doi.org/10.1007/s10822-017-0019-4

50. Lin A, Horvath D, Afonina V, Marcou G, Reymond JL, and Varnek A (2018) Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. ChemMedChem 13:540–554. https://doi.org/10.1002/cmdc.201700561

51. Gaspar HA, Sidorov P, Horvath D, Marcou G, Baskin II, and Varnek A (2016) Generative topographic mapping approach to chemical space analysis. In: ACS Symposium Series. pp 211–241

52. Kireeva N, Kuznetsov SL, and Tsivadze AY (2012) Toward navigating chemical space of ionic liquids: Prediction of melting points using generative topographic maps. Ind Eng Chem Res 51:14337–14343. https://doi.org/10.1021/ie3021895

53. Liaw A, and Wiener M (2002) Classification and regression by randomForest. R news 2:18–22

54. Nadaraya EA (2005) On Estimating Regression. Theory Probab Its Appl 9:141–142. https://doi.org/10.1137/1109020

55. Quinlan JR (1992) Learning With Continuous Classes. In: Proceedings AI'92, 5th Australian Conference on Artificial Intelligence.World Scientific. World Scientific, pp 343–348

56. Sjöström M, Eriksson L, and Wold S (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

57. Gimadiev TR, Madzhidov TI, Marcou G, and Varnek A (2016) Generative Topographic Mapping Approach to Modeling and Chemical Space Visualization of Human Intestinal Transporters. Bionanoscience 6:464–472. https://doi.org/10.1007/s12668-016-0246-5

58. Casciuc I, Zabolotna Y, Horvath D, Marcou G, Bajorath J, and Varnek A (2019) Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? J Chem Inf Model 59:564–572. https://doi.org/10.1021/acs.jcim.8b00650

59.	Andrade AO, Nasuto S, Kyberd P, and Sweeney-Reed CM (2005) Generative topographic mapping applied to clustering and visualization of motor unit action potentials. BioSystems 82:273–284. https://doi.org/10.1016/j.biosystems.2005.09.004

60.	Escobar MS, Kaneko H, and Funatsu K (2015) Combined generative topographic mapping and graph theory unsupervised approach for nonlinear fault identification. AIChE J 61:1559–1571. https://doi.org/10.1002/aic.14748

61.	Kayastha S, Kunimoto R, Horvath D, Varnek A, and Bajorath J (2017) 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 31:961–977

62.	Liu T, Lin Y, Wen X, Jorissen RN, and Gilson MK (2006) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res 35:D198–D201

63.	Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, and Chong J (2015) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053

64.	Gamo F-J, Sanz LM, Vidal J, De Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, and Hasan S (2010) Thousands of chemical starting points for antimalarial lead identification. Nature 465:305

65.	Klimenko K, Marcou G, Horvath D, and Varnek A (2016) Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. J Chem Inf Model 56:1438–1454. https://doi.org/10.1021/acs.jcim.6b00192

66.	Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS, and Chang RSL (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 31:2235–2246

67.	Kayastha S, Horvath D, Gilberg E, Gütschow M, Bajorath J, and Varnek A (2017) Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. J Chem Inf Model 57:1218–1232. https://doi.org/10.1021/acs.jcim.7b00128

68.	Liwo A, Czaplewski C, Ołdziej S, and Scheraga HA (2008) Computational techniques for efficient conformational sampling of proteins. Curr Opin Struct Biol 18:134–139

69.	Good AC, and Cheney DL (2003) Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. J Mol Graph Model 22:23–30

70.	Agrafiotis DK, Gibbs AC, Zhu F, Izrailev S, and Martin E (2007) Conformational sampling of bioactive molecules: a comparative study. J Chem Inf Model 47:1067–1086

71.	Horvath D, Baskin I, Marcou G, and Varnek A (2017) Generative topographic mapping of conformational space. Mol Inform 36:1700036

72.	Wang J, Wolf RM, Caldwell JW, Kollman PA, and Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174

73. Horvath D, Marcou G, and Varnek A (2018) Monitoring of the Conformational Space of Dipeptides by Generative Topographic Mapping. Mol Inform 37:1700115

74. Horvath D, Marcou G, and Varnek A (2019) Generative Topographic Mapping of the Docking Conformational Space. Molecules 24:2269

75. Mishima K, Kaneko H, and Funatsu K (2014) Development of a New De Novo Design Algorithm for Exploring Chemical Space. Mol Inform 33:779–789. https://doi.org/10.1002/minf.201400056

76. Takeda S, Kaneko H, and Funatsu K (2016) Chemical-Space-Based de Novo Design Method To Generate Drug-Like Molecules. J Chem Inf Model 56:1885–1893. https://doi.org/10.1021/acs.jcim.6b00038

77. Gaspar HA, Baskin II, Marcou G, Horvath D, and Varnek A (2015) Stargate GTM: bridging descriptor and activity spaces. J Chem Inf Model 55:2403–2410

78. Sattarov B, Baskin II, Horvath D, Marcou G, Bjerrum EJ, and Varnek A (2019) De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. J Chem Inf Model 59:1182–1196. https://doi.org/10.1021/acs.jcim.8b00751

79. Hoffer L, and Horvath D (2012) S4MPLE–Sampler For Multiple Protein–Ligand Entities: simultaneous docking of several entities. J Chem Inf Model 53:88–102

80. Ruggiu F, Marcou G, Varnek A, and Horvath D (2010) ISIDA Property-labelled fragment descriptors. Mol Inform 29:855–868. https://doi.org/10.1002/minf.201000099

81. ChemAxon JChem. https://chemaxon.com/products/jchem-engines

82. Muegge I, and Oloff S (2006) Advances in virtual screening. Drug Discov today Technol 3:405–411

83. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20:318–331

84. Hristozov D, Oprea TI, and Gasteiger J (2007) Ligand-based virtual screening by novelty detection with self-organizing maps. J Chem Inf Model 47:2044–2062

85. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, and Gasteiger J (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. J Med Chem 50:1698–1702

86. Schneider G, and Nettekoven M (2003) Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. J Comb Chem 5:233–237

87. Gabriel E, Fagg GE, Bosilca G, Angskun T, Dongarra JJ, Squyres JM, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain RH, Daniel DJ, Graham RL, and Woodall TS (2010) Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. In: European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting. Springer, pp 97–104

88. Qiu X, Fox GC, Yuan H, and Bae S (2008) Parallel Data Mining on Multicore Systems. In: 2008 Seventh International Conference on Grid and Cooperative Computing. IEEE, pp 4–11

89. Choi JY, Bae SH, Qiu X, and Fox G (2010) High performance dimension reduction and visualization for large high-dimensional data analysis. In: CCGrid 2010 - 10th IEEE/ACM International Conference on Cluster, Cloud, and Grid Computing. IEEE Computer Society, pp 331–340

90. Ormoneit D, and Tresp V (1998) Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. IEEE Trans Neural Networks 9:639–650. https://doi.org/10.1109/72.701177

91. Franklin JN (2012) Matrix theory. Courier Corporation

92. Givon LE, Unterthiner T, Erichson NB, Chiang DW, Larson E, Pfister L, Dieleman S, Lee GR, van der Walt S, and Moldovan TM (2018) scikit-cuda 0.5.2: a Python interface to GPU-powered libraries

93. Gaulton A, Hersey A, Nowotka ML, Patricia Bento A, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, and Leach AR (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954. https://doi.org/10.1093/nar/gkw1074

94. Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M, and Hagler AT (1988) Structure and energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. Proteins Struct Funct Bioinforma 4:31–47. https://doi.org/10.1002/prot.340040106

95. Marcou G, Solov'ev VP, Horvath D, and Varnek A (2017) ISIDA Fragmentor - User Manual

96. Miteva MA, Violas S, Montes M, Gomez D, Tuffery P, and Villoutreix BO (2006) FAF-Drugs: free ADME/tox filtering of compound collections. Nucleic Acids Res 34:W738–W744. https://doi.org/10.1093/nar/gkl065

97. Giganti D, Guillemain H, Spadoni J-L, Nilges M, Zagury J-F, and Montes M (2010) Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment. J Chem Inf Model 50:992–1004. https://doi.org/10.1021/ci900507g

98. Basse N, Montes M, Maréchal X, Qin L, Bouvier-Durand M, Genin E, Vidal J, Villoutreix BO, and Reboud-Ravaux M (2010) Novel Organic Proteasome Inhibitors Identified by Virtual and in Vitro Screening. J Med Chem 53:509–513. https://doi.org/10.1021/jm9011092

99. Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Ivanenkov Y, Putin E, Zhavoronkov A, and Asadulaev A (2018) Reinforced Adversarial Neural Computer for de Novo Molecular Design . J Chem Inf Model 58:1194–1204. https://doi.org/10.1021/acs.jcim.7b00690

100. Kang S, and Cho K (2019) Conditional Molecular Design with Deep Generative Models. J Chem Inf Model 59:43–52. https://doi.org/10.1021/acs.jcim.8b00263

101. Schneider P, and Schneider G (2016) De novo design at the edge of chaos: Miniperspective. J Med Chem 59:4077–4086

102. Chang J-W, and Jin D-S (2003) A new cell-based clustering method for large, high-dimensional data in data mining applications. In: Proceedings of the 2002 ACM

symposium on Applied computing. ACM, p 503

103. Medina-Franco JL, Maggiora GM, Giulianotti MA, Pinilla C, and Houghten RA (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Des 70:393–412. https://doi.org/10.1111/j.1747-0285.2007.00579.x

104. Bernard P, Golbraikh A, Kireev D, Chrétien JR, and Rozhkova N (1998) Comparison of chemical databases: Analysis of molecular diversity with Self Organising Maps (SOM). Analusis 26:333–341. https://doi.org/10.1051/analusis:1998182

105. ChemAxon Standardizer. https://docs.chemaxon.com/display/docs/Standardizer

106. Monev V (2004) Introduction to Similarity Searching in Chemistry. Match-Communications Math Comput Chem 51:7–38

107. (2019) RDKit: Open-source cheminformatics. http://www.rdkit.org

108. Rogers D, and Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t

109. Volochnyuk DM, Ryabukhin S V., Moroz YS, Savych O, Chuprina A, Horvath D, Zabolotna Y, Varnek A, and Judd DB (2019) Evolution of commercially available compounds for HTS. Drug Discov Today 24:390–402. https://doi.org/10.1016/j.drudis.2018.10.016

110. Hariharan R, Janakiraman A, Nilakantan R, Singh B, Varghese S, Landrum G, and Schuffenhauer A (2011) MultiMCS: A fast algorithm for the maximum common substructure problem on multiple molecules. J Chem Inf Model 51:788–806. https://doi.org/10.1021/ci100297y

111. Oliphant TE, and January UT (2010) Guide to NumPy. Tregol Publishing

112. Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9:10–20. https://doi.org/10.1109/MCSE.2007.58

113. Inc. PT (2015) Collaborative data science. In: Plotly Technol. Inc. https://plot.ly

114. Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, and Wyatt PG (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem Chem Enabling Drug Discov 3:435–444

115. Baell JB, and Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740

116. Doveston RG, Tosatti P, Dow M, Foley DJ, Li HY, Campbell AJ, House D, Churcher I, Marsden SP, and Nelson A (2015) A unified lead-oriented synthesis of over fifty molecular scaffolds. Org Biomol Chem 13:859–865

117. Jadhav A, Ferreira RS, Klumpp C, Mott BT, Austin CP, Inglese J, Thomas CJ, Maloney DJ, Shoichet BK, and Simeonov A (2009) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. J Med Chem 53:37–51

118. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,

Prettenhofer P, Weiss R, and Dubourg V (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

119. Holovaty A, and Kaplan-Moss J (2009) The definitive guide to Django: Web development done right. Apress

11 List of Abbreviations

AD	Applicability Domain
ADME	An abbreviation in pharmacokinetics and pharmacology for "Absorption, Distribution, Metabolism, and Excretion"
AMS	Aldrich-Market Select
ANN	Artificial Neural Network
AUC (ROC AUC)	Area Under the Receiver Operating Characteristics Curve
BA	Balanced Accuracy
BI	Boehringer Ingelheim
CLF	Class-Likelihood Factor
CPF	Class Prevalence Factor
CPU	Central Processing Unit
CS	Constrained Screening
CVFF	Consistent Valence Force Field
DUD	Directory of Useful Decoys

EC50	Half maximal Effective Concentration
EM	Expectation-Maximization algorithm
FS	Frame Set
GA	Genetic Algorithm
GTM	Generative Topographic Mapping
IC50	Half maximal Inhibitory Concentration
kNN	k-Nearest Neighbors
LLh	Logarithm of Likelihood
MCS	Maximum Common Substructure
MDS	Multi-Dimensional Scaling
MPI	Message Passing Interface technique
PCA	Principal Component Analysis
PGTM	Parallel Generative Topographic Mapping
PSM	Privileged Structural Motif
Q^2	Determination coefficient in cross-validation
QSAR	Quantitative Structure-Activity Relation
QSPR	Quantitative Structure-Property Relation
R^2	Determination coefficient
RAM	Random Access Memory

RBF   Radial Basis Function

RF   Random Forest

RMSE   Root Mean Square Error

RP   Responsibility Pattern

SAR   Structure-Activity Relationship

SOM   Self-Organizing Map

SRC   Tyrosine kinase inhibitors

SVM (format)   Support-Vector Machine

SVM (method)   Sparse Vector Matrix

TPSA   Topological Polar surface Area

uGTM   Universal Generative Topographic Map

VS   Virtual Screening

Appendix 1

Supplementary Material for section 5.1

Table S1. List of 618 ChEMBL (v. 23) targets used for unievrsal GTM 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		

Appendix 2

Supplementary Material for section 6

Table S2. PubChem biological targets used for GTM map selection.

PubChem ID	PubChem BioAssay name *
1012	Tissue non-specific alkaline phosphatase precursor [Homo sapiens]
1159524	HTS for Foot and Mouth Disease Virus Antivirals
1490	QHTS Assay For Inhibitors Of Bacillus Subtilis Sfp Phosphopantetheinyl Transferase (PPTase)
1721	QHTS Assay For Inhibitors Of Leishmania Mexicana Pyruvate Kinase (LmPK)
1981	A Screen For Inhibitors Of The PhoP Regulon In Salmonella Typhimurium Using A Modified Counterscreen
2100	qHTS Assay for Inhibitors and Activators of Human alpha-Glucosidase Cleavage of Glycogen
2289	qHTS Assay for Modulators of miRNAs and/or Inhibitors of miR-21
2314	Cycloheximide Counterscreen For Small Molecule Inhibitors Of Shiga Toxin
2315	A QHTS For Small Molecule Inhibitors Of Shiga Toxin
2451	qHTS Assay for Inhibitors of Fructose-1,6-bisphosphate Aldolase from Giardia Lamblia
2546	VP16 Counterscreen QHTS For Inhibitors Of ROR Gamma Transcriptional Activity
2551	QHTS For Inhibitors Of ROR Gamma Transcriptional Activity
2842	HTS Of A Putative Kinase Compound Library To Identify Inhibitors Of Mycobacterium Tuberculosis H37Rv
410	Cytochrome P450, family 1, subfamily A, polypeptide 2 [Homo sapiens]
485313	Niemann-Pick C1 protein precursor [Homo sapiens]
485364	Thioredoxin glutathione reductase [Schistosoma mansoni]
485367	ATP-dependent phosphofructokinase [Trypanosoma brucei]
504466	ATAD5 protein [Homo sapiens]

PubChem ID	PubChem BioAssay name *
504847	Vitamin D3 receptor isoform VDRA [Homo sapiens]
521	Protein tyrosine phosphatase, non-receptor type 7 isoform 2 [Homo sapiens]
588342	Luciferase [Photinus pyralis]
624173	Hypothetical protein, conserved [Trypanosoma brucei]
624330	Rac GTPase-activating protein 1 [Homo sapiens]
651635	Ataxin-2 [Homo sapiens]
651724	CtBP interacting protein CtIP [Homo sapiens]
652105	qHTS for Inhibitors of phosphatidylinositol 5-phosphate 4-kinase (PI5P4K)
686971	qHTS for induction of synthetic lethality in tumor cells producing 2HG: qHTS for the HT-1080-IDH1KD cell line
686978	TDP1 protein [Homo sapiens]

* PubChem BioAssay target name corresponds to its description or target name on PubChem

Table S3. PubChem biological targets used for GTM map validation.

PubChem ID	PubChem BioAssay name *
686979	qHTS for Inhibitors of human tyrosyl-DNA phosphodiesterase 1 (TDP1): qHTS in cells in presence of CPT
720504	qHTS for Inhibitors of PLK1-PDB (polo-like kinase 1 - polo-box domain): Primary Screen
720580	qHTS for Stage-Specific Inhibitors of Vaccinia Orthopoxvirus: Venus Reporter Primary qHTS
720708	qHTS for Antagonist of cAMP-regulated guanine nucleotide exchange factor 2 (EPAC2): primary screen
743255	Inhibitors Of USP1/UAF1: Primary Screen

PubChem ID	PubChem BioAssay name *
743279	qHTS for Inhibitors of Inflammasome Signaling: IL-1-beta AlphaLISA Primary Screen
778	Cytochrome P450, family 2, subfamily C, polypeptide 19 [Homo sapiens]
1443	uHTS for the identification of compounds that potentiate TRAIL-induced apoptosis of cancer cells
1619	Inhibitors of Plasmodium falciparum M17- Family Leucine Aminopeptidase (M17LAP)
1903	Identification of SV40 T antigen inhibitors: A route to novel anti-viral reagents
2401	A Counter Screen To Identiry Small Molecule Screen For Inhibitors Of The PhoP Regulon In Salmonella Typhimurium
485297	QHTS Assay For Rab9 Promoter Activators
504327	QHTS Assay For Inhibitors Of GCN5L2
504329	Discovery Of Small Molecule Probes For H1N1 Influenza NS1A
504332	QHTS Assay For Inhibitors Of Histone Lysine Methyltransferase G9a
504333	QHTS Assay For Inhibitors of bromodomain adjacent to zinc finger domain 2B [Homo sapiens]
504339	Chain A, Jmjd2a Tandem Tudor Domains In Complex With A Trimethylated Histone H4-K20 Peptide
504832	Primary QHTS For Delayed Death Inhibitors Of The Malarial Parasite Plastid, 48 Hour Incubation
540267	Small Molecules That Selectively Kill Giardia Lamblia: QHTS
588453	QHTS Assay For Inhibitors Of Mammalian Selenoprotein Thioredoxin Reductase 1 (TrxR1): QHTS
588579	QHTS For Inhibitors Of Polymerase Kappa
624171	QHTS Of Nrf2 Activators
624202	QHTS Assay To Identify Small Molecule Activators Of BRCA1

PubChem ID	PubChem BioAssay name *
	Expression
651725	QHTS Assay For Inhibitors Of The Six1/Eya2 Interaction

* PubChem BioAssay target name corresponds to its description or target name on

PubChem
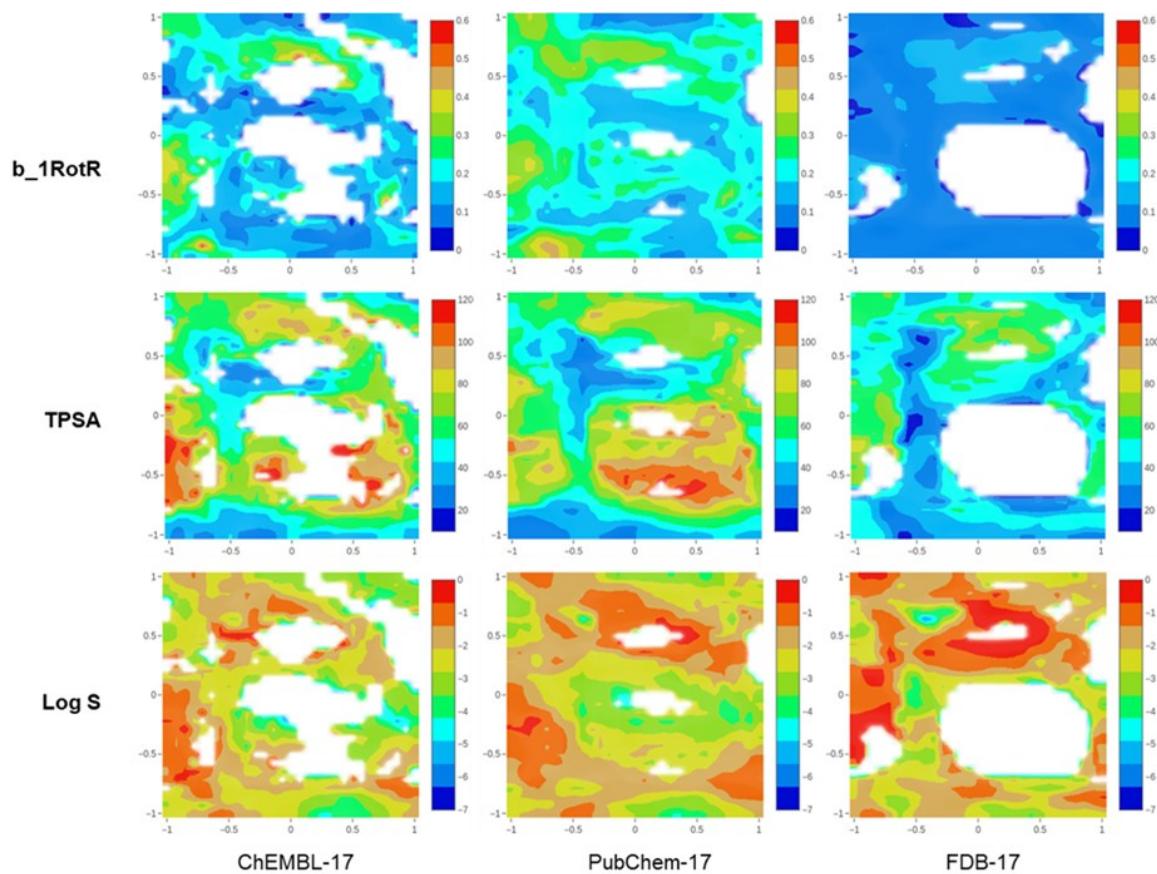
Figure S1. GTM property landscapes for **b_1RotR** (fraction of rotatable single bonds), **TPSA**, and **Log S**.
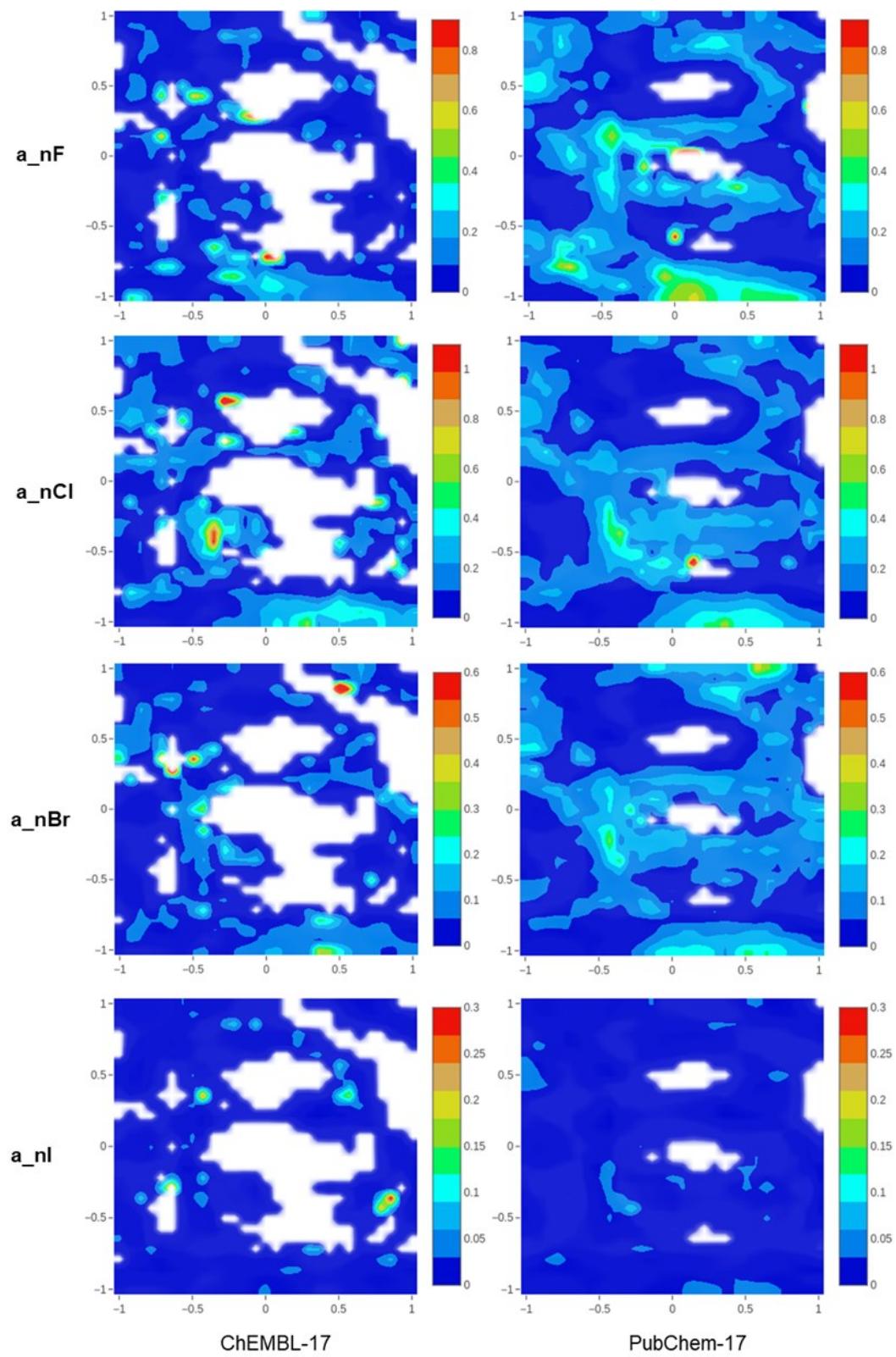
Figure S2. GTM property landscapes for **a_nF** (number of fluorine atoms), **a_nCl** (number of chlorine atoms), **a_nBr** (number of bromine atoms), and **a_nI** (number of iodine atoms).

Arkadii LIN

Cartographie Topographique Générative: un outil puissant pour la visualisation, l'analyse et la modélisation de données chimiques volumineuses

Résumé

Cette thèse concerne l'utilisation de Cartographie Topographique Générative (Generative Topographic Mapping – GTM) pour l'analyse, la visualisation et la modélisation de grands volumes de données chimiques. Les principaux sujets traités dans ces travaux sont le criblage virtuel multi-cibles dans la conception de médicaments et la visualisation, l'analyse et la comparaison de grandes chimiothèques. Plusieurs développements méthodologiques ont été proposés : (*i*) un algorithme de zoom hiérarchique automatisé pour la GTM afin d'aider à résoudre le problème de la résolution des cartes ; (*ii*) un protocole d'extraction automatisé des Sous-structures Maximum Communes (MCS) pour améliorer l'efficacité de l'analyse de données ; (*iii*) un criblage contraint basé sur la GTM permettant de détecter les molécules avec un profil pharmacologique souhaité, et (*iv*) une technique de GTM parallèle, qui réduit significativement le temps nécessaire pour construire une carte. Les méthodologies développées ont été implémentées sous forme de logiciel, utilisé à la fois dans des projets académiques (Université de Strasbourg, France) et industriels (Compagnie Boehringer Ingelheim Pharma, Allemagne).

Mots-clés : GTM, grand volumes de données en chimie, comparaison de grandes chimiothèques, visualisation de données, QSAR, criblage virtuel

Résumé en anglais

This thesis concerns the application of the Generative Topographic Mapping (GTM) approach to the analysis, visualization, and modeling of Big Data in chemistry. The main topics covered in this work are multi-target virtual screening in drug design and large chemical libraries visualization, analysis, and comparison. Several methodological developments were suggested: (i) an automatized hierarchical GTM zooming algorithm helping to resolve the map resolution problem; (ii) an automatized Maximum Common Substructure (MCS) extraction protocol improving efficiency of data analysis; (iii) constrained GTM-based screening allowing to detect molecules with a desired pharmacological profile, and (iv) a parallel GTM technique, which significantly increases the speed of GTM training. Developed methodologies were implemented in a software package used in both academic (University of Strasbourg, France) and industrial (Boehringer Ingelheim Pharma company, Germany) projects.

Key words: GTM, Big Data in chemistry, libraries comparison, data visualization, QSAR, virtual screening