Chemical cartography and complex systems modeling

Iuri Casciuc

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :

Iuri CASCIUC

Soutenue le : **15 septembre 2020**

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

Discipline/ Spécialité : **Chimie / Chémoinformatique**

Cartographie chimique et modélisation de systèmes complexes

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. LEHN Jean-Marie	Professeur, Université de Strasbourg

RAPPORTEURS :

M. MORELLI Xavier	Directeur de recherche, Centre de Recherche en Cancérologie de Marseille (CRCM)
M. VILLOUTREIX Bruno	Directeur de recherche, U1177-Pasteur Lille et UMR 1141 Hopital Universitaire Robert-Debré, Paris

---

AUTRES MEMBRES DU JURY :

Mme. CAMPROUX Anne-Claude	Professeur, Université Paris Diderot

Dedication

In the memory of my father Casciuc Valeriy, who passed away on the 24th of November 2018.

Abstract

This work concerns application of Generative Topographic Mapping method to different tasks including data analysis and visualization, virtual screening and library design. Performance of multi-target GTM-based classification models (uGTM) in virtual screening was investigated and consensus usage of several uGTMs has been suggested. Virtual screening involving a combination of GTM with some other chemoinformatics techniques allowed to discover 29 new BRD4 inhibitors, activities of which were experimentally confirmed. As a library design tool, GTM was compared to the MaxMin method. Although diversity of MaxMin libraries is systematically larger than those obtained with GTM, the latter is much faster and, therefore, can be recommended for large datasets. A modeling workflow for speciation analysis in imine-based Dynamic Combinatorial Libraries in absence and presence of a protein has been suggested. Developed models are publicly available at the site of the Laboratory of Chemoinformatics.

Acknowledgements

I would like to express my sincere gratitude to all my colleagues from the Laboratory of Chemoinformatics at the University of Strasbourg. Particular thanks to my supervisors Professor Alexandre Varnek and Professor Jean-Marie Lehn, for their patience, a plethora of professional advice and also for their kindness. Special thanks to Dr. Gilles Marcou for our talks related to the research and educational sides of the work. Special *mulțumesc* to Dr. Dragos Horvath for all the tools that he has implemented, for his help with the scripts and last but not least for not only helping me to not forget the hardly-learned Romanian language but also for teaching me some new words. Dr. Igor Baskin, thank you for your advice and for our productive discussions while I was working on a diverse-library project. I appreciate the help of Dr. Fanny Bonachera and Dr. Olga Klimchuk in organizing my working process and documents and for their help in some bureaucracy unrelated to work. I am grateful to Dr. Arkadii Lin, Dr. Alexey Orlov, Yuliana Zabolotna and William Bort for our friendship and for all the discussions that we had during this time. I thank all the former PhD students from our lab: Dr. Pavel Sidorov, Dr. Timur Gimadiev and Dr. Marta Glavatskikh. I cannot find the words to express how deeply grateful I am to all the members of the lab for their support after my father's loss. Finally, I would like to thank all members of Lehn's laboratory, especially Dr. Artem Osypenko and Bohgdan Kozybroda, for our fruitful discussions on the "purely-chemical" topics. It was worth mentioning another person: Dr. Julien Diharce, a friend of mine who encouraged and supported me from my first year at the University of Nice. Last but not least, I would like to thank the Région Grand Est for the PhD fellowship.

Contents

1 Résumé en français

1.1 Introduction

Actuellement les bases de données chimiques incluent des millions de structures de composés chimiques [1, 2]. Grâce à la synthèse combinatoire et aux réacteurs en flux continu ce nombre augmente exponentiellement. Néanmoins ces chiffres sont « négligeables » en comparaison du nombre de composés que contiendrait l'espace chimique même en se limitant aux molécules d'intérêt thérapeutique, celui-ci étant estimé à 10^{33} [3]. L'exploration et l'analyse de cet espace permet aux chimistes de mieux comprendre les relations structure-activité ; de plus, grâce l'analyse des régions inexplorées de l'espace chimique facilite l'innovation, en particulier pour la recherche de nouveaux candidats médicaments.

L'une des approches qui permet d'effectuer une telle analyse est la méthode de cartographie topographique générative (Generative Topographic Mapping, ou GTM) [4]. Elle localise les structures chimiques, représentées par un espace de descripteur multidimensionnel initial sur un espace bidimensionnel plat, appelé une carte (**Figure 1.1**). La GTM établi en fait une correspondance entre une distribution de probabilité dans l'espace initial avec une distribution bidimensionnelle sur la carte. Cette dernière est quantifiée en des points spécifiques de la carte, appelés « nœuds ». À une molécule localisée dans l'espace initial correspond une distribution de probabilité sur la carte. Les coordonnées de la molécule sur la carte correspondent au centre de gravité de la distribution de probabilité qui représente le composé sur la carte. Inversement, à chaque lieu de la carte correspond une distribution de l'espace initial qui représente une population de structures chimiques. Précédemment, cette approche a été appliquée avec succès pour visualiser et

analyser des données chimiques [5] ainsi que pour préparer des modèles prédictifs de régression ou de classification [6].

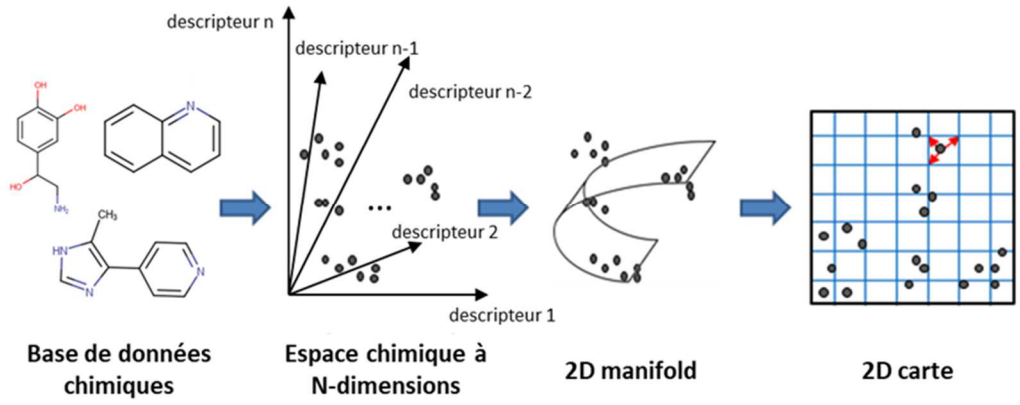


Figure 1.1 : Préparation d'une carte générative topographique (GTM) pour un espace chimique définie par les descripteurs moléculaires

L'objectif de cette thèse est d'explorer l'application de la méthode GTM à plusieurs tâches de la chémoinformatique : le criblage virtuel, l'analyse de l'espace chimique de systèmes complexes et la constitution d'une bibliothèque de composés chimiquement divers. La thèse est divisée en 6 Chapitres. Les chapitres 1 et 2 sont l'introduction et les méthodes, respectivement. Le chapitre 3 résume les résultats de l'utilisation simultanée de plusieurs cartes GTM dans le criblage virtuel de différentes cibles biologiques à partir de données issues de la base de données ChEMBL. Le chapitre 4 décrit le projet dédié à la conception assisté par ordinateur visant à trouver de nouveaux inhibiteurs de Bromodomaine. Le chapitre 5 est est dédié à la modélisation de la spéciation de bibliothèques combinatoires dynamiques [7, 8] en absence ou en présence d'un effecteur (protéine ou métal). Ici, la préparation de modèles prédictifs sont décrits, pour le calcul du logarithme de la constante d'équilibre $\log K_{eq}$ pour les réactions de formation d'imines, et pour la formation de complexes entre des effecteurs variés et des molécules organiques. Dans le même chapitre la constitution d'un jeu de données structurellement divers à partir d'une base de données contenant plus de 42000 composés à l'aide d'une GTM, ainsi que la

comparaison de sa performance avec la méthode traditionnelle MaxMin [9] est décrite. Dans le chapitre 6 les conclusions générales ainsi que les perspectives sont décrites.

1.2 Résultats et discussions

1.2.1 Application du modèle consensus GTM au criblage virtuel

Il a été démontré que la méthode GTM peut être utilisée pour préparer des modèles prédictifs de régression et de classification [6]. Dans ce projet, nous montrons que le consensus de cartes construites à partir de différents descripteurs moléculaires mais sur un même jeu de données, fournis des prédictions plus fiables par rapport à l'utilisation d'une carte unique basée sur un seul ensemble de descripteurs moléculaires. Nos études ont été menées sur les « cartes universelles » [10] qui sont capables de distinguer les composés actifs de composés inactifs pour plus de 600 cibles biologiques, simultanément [2]. Huit cartes universelles ont été obtenues à partir de différents espaces de descripteurs utilisés. Chaque carte universelle est capable à prédire l'activité de ligands pour plusieurs cibles avec de bonnes performances. Néanmoins des performances de prédiction pour une même cible divergent considérablement d'une carte à l'autre. Ces performances sont estimées par un paramètre statistique appelé « précision balancée » (Balanced Accuracy, ou BA) qui varie entre 0.5 (si les prédictions sont aléatoires) et 1 (si les prédictions correspondent à la l'expérience). Il a été trouvé qu'aucune des 8 cartes n'est capable de séparer à elle seule les composés actifs et inactifs pour toutes les cibles avec une BA suffisamment élevée. Toutefois, les prédictions de chaque carte peuvent être combinées en un consensus. Ainsi, il a été observé qu'un consensus de 7 cartes est suffisant pour prédire l'activité des molécules avec une haute fiabilité (BA>0.75) sur plus de 85% des cibles, simultanément (**Figure 1.2**). Ces cartes sont complémentaires, car les cibles moins bien prédites par une carte sont mieux prédites sur une autre, en exploitant les points de vus alternatifs que représentent les descripteurs moléculaires utilisés pour chacune.

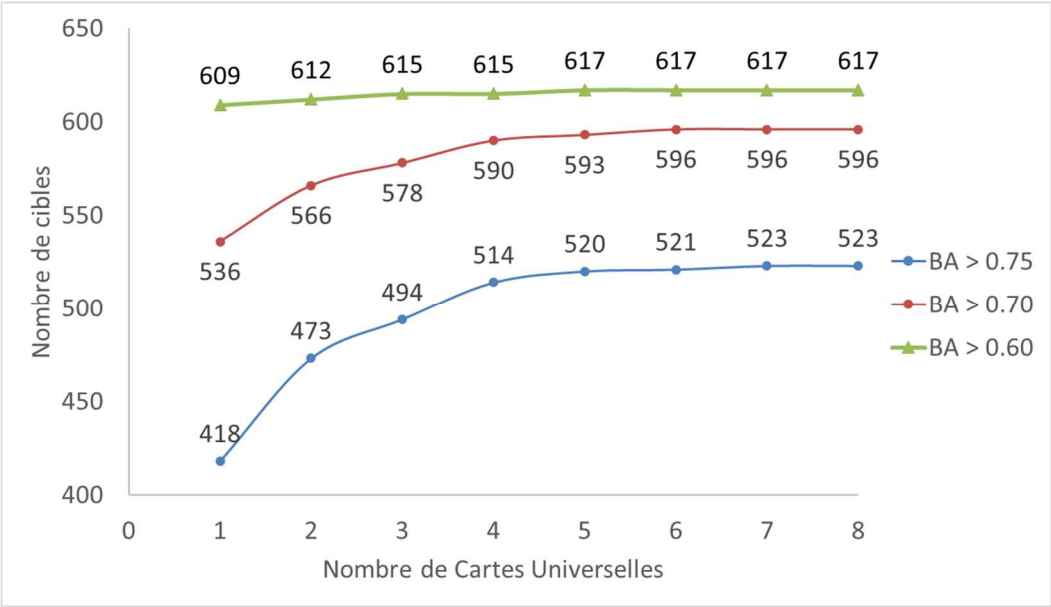


Figure 1.2 : La performance cumulée de cartes exprimée en nombre de cibles ayant un BA supérieur au seuil établi en fonction de nombre des cartes utilisées au criblage

1.2.2 Conception assistée par ordinateur de nouveaux inhibiteurs de Bromodomaine

L'étude par criblage virtuel (Virtual Screening, ou VS) décrite ici visait à identifier de nouveaux ligands de Bromodomaine BRD4. Elle s'est appuyée sur le contenu de bases de données publiques (ChEMBL, REAXYS) pour établir, dans un premier temps, un modèle prédictif de l'activité BRD puis, dans un second temps, l'utilisation de ces modèles pour la sélection de ligands putatifs. Différentes approches chémoinformatiques (SVM, pharmacophores, GTM, docking) ont donc été utilisées pour filtrer la collection de 2 millions de composés de la société Enamine. Ce partenaire industriel a ensuite testé expérimentalement un sous-ensemble de 2992 molécules de cette sélection.

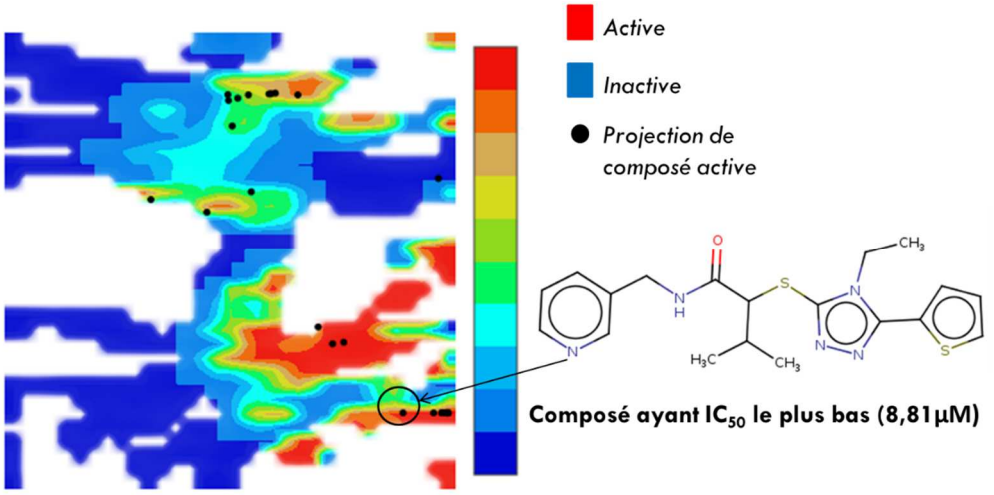


Figure 1.3 : Une des cartes génératives topographiques utilisées lors du VS. Les zones rouges et bleus sont peuplées par des composés actifs et inactifs, respectivement. Les régions ayant des couleurs intermédiaires correspondent aux zones peuplées simultanément par des composés de deux classes.

Ainsi, 29 composés actifs ont été confirmés après ces tests expérimentaux, ce que représente 1% des candidats sélectionnés, ce qui représente une amélioration d'un facteur 2,6 par rapport à une sélection aléatoire de composés. Ce succès est d'autant plus remarquable que les modèles ont été conçus pour prédire la concentration inhibitrice à 50% du BRD4 (IC_{50}) et que la société ENAMINE a utlisé pour la validation expérimentale, le Thermal Shift Assay. Quoique apparentées, ces deux mesures sont distinctes. Cet exemple illustre donc comment les modèles obtenus par apprentissage automatique sont en mesure d'identifier des relations structure-activité pertinentes pour être exploitées dans un contexte de recherche industriel.

1.2.3 Modélisation *in silico* des bibliothèques combinatoires dynamiques des imines

1.2.3.1 Application de la méthode GTM à la constitution d'une bibliothèque de composés chimiquement diverse

Les bibliothèques de composés chimiquement diverses sont particulièrement importantes pour recherche en chimie médicinale car un tel ensemble de composés vise à

tester le plus grand nombre d'hypothèses quand une nouvelle activité biologique est recherchée et qu'il existe peu de connaissances à priori pour faire des choix rationnels. Habituellement, une telle collection est obtenue à partir d'une sélection des composés les plus différents les uns des autres, à partir d'une grande chimiothèque de composés accessibles commercialement ou par voie de synthèse. La dissimilarité entre deux structures chimiques est assimilée à la distance séparant celles-ci dans l'espace chimique, c'est-à-dire la distance entre les deux vecteurs de descripteurs moléculaires qui les représentent.

Trois jeux de données ont été utilisés dans ce travail : l'un contenant 154 amines, l'autre – 277 aldéhydes et le troisième contenant 42658 imines qui sont les hypothétiques produits de réaction entre chacune des amines avec chacun des aldéhydes. Le but est de proposer une chimiothèque diverse de 225 imines. Deux algorithmes ont été comparés : l'algorithme « traditionnel » MaxMin et un algorithme innovant basé sur la GTM. L'algorithme MaxMin maximise les distances entre les individus de la chimiothèque diverse et sélectionne la molécule la plus éloigné de l'ensemble des molécules déjà choisis. Pour cette raison, son utilisation pour échantillonner une très grande chimiothèque conduit rapidement à des temps de calcul prohibitifs. L'algorithme innovant basé sur une carte GTM exploite les deux dimensions de la carte. En effet, sur un nombre réduit de dimension, une approche directe qui consiste à diviser la carte en cellules de surface égale puis à tirer au hasard des représentants dans chacune d'elles, est très efficace. La notion de distance entre individus dans la sélection est écartée, mais l'échantillon est bien plus représentatif. Ici, la carte était divisée en 225 zones égales, puis dans chaque zone un composé a été extrait aléatoirement.

Deux critères de performance ont été utilisés : distance moyenne de Soergel (<S>) et le taux de couverture de données. Le paramètre <S> est calculé comme la distance de Soergel moyenne entre tous les 225 composés sélectionnés. Le taux de couverture désigne le pourcentage de jeu de données initial (42658 imines) qui a un analogue parmi les 225

composés sélectionnés. Une sélection aléatoire de 225 imines a été réalisée pour servir de référence. Les résultats présentés dans la **Figure 1.4** montrent que MaxMin permet de sélectionner les bibliothèques plus diverses et, pour peu que le choix ait été réalisé sur les produits de réaction, la sélection offre une excellente couverture du jeu de données, mais cela nécessite beaucoup plus de temps et de ressources informatiques. Toutefois, le taux de couverture des chimiothèques sélectionnées à l'aide de la GTM (98%) est supérieur à celui de MaxMin, mais offre des garanties sur la distance séparant les composés sélectionnés. Comme la méthode basée sur la GTM est une méthode bien plus rapide que MaxMin, elle offre des perspectives intéressantes.

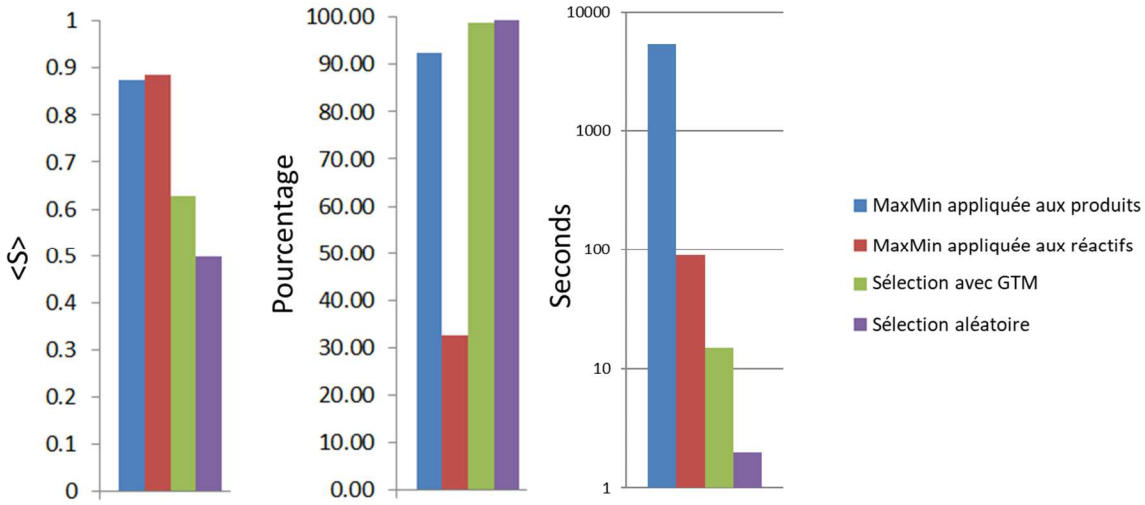


Figure 1.4 : La comparaison de la performance de GTM (en vert) en constitution de la bibliothèque diverse des imines avec la méthode traditionnelle (MaxMin) appliqué aux jeux de données de réactant (rouge) et produits (bleu), ainsi qu'avec le choix aléatoire (violet). Le diagramme de gauche représente la distance de Sorgel moyenne, celui de milieu la couverture de l'échantillon et celui de droite la vitesse de constitution d'une bibliothèque sélectionnée.

1.2.4 Modélisation des équilibres dans une bibliothèque combinatoire dynamique

Une bibliothèque de composés qui peuvent interagir de manière réversible les uns avec les autres s'appelle une bibliothèque combinatoire dynamique (*Dynamic*

Combinatorial Library, ou DCL). Tous les constituants d'une DCL [7] sont en équilibre et leur distribution est déterminée par leur stabilité thermodynamique. Une telle bibliothèque requière que les réactions impliquées dans sa formation soient réversibles et que les produits de réaction soient en proportions comparables. Lorsqu'une DCL est exposée à un « effecteur » externe (comme des protéines ou des métaux) qui agit sélectivement sur un ou plusieurs membres de la DLC, l'équilibre se déplace selon le principe de Le Chatelier et la composition du mélange change. Le but de ce projet est de pouvoir prédire la spéciation, c.à.d., la composition d'une solution à partir des structures chimiques des composants de DCL. Ce calcul passe par une étape d'estimation des constantes d'équilibre pour toutes réactions de la DCL. Dans ce projet les DCL basées sur la réaction de formation des imines ont été considérées en présence ou en absence de l'enzyme anhydrase carbonique 2 (AC2). Pour cela, des modèles capables de prédire les constantes de formations des imines (K_{eq}) d'une part, et les constants de complexation de l'AC2 avec des imines, des aldéhydes et des amines (K_i), d'autre part ont été développés. Les données expérimentales sur les constantes de formations des imines dans le chloroforme ont été obtenues par RMN dans le laboratoire du prof. J.-M. Lehn pour 276 réactions.

Le modèle consensus de régression obtenu pour $\log K_{eq}$ regroupe 16 modèles individuels SVR [11, 12] (Support Vector Regression). La performance prédictive de ce modèle évalué en validation croisée répétée 5 fois (5*5CV) est bonne : le coefficient de détermination $R^2 = 0.93$ ($R^2 = 1$ désigne un modèle parfait) et l'erreur quadratique moyenne RMSE = 0.63 unité $\log K_{eq}$. Le domaine d'application du modèle est très large : pour une chimiothèque virtuelle de 120000 imines issue de réactions entre 300 aldéhydes et 400 amines les plus cités dans la littérature, ce modèle peut être appliqué à plus de 80700 réactions. Afin de pouvoir utiliser ce modèle dans les solutions aqueuses, le rapport de $\log K_{eq}$ dans l'eau et dans le chloroforme a été estimé en utilisant les énergies libres de solvatation des réactants et des produits dans ces deux solvants. De plus, l'espace chimique de la chimiothèque virtuelle de 120000 imines a été analysée à l'aide de GTM. Un modèle

SVR pour des logarithmes de constantes de complexation de molécules organiques avec l'AC2 ($\log K_i$) dans un milieu aqueux a été entrainé sur les données expérimentales extraites de la base de données ChEMBL. La performance prédictive de ce modèle est assez bonne (R^2>0.7). Pour preuve de concept, l'ensemble de constantes K_{eq} et K_i prédites pour différents équilibres individuels a été utilisé pour estimer les concentrations pour une DCL de 2 amines et 2 aldéhydes en absence et en présence de la protéine.

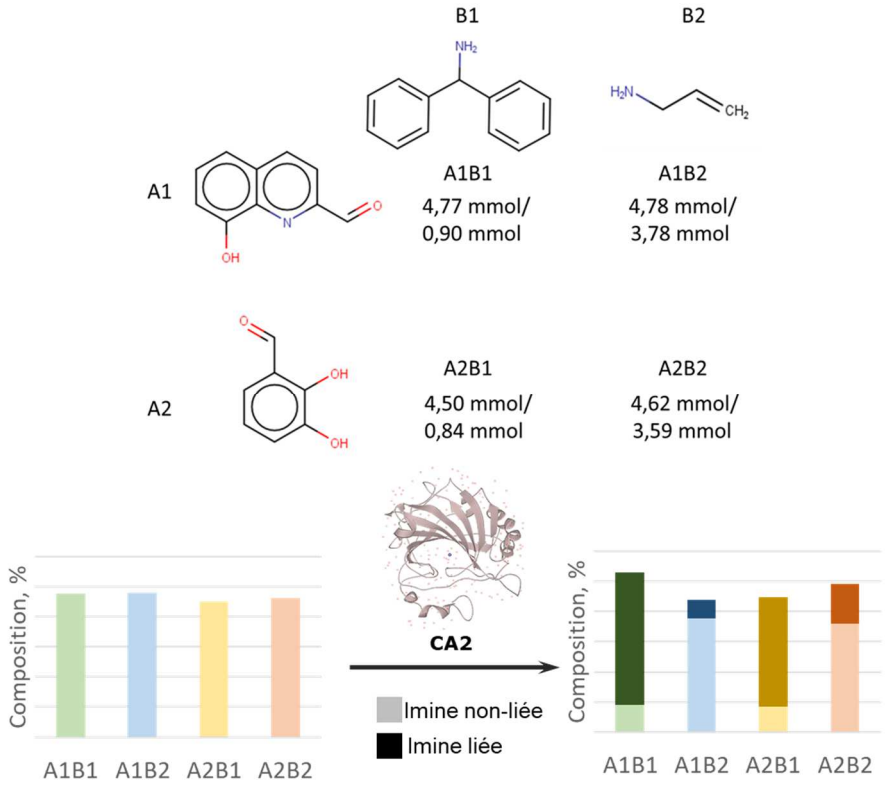


Figure 1.5 : Spéciation dans un DCL hypothétique de 2 amines et 2 aldéhydes dont la spéciation avant / après l'addition de l'anhydrase carbonique est estimée en utilisant les constantes d'équilibre prédites.

1.3 Conclusions

La méthode GTM a été utilisée avec succès pour l'analyse de l'espace chimique, le criblage virtuel et la constitution d'une chimiothèque diverse. Nous avons démontré qu'un consensus de modèles individuels GTM fournis des prédictions plus fiables que chaque

modèle considéré séparément. Ainsi, un modèle de classification basé sur l'application simultanée de sept cartes universelles peut distinguer les composés actifs et des inactifs sur 617 cibles biologiques avec une BA> 0,6 et pour 523 cibles avec une BA> 0,75.

Afin de rechercher de nouveaux inhibiteurs de BRD4, un ensemble de méthodes chémoinformatiques (SVM, pharmacophores, GTM, docking) a été utilisé pour cribler 2 millions de composés. Ceci a permis d'identifier 29 inhibiteurs de BRD4 confirmés expérimentalement.

Pour la première fois, la GTM a été utilisée pour la sélection d'un échantillon divers d'une chimiothèque. Cette méthode a été comparée à la méthode classique MaxMin lors de la sélection d'une chimiothèque diverse d'imines. Les résultats montrent que la GTM a un comportement spécifique comparé à MaxMin. Le taux de couverture du jeu de données initiales par la GTM (98%) est excellent tout en garantissant que les structures chimiques sélectionnées soient bien séparées. De plus, les calculs avec GTM sont beaucoup plus rapides par rapport à ceux de MaxMin.

La méthode SVR a été utilisée pour préparer des modèles de régression pour la constante de formation d'imines et la constante de complexation de molécules organiques avec l'enzyme anhydrase carbonique 2 ($\log K_i$). La combinaison de ces modèles permet de modéliser la spéciation d'une DCL d'imines sans et avec l'enzyme.

1.4 Liste des presentations

1. Zabolotna Y., **Casciuc Iuri**, Horvath D., Marcou G., Bajorath J., Varnek A. Generative Topographic Mapping in Virtual Screening: why ensemble of maps is needed? Strasbourg Summer School in Chemoinformatics (26/06/2018) (**poster**)
2. Gimadiev T.R., Madzhidov T.I., Nugmanov R.I., Klimchuk O., **Casciuc Iuri**, Baskin I.I., Antipin I.S.,Varnek A.A. Reaction data treatment by means of

Condensed graph of reaction Strasbourg Summer School in Chemoinformatics (26/06/2018) (**poster**)

3. **Casciuc Iuri**, Horvath D., Varnek A. Computer-aided design of new inhibitors of Bromodomain, Journée Scientifique Doctorant et Master UMR7140, Strasbourg 17/04/2018 (**oral**)
4. **Casciuc Iuri,** Zabolotna Y., Horvath D., Marcou G., Varnek A., Virtual Screening with Generative Topographic Maps: how many maps are needed? Journée Scientifique Doctorant et Master UMR7140, Strasbourg 07/05/2019 (**oral**)
5. **Casciuc Iuri,** Zabolotna Y., Horvath D., Marcou G., Varnek A., Pharmacological profiling with universal Generative Topographic Maps, 9e journées de la Société Française de Chémoinformatique, Paris 22/11/2019 (**oral**).

1.5 Liste des publications

1. Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis. *Mol. Inform.* **2019**, *38* (4), 1800104.
2. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J. Chem. Inf. Model.* **2018**, *59* (1), 564–572.
3. Casciuc, I.; Horvath, D.; Gryniukova, A.; Tolmachova, K. A.; Vasylchenko, O. V; Borysko, P.; Moroz, Y. S.; Bajorath, J.; Varnek, A. Pros and Cons of Virtual Screening Based on Public “Big Data”: In Silico Mining for New Bromodomain Inhibitors. *Eur. J. Med. Chem.* **2019**, *165*, 258–272.

2 Introduction

Nowadays, chemical databases include millions of structures of chemical compounds [1, 2]. Thanks to combinatorial synthesis and continuous flow reactors, this number increases exponentially. However, these numbers are "negligible" in comparison to the number of compounds that the chemical space would contain even if it were limited to molecules of therapeutic interest, this being estimated at 10^{33} [3]. The exploration and analysis of chemical space allow chemists to understand structure-activity relationships better; moreover, the study of the unexplored regions of the chemical space facilitates innovation, in particular for the research of new drug candidates.

When it comes to chemical data visualization, analysis and modeling, the applied methods could be either descriptor-based or graph-based. In a descriptor-based approach, a compound is represented as a vector of descriptors, and each descriptor is describing the molecule in terms of physical or chemical properties (molecular weight, logP) and/or purely structural (number of atoms, types of bonds). These vectors of descriptors are serving as input for machine-learning algorithms. Some of these algorithms, like dimensionality reduction techniques, are designed specifically for the visualization and modeling of multi-dimensional data. Multiple dimensionality reduction techniques are reported in the literature: Principal Component Analysis (PCA [13–15]), Multi-Dimensional Scaling (MDS [16]), Sammon mapping, Self-Organizing Maps (SOM) [17] due to their efficiency. In 2001 T. Oprea [18] proposed a new term – "… *chemography*, by analogy with geography, as the art of navigating chemical space." In other words, Oprea suggested the usage of maps for navigation in chemical space. Such maps of chemical space, by analogy with the world map, should posses a universal character, i.e. the compounds are defining the "contours of the continents" while their properties will be defining the colors. The

above-mentioned dimensionality reduction methods are efficient, nevertheless they are not perfect. For example, SOM is producing a 2D map which is based on a non-linear model, PCA is also able to produce a 2D map if two principal components will be taken. However, PCA is efficient with datasets having internal linear correlations [19], but it could fail while representing vast multidimensional data [20]. MDS is another dimensionality reduction technique that is also linear that is using Euclidean distances [21]. Sammon maps do not allow the addition of the new data on already existing map, forcing the user to rebuild the map if the new compounds should be added [22].

On the other hand, graph-based approaches represent a molecule as a graph, the atoms and bonds corresponding to the graph's nodes and edges, respectively. One way to work in graph-based chemical space is to rely on the concept of a scaffold that is defined as the "core part" of the structure with all the terminal chains removed [23]. These could be regrouped in a so-called hierarchical scaffold tree, which allows the data visualization and modeling [24].

Generative Topographic Mapping (GTM) [4] is a probabilistic extension of SOM that considers the likelihood of the training data as the objective function. Moreover, unlike to SOM, the object is not associated with one particular node. In GTM, the objects are represented as a probability distribution over all the nodes, thus creating a vector of probabilities for each data point. This vector of probabilities is used for data visualization as well as for the building classification and regression models. GTM is a versatile method that can be applied in different everyday tasks of chemoinformatics like data visualization, libraries comparison, QSAR, *de novo* design; therefore, GTM could be compared to a Swiss army knife.

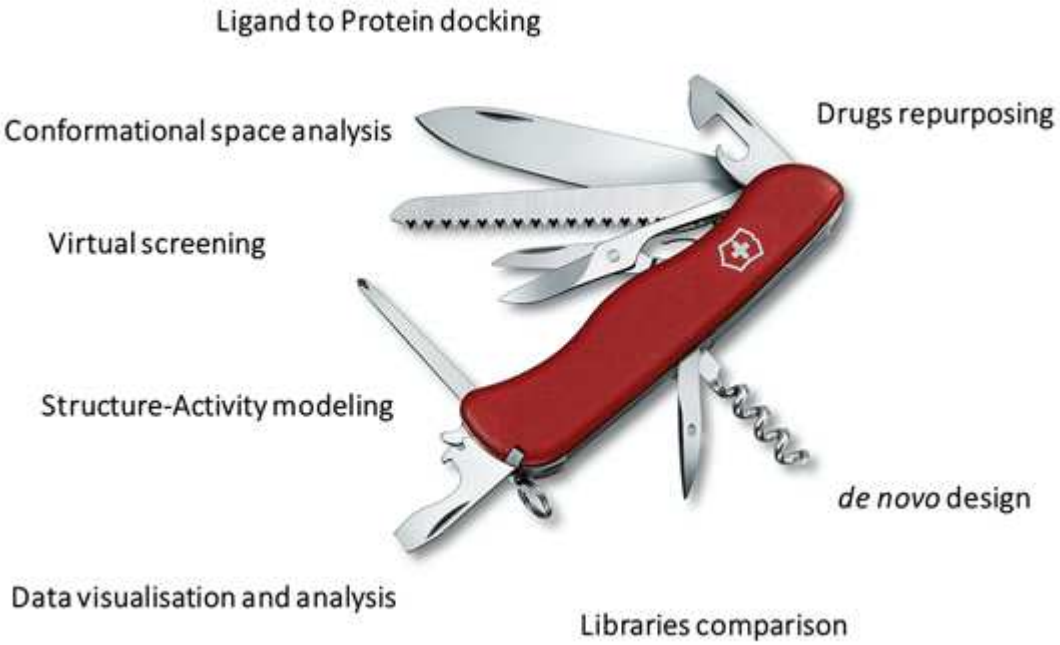


Figure 2.1: Ares of successful applications of GTM.

GTM has been used in several projects for data visualization. The hierarchical GTM algorithm [25] has been used [24] for the visualization of the active/inactive classes distribution in five different datasets issued from high-throughput screening. Large datasets of compounds (2.2M compounds) have been visualized for the first time by applying the incremental GTM [5]. In the same work, a comparison of libraries has been made. Each library was considered as a single object of cumulated responsibilities or properties. A *Responsibility Pattern* (RP) term has been introduced by Klimenko et al. [26]; RP allowed to automatically detect and extract the compounds similar in the latent space. The concept of "privileged substructures" (PSM) was initially introduced by BE. Evans et al. [27], referring to core structures that are recurrent in compounds active against a given target family and, therefore, associated with that biological activity. This approach has been applied in the analysis and modeling of antimalarial compounds [28]. PSM has been modified by applying the retrosynthetic rules (RECAP) [29]. The authors tried to extract the "frequent" RECAP cores to identify PSMs for inhibitors of protease, kinase and GPCRs.

In recent work, GTM has been applied to visualization, analysis and comparison of the compounds tested against virus species, representatives of the *Coronaviridae* family [30].

GTM has been successfully applied for QSAR and QSPR modeling. Kireeva et al. [31] used GTM-based classification models for the prediction of the melting point of ionic liquids. Gaspar et al. have used GTM-based regression models [6] for the modeling of stability constants for metal binders, the activity of thrombin inhibitors and aqueous solubility. In multiple works, GTM has been compared to other popular machine-learning methods: SOM [32], Random Forest [33], Partial Least Squares [34], M5P regression tree [35], SVM [11]. It has been shown across many projects that GTM is a method that can compete in terms of the performance of produces target- and property-specific models with other machine-learning methods. In their work, Sidorov et al. [10] have shown that GTM can be successfully used as a multi-target predictive model. For instance, a dataset of 1.3M ChEMBL compounds (version 20) corresponding to 410 targets have been covered, and approximatively 80% of these targets have been predicted with relatively high Balanced Accuracy (> 0.7). Lately, Lin et al. [36] have applied the same protocol on the ChEMBL 23 dataset and benchmarked the obtained model with popular machine-learning methods.

Conformational sampling plays an important role in medicinal chemistry. Horvath et al. have described the application of GTM to conformational sampling [37]. In this work, a set of conformers with previously calculated total energies has been used (calculations were done using the general AMBER force field [38]). Torsion angles and some non-bonded contact energies have been used as descriptors to describe the conformers interaction fingerprints. The obtained map can be used for visualization and analysis of the "training" conformational space as well as to predict the energies for new conformers. This approach has been applied to the conformational space of dipeptides [39]. It was also used in a docking study of the ATP-binding site of CDK2 [40]. The maps have been trained to be

able to discriminate native from non-native ligand poses as well as to distinguish the potency of ligands.

GTM has also been applied in *de novo* design. For the first time in 2014, Mishima et al. [41] used GTM for an assessment of biological activities for virtually enumerated structures. Another attempt of the compound generation with specific activity(ies) has been made with Stargate GTM [42]. Stargate GTM uses two manifolds: one is built in the descriptor space and another in the activities space; thus, the two spaces are bound. A specific mapping function allows to "warp" from the activities space to descriptors space, therefore identifying the values of the desirable descriptors. Once the values of the descriptors are found, it is needed to generate structures with high similarity to the detected descriptors vector, thus assuming that the generated structures would posses the searched activity. Recently GTM has been combined with auto-encoder, where the map was trained on the generated latent descriptors. Sattarov et al. [43] used this approach to generate and analyze the binding potency of ligands of Adenosine A2a receptors. A similar approach was applied to the discovery of novel chemical reactions [44]. The authors have trained a sequence-to-sequence autoencoder on the USPTO [45] reaction database. The autoencoder latent space has been visualized using GTM, the zones of the map populated by Suzuki reactions have been targeted. Many of the generated chemical reactions possessed reaction centers not present in the training set.

During this thesis, a broader exploration of GTM capabilities in chemoinformatics tasks such as virtual screening and diverse library selection has been done. Sidorov et al. [10] have successfully built a *universal map* – a GTM-based multi-target classification model. In the first project of this thesis described in chapter 4, entitled "Consensus modeling using universal maps", we have applied the same model building protocol on the data extracted from ChEMBL v.23 (1.5M compounds with known activities on 618 targets). In this project, several universal generative topographic maps have been obtained, each

map being built in different descriptor spaces (hence encoding different distinct structural features). For each target-specific subset of 618 targets, the balanced accuracy (BA) has been computed; the score used to quantitatively describe the predictive performance of the map was calculated by averaging the BA over all the 618 target-specific subsets. The obtained universal maps have shown similar scores. The results are shown in chapter 4, and they answer on the following questions: i) For a virtual screening task, is it better to use one sole "best map" or several maps in consensus? ii) If the latter – how many maps should be applied?

The second project is presented in chapter 5, "*In silico* mining for new Bromodomain 4 inhibitors". Bromodomain 4 (BRD4) [46, 47] can be considered a difficult target for virtual screening because of its flexible structure – 2 α-helixes bound with two loops. Known inhibitors of BRD4 usually form 1 H-bond with the target, the rest of the protein-ligand interactions being hydrophobic [46]. This project was carried out in collaboration with the Enamine company [48], its goal was to find new inhibitors of BRD4. Two public sources of data have been used to form a training set – Reaxys and ChEMBL. The obtained models were used for virtual screening of 2M compounds that are available at Enamine. Our collaborators agreed to test 3000 compounds; hence our goal was to find the top 3000 compounds that are most likely to be BRD4 inhibitors. Here, GTM was used in rather complex virtual screening funnel, including SVM models [12], ligand-based pharmacophores [49, 50] and docking [51].

The third project is described in chapter 6, "*In silico* speciation assessment of Dynamic Combinatorial Libraries". In this project, GTM performance in diverse library selection has been compared to the classical dissimilarity-based method – MaxMin [9]. Dynamic combinatorial libraries (DCL) [7, 8] are the cornerstone of the dynamic combinatorial chemistry. It relies on the reversible nature of the involved reactions between the constituents of the libraries. Briefly, a DCL is usually represented as a solution of *m*x*n*

reactants that can reversibly interact one with another, leading to a formation of multiple reaction products and their thermodynamic stability dictates their distribution. An external effector (e.g., biological target) is introduced in DCL; it can lead to a global change of the distribution of the products in solution according to Le Chatelier principle if the effector will selectively bind to one (or a few) members of the DCL. It can be done in order to find the "best binder" with a given target. In this case, the reaction products have to be in almost equal proportions, because for a solution where one or a few constituents are overrepresented, the preferred interaction of a minor DCL constituent with the target may not be strong enough to overturn the equilibrium. There is software that can predict the numerical distributions of the amount of substances in the solution. However, all of them require precise thermodynamic data on equilibrium constants of the involved reactions, which complicates their usage because of the availability of the thermodynamic data for the specific case. One way to overcome this constraint is the usage of QSPR models that will predict the equilibrium constants, but again for the training of the QSPR model, the data is needed. In this project, we present the first steps toward *in silico* DCL modeling on the example of imine-based DCL with human Carbonic Anhydrase II as the effector. The presented workflow requires data on both equilibrium constants of the involved reactions and binding affinities with the biological target. The data on binding affinities can be extracted from public databases like ChEMBL, but there is less data on equilibrium constants of reactions. We have selected a diverse library of imines and our collaborators from prof. Lehn's laboratory synthesized and measured the equilibrium constants for over 250 reactions corresponding to the selected pool of imines, which became the training set in this study.

3 Methods

In this section of the thesis, several aspects of the applied methods will be discussed. First of all, the basics of QSAR methodology are described, which include a brief description of the QSAR paradigm, some words on different types of descriptors and popular machine learning methods used in model building. It will be followed by a description Support Vector Machine (SVM) and Generative Topographic Mapping (GTM) used as machine-learning methods in the presented projects.

3.1 QSAR /QSPR methodology

Quantitative structure-activity relationship (QSAR)/Quantitative structure-property relationship (QSPR) modeling is one of the cornerstones of the chemoinformatics tools regularly used in many fields such as medicinal chemistry and material science [52]. The principle of the modeling is to find a mathematical function that relates a chemical structure to the studied property (such as logP) or activity (for a given biological target). Such modeling implies that the structural information of the molecule is encoded in numerical form – in a vector of *molecular descriptors* [53], and the values of these molecular descriptors are used to define the position of the compounds in the *chemical space*. All of the above said could be expressed by the following equation:

$$Property_{molecule} = f(structure_{molecule})$$

To model any possible property/activity, one needs a way to encode all (or at least most of) the essential structural information and a big enough dataset of compounds with known activity. Last but not least, a machine learning algorithm that will eventually build a predictive model.

When it comes to the calculation of molecular descriptors, there are several ways to do it. For instance, **1D molecular descriptors** are directly obtained from the chemical formula of the compound. These descriptors are the most straightforward descriptors related to the molecule's "fundamental properties," such as the number of atoms and molecular weight. An obvious flaw of these descriptors is the impossibility to discriminate between isomers. **2D molecular descriptors** are based on a two-dimensional representation of the molecule (*a molecular graph*). This representation provides the information on atoms connectivity, therefore overcoming the flaw of 1D descriptors. Topological indices and molecular fragments are good examples of this type of descriptors. **3D molecular descriptors** are obtained from the 3D structure of the compound. They include the quantitative values obtained by quantum mechanics (such as HOMO/LUMO energies of the compound, dipole moment or electrostatic potential), ovality of the compound and van der Waals volume.

In this work, **ISIDA** [54–56] descriptors are used. They are 2D descriptors that encode a compound structure by counting the number of occurrences of different substructural fragments. These fragments could be linear sequences, augmented atoms (central atoms with their environment) or triplets that encode the compound's atoms and/or bond types. In addition to this, the fragments may be colored, adding some additional information: pharmacophoric types of atoms, formal charges force-field atom types, etc.

The dataset that will be used for model training [57] is called the *training set*. The models can be divided into two groups basing on the modeled property or activity: *regression model* (when the modeled/predicted property is a numerical value) and *classification model* (when the modeled/predicted property is categorical). The machine learning algorithms that require a known (experimental) property are called **supervised learning** algorithms. A non-exhaustive list of these methods includes Multilinear Regression (MLR) [58], Random Forest (RF) [33], Support Vector Machine (SVM) [11]

and Artificial Neural Network (ANN) [59]. Supervised modeling occurs when the modeled property/activity is known for all the entries in the training set, and the training *per se* consists of fitting a function in order to minimize the prediction error. Another type of machine-learning algorithms is **unsupervised learning**, where the compound's property is not used. Usually, the unsupervised methods are used in order to extract some "inherited" structural information basing on data distribution, to describe and interpret the data and, in some cases, to visualize it. This class of methods includes clustering (hierarchical clustering [60], k-means [61]) and dimensionality reduction (Principal Components Analysis (PCA) [14], Self-Organizing Maps (SOM) [17], Generative Topographic Maps (GTM) [4]).

An important parameter of any model is its quality [62]. Usually, for regression models, Root Mean Squared Error (RMSE) and determination coefficient (R^2) are used. RMSE is calculated according to the formula below, where N is the number of compounds, $y_{exp,i}$ and $y_{pred,i}$ are experimental and predicted property values of ith molecule respectively:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{exp,i} - y_{pred,i}\right)^2}{N}}$$

R^2 estimates the correspondence of experimental and predicted values. The maximal value of R^2 is 1, which corresponds to an ideal fit of the model, i.e., all the predicted values are equal to experimental ones. Lower values of R^2 correspond to a worse model and "acceptable" value of $R^2 > 0.5$ [63]. Determination coefficient is calculated by the formula below, where $y_{exp,i}$ and $y_{pred,i}$ are experimental and predicted property values of ith molecule respectively and $\langle y_{exp} \rangle$ is the average property value of the dataset:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{exp,i} - y_{pred,i}\right)^2}{\sum_{i=1}^{N}\left(y_{exp,i} - \langle y_{exp} \rangle\right)^2}$$

When it comes to classification models, the model's quality is related to the number of compounds with the correctly assigned category. Usually, the classification tasks are

reduced to binary classification, where the compounds are split into 2 classes (active/inactive). In this case, to evaluate the model's quality, a *confusion matrix* is used. It represents a table where the predicted class of the compounds is matched with their actual value. The cells of the table contain the number of compounds that have had a correct ("True") or an incorrect ("False") class assignment that is denoted as "Positive" and "Negative".

		Actual	
		Class 1	Class 2
Predicted	Class 1	True Positive (TP)	False Positive (FP)
	Class 2	False Negative (FN)	True Negative (TN)

Using the confusion matrix, one can calculate the **balanced accuracy (BA)**, which is a numerical characteristic of a classification model. BA takes the rate of correct predictions of both classes in equal proportions, and it takes values from 1 (ideal case) to 0.5 (random predictions). BA is used for the datasets where the predomination of one class over another is observed:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{FP + TN}\right)$$

Another essential factor that is intrinsically related to model quality is its applicability domain (AD). The machine-learning methods perform well in interpolating tasks because the model is trained on a limited set of compounds. The model provides *reliable* predictions for the compounds that are similar to the ones from the training set. In every project of this thesis, ISIDA descriptors were used; fragment control was generally used as AD [55]. This approach considers any compound to be out of the AD if it has at least one descriptor (i.e., substructural molecular fragment) that was not present within the training set.

One cannot tell how well a model will perform in a "real-life scenario", nonetheless basing on its performance during *cross-validation* (CV) [64], one can estimate the model's behavior. In this thesis, k-fold cross-validation has been used. It consists of the dataset division into k subsets (folds); k-1 folds are used alternatively for model training, and the last fold is used for testing. At the end of this procedure, every molecule has received a prediction exactly once, and these values are used for the calculation of the model's performance parameters (BA for classification, RMSE and/or R^2 for regression).

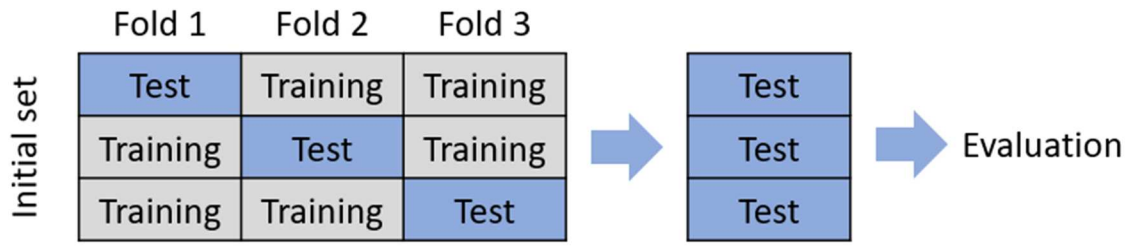


Figure 3.1: Schematic representation of a 3-fold cross-validation procedure. The initial dataset is divided into three parts; on each fold, the model is trained on two parts, and it is applied to the associated test part. At the end of the procedure, all the predicted values of "test" subsets are used for the model's evaluation.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a popular supervised machine learning method that can be used in classification and regression tasks. The method has been developed and published by Vapnik in 1995 [11] following the idea to find a hypersurface that separates two classes of objects with the "gap" (called *margin*) as wide as possible. New objects that will be mapped to the same space will be assigned to one of the classes according to the side of the surface where they fall on. Additionally, SVM can non-linearly map the input vectors into higher dimensional feature space using a kernel function (the so-called *kernel trick*) and then to linearly separate them in this new feature space.

However, perfect separation of classes is not possible, since "an ideal SVM" should produce a hyperplane able to completely separate the objects of different classes. Such hyperplane may result in an overfit model; therefore, new objects might be wrongly classified. The SVM algorithm is maximizing the margin, and in the meantime, it minimizes the misclassifications using the *slack variable* ξ_i. The goal of the algorithm becomes to maintain the slack variable at a minimal value while maximizing the margin; therefore the constraint and objective function becomes

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i, \forall \mathbf{x}_i \; \xi_i \geq 0$$

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i \text{ , with C being trade-off margin}$$

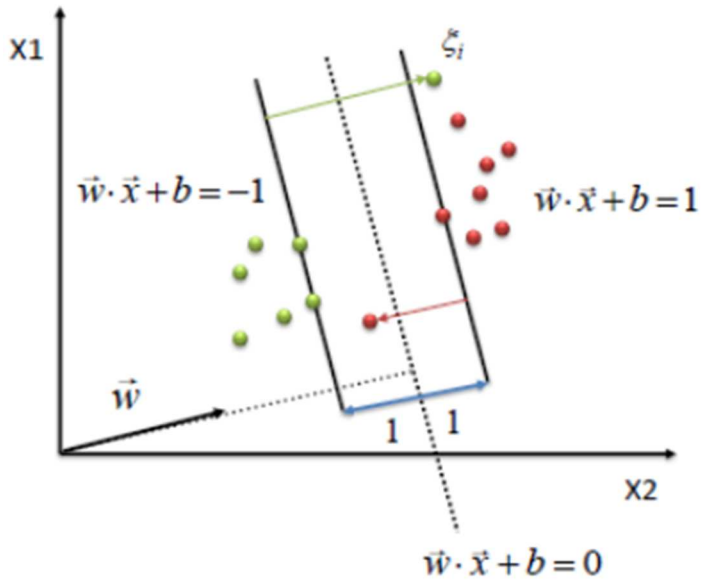


Figure 3.2: A schematic representation of a separable problem in 2D space. The margin providing the widest separation as well as the hyperplane are defined by the support vectors [65].

In 1996 Vapnik proposed a support vector regression (SVR) [66] as the development of the SVM method for the prediction of a continuous variable. SVR maintains the main feature that characterizes the original algorithm – the maximal margin. In cases of regression task ($y = \mathbf{w} \cdot \mathbf{x} + b$), a margin of error tolerance ε is set in approximation to the hyperplane and the algorithm is minimizing:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i + \xi_i^*)_i$$

Where C is the *cost* that defines the penalty for objects whose predicted value deviates from the experimental value for more than ε. The constraints become:

$$y_i - (\mathbf{w} \cdot \mathbf{x_i} + b) \leq \varepsilon + \xi_i$$

$$(\mathbf{w} \cdot \mathbf{x_i} + b) - y_i \leq \varepsilon + \xi_i^*$$

$$\xi_i^*, \xi_i \geq 0$$

3.3 Generative Topographic Mapping

Generative Topographic Mapping (GTM) has been presented by Bishop in 1998[4] as a method of data visualization. GTM is a probabilistic extension of Self-Organizing Maps (SOM) [17], but unlike SOM, it considers the likelihood of the training data as the objective function. Moreover, in GTM, a single object is not associated with one particular node (and its neighbors), but it is associated with the probability distribution over the entire latent space.

GTM finds a representation of the data distribution in the initial *D*-dimensional data space on an *L*-dimensional hypersurface called a *manifold*. Although the dimensionality of the manifold is usually *L*=2, *L* can take any natural value from 1 to *D*. To map the objects from the initial space to the latent space, a mapping function **y**(x, **W**) given as a grid of M Gaussian activation functions (radial basis functions, or RBF) is applied:

$$y_d(x, \mathbf{W}) = \sum_{m=1}^{M} \mathbf{W}_{md} \exp\left(\frac{\|x - x_m\|^2}{2\sigma}\right)$$

Where M is the number of RBFs, D is the initial space dimensionality, **W** is the matrix ($M \times D$) of weights connecting the initial data space and RBF grid, x_m is the center of the m-th RBF; d takes values from 1 to D, M and σ being the parameters of the method.

Since the function **y**(x, **W**) is smooth (and therefore continuous), the so-called *neighborhood behavior* is observed – the objects that are close in the initial space remain neighbors in the latent space. Every node of the grid is associated with the center of a normal distribution function with inverse variance β, that corresponds to the sampling of the random variable t with the following probability density function:

$$p(t|\mathbf{W},\beta) = \frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|t - y(x_k,\mathbf{W})\|^2\right)$$

Where K is the number of nodes, x_k the coordinate of the k-th grid node in the latent space, $y(x_k,\mathbf{W})$ – the coordinates to which it has been mapped in the initial data space and t covers the whole data space representing any object. The logarithm of the probability with which the data could be generated is called *log-likelihood*, and it is denoted as LLh. The higher the value of LLh, the better the manifold represents the data. LLh is used as the maximized function in the Expectation-Maximization (EM) algorithm. LLh is a function of two parameters, **W** and β:

$$\mathrm{LLh}(\mathbf{W},\beta) = \sum_{n=1}^{N}\mathrm{LLh}_n = \sum_{n=1}^{N}\ln\left(\frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|t_n - y(x_k,\mathbf{W})\|^2\right)\right)$$

t_n is the position of the n-th object in the initial data space. Every object has a non-zero probability of being mapped into any node of the grid. This probability is called *responsibility,* and it is calculated using Bayes's theorem:

$$r_{nk} = p(x_k|\,t_n) = \frac{\exp\left(-\frac{\beta}{2}\|t_n - y(x_k,\mathbf{W})\|^2\right)}{\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\|t_n - y(x_k,\mathbf{W})\|^2\right)}$$

For every object, the responsibility is normalized over the grid of nodes; therefore, the sum of responsibilities for a given object is 1. This vector is used for visualization **and modeling** purposes.

Standard GTM (sGTM) [67] algorithm is computationally expensive in the case when the number of the compounds and descriptors is relatively high since it takes too long to calculate the Euclidean distances for all the pairs "object-node". For a small dataset of 1000 compounds having 500 descriptors, the computing time of all the distances between all the compounds and 900 nodes is around 4.5s on a single CPU (Intel Core i7 - 6700HQ) [68]. However, the computational time rises to 135s for the dataset of just 30k compounds. Since this procedure is done at each iteration of the EM algorithm, it renders the sGTM algorithm rather slow. Moreover, when it comes to extensive collections of data, a memory problem might appear. To overcome the constraints of sGTM, an incremental variation of GTM (iGTM) [5] has been proposed. In this case, the initial dataset, instead of being processed as a whole, is divided into many blocks, then each block is processed consecutively. The manifold will be trained on one single block at a time, which accelerates the procedure. In the current work, iGTM has been used almost exclusively.

Recently a new approach in GTM methodology has been proposed – parallel GTM (pGTM) [68]. The main idea of pGTM is to extend iGTM over multiple CPUs. It is done by manifold initialization over all dataset, then this dataset is split into blocks, and the initialized manifold is fit on each block independently and simultaneously. Thus, each block provides an intermediate manifold that has been fitted on a given part of the data. In the end, an averaging of the $\mathbf{W}$ and β over all blocks is done.

3.3.1 GTM as a visualization method and modeling tool

For the visualization and modeling of the data, GTM uses the concept of *landscape*. For every compound, GTM generates a vector of normalized responsibilities that can be treated in the same way as molecular descriptors. The number of these descriptors will be equal to the number of the nodes used in the GTM grid. The landscape is obtained by adding a specific property of interest for the given dataset as a third dimension of the 2D

map. Three types of landscapes can be defined: *class landscape* [69], *property landscape* [6] and *density landscape* [68].

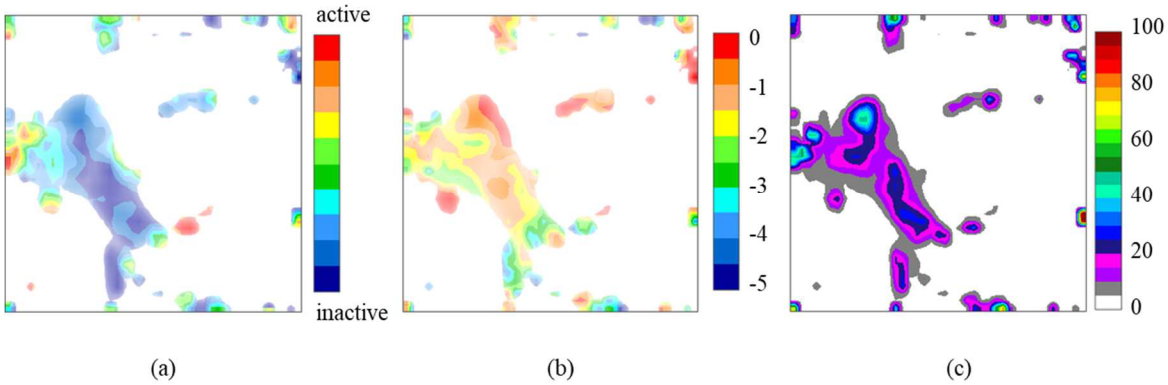


Figure 3.3: Examples of three types of landscapes. The GTM has been applied to the dataset containing 6.7k compounds with known activities for vascular endothelial growth factor receptor 2 (CHEMBL279). Class landscape (a) shows the distribution of compounds of two classes: active (red) and inactive (blue). logS (solubility) has been used as a 3rd axis to build the property landscape (b). Density landscape (c) shows the population of the zones of the map.

The class landscape represents the **GTM-based classification model**. To obtain a class landscape, a class is attributed to each node of the grid by averaging the responsibilities $r_{kn}(C_i)$ over the number of compounds Nc_i of the training set that belongs to the i-th class. The conditional probability $P(k|C_i)$ of the new object close a node k is calculated:

$$P(x_k|C_i) = \frac{\sum_{n=1}^{N} r_{nk}(C_i)}{Nc_i}$$

$$P(c_i|\boldsymbol{x}_k) = \frac{P(\boldsymbol{x}_k|c_i) \times P(c_i)}{\sum_j P(\boldsymbol{x}_k|c_j) \times P(c_j)}$$

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

Where N_{tot} is the total number of training items, and r_{kn} is the responsibilities of the members of class c_i in the node k. To predict a class for a new compound q, the following equation is used:

$$P(c_i|\boldsymbol{t}_q) = \sum_{k=1}^{K} P(c_i|\boldsymbol{x}_k) \times r_{kq}$$

To visualize a class landscape, the normalized probability of class c_i is used as a 3rd axis (color code). The population of the nodes is taken into account by the addition of the transparency to the used colors.

GTM-based regression models are relying on the property landscapes, which represent the distribution of a property over the latent space. The definition of property landscape is done by using a list of property values of compounds that correspond to a particular node:

$$p_k = \frac{\sum_{n=1}^{N} p_n \times r_{kn}}{\sum_{n=1}^{N} r_{kn}}$$

Where p_n is the property value for the compound n, and p_k is the mean property value for the node k. The visualization of the property landscape is done by using the p_k values as the 3rd axis and interpreting them as color code. In contrast, transparency is used (same as in the class landscapes) to take into account the nodes population.

The property of a new compound q is predicted similarly to the class prediction:

$$p_q = \sum_{k=1}^{K} r_{kq} \times p_k$$

The density landscape could be viewed as a "subtype" of property landscape, where p_k represents the sum of all compound responsibilities in the node k. This landscape is usually applied for the analysis of the data distribution when there is no particular property

to use (see chapter 6.1.2.2 Cell-based diverse-library selection using GTM) or when the landscape transparency is not easily readable.

3.3.2 Applicability domain of GTM-based QSAR models

While using GTM for modeling, one can use classical AD approaches, but GTM is offering several definitions of AD. Several different GTM-based AD definitions have been reported [6]: likelihood-based, density-based and class-dependent density.

When applying *likelihood-based* AD, a compound is considered to be out of the GTM AD if its position is too far away from the manifold in the initial data space. To apply this AD concept, the LLh cutoff is determined by sorting the training compounds accordingly to their LLh from the higher to lower LLh value. The cutoff is set at n% of compounds (usually 5%) having the smallest LLh; 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 AD allows using only populated zones to make the predictions. This concept is the easiest to visualize and interpret since all the testing compounds that have been projected on the white zones of the map are considered to be out of AD.

The *class-dependent density* AD is similar to the density-based AD. The difference is that the density of the winning class c_{best} in the node is checked, which has the highest conditional node probability $P(x_k|c_{best})$. The ratio for this predominance is a user-defined variable. This AD concept is especially useful for classification models.

3.3.3 Universal GTM

Universal Generative Topographic Maps (uGTM) are GTM-based classification models that have been introduced by Sidorov et al. [10]. In this work, the authors aimed to

cover large chemical space defined by the ChEMBL database of version 20 using a single map. The main difference between uGTM and "local" GTM (GTM explicitly built for a dataset of compounds manifesting a particular activity) is the data used for manifold fitting. Usually, when one is applying GTM, the compounds of the training set share a common activity/property; thus, the obtained map will describe its landscape. The data used for uGTM contains more than 1.2M ligands with known activity for more than 400 biological targets. The descriptors space and the GTM parameters were selected using the Genetic algorithm [67, 70] described in chapter SVM/GTM parameters tuning. The results showed that the uGTM approach could efficiently cover a broad range of chemotypes. The best map selected by GA was cross-validated on 410 ChEMBL targets, showing that approximately 80% of the targets were predicted with the mean Balanced Accuracy of 0.7.

3.4 SVM/GTM parameters tuning

The performance of the machine learning methods is parameter dependent. For instance, the SVM/SVR performance depends on the type of kernel and the regularization coefficient C, while the GTM has four parameters: number of nodes *k*, number of RBFs *m*, regularization coefficient *l*, and RBF's width *w*. Besides these parameters, an "optimal" descriptor space is also needed to be found. To tune all these parameters, the genetic algorithm (GA) has been used in all of the projects. GA is a stochastic approach that allows achieving "the maximal" model performance while trying a variety of combinations of parameters from a pre-defined range.

The algorithm's details have already been described in several publications [67, 70]. Shortly, GA generates a set of chromosomes; each chromosome is presented by a vector of model's parameter values, as well as some meta-parameters like descriptor space. Each attempt (chromosome) is validated using n-fold cross-validation repeated m times, and a fitness score (FSc) is associated with the attempt. The FSc is related to the model's success with a given chromosome; for classification tasks, FSc is related to BA and for regression

tasks to R^2. Higher scored chromosomes will be allowed to generate "children" using cross-overs and mutations, which might result in potentially better FSc. The GA stops in two cases: either no FSc improvement has been observed during the last two generations, or the maximal number of attempts has been achieved.

In the case of SVR- or GTM-based regression model, the FSc is defined by a cross-validated determination coefficient. For each *repetition* of the n-fold cross-validation the R_n^2 is computed. By default, the algorithm is doing 3-fold cross-validation repeated 12 times. Then, the mean value of R_n^2 ($\langle R^2 \rangle$) as well as its standard deviation σ is calculated. The FSc is defined as:

$$FSc = \langle R^2 \rangle - 2 \times \sigma$$

When the GA is run for the optimization of the classification model, the FSc is calculated as follows:

$$FSc = \langle BA \rangle - 2 \times \sigma$$

Where $\langle BA \rangle$ is the mean value of the cross-validated BA of each repetition and σ being its standard deviation.

4 Consensus modeling using universal maps

4.1 Introduction

Virtual Screening (VS) [71] is a technique applied in drug discovery to search libraries of molecules in order to identify the compounds with the property/activity of interest using knowledge retrieved from the existing data. Usually, the so-called VS funnel has several layers of applied methods differentiating them in terms of accuracy/speed ratio. For instance, the methods having low accuracy but high computational speed (like filters or similarity search) will usually be applied in the first place in order to eliminate the compounds that are less likely to be active. On the other hand, methods providing high accuracy with the cost of slower computational speed (like docking) will be applied at "the bottom of the funnel" on a more restricted set of compounds since they are more likely to be active.

GTM has proven to be able to produce target/property-specific models having comparable performance to other machine learning methods like SVM and RF. However, in contrast to these methods, the manifold fitting is an unsupervised process. Therefore, with GTM, one manifold can fit any database containing thousands or even millions of compounds. These compounds may have various activities/properties; hence with only one fitted manifold, one have access to all the landscapes representing the present compounds' activities/properties. This has been applied and tested in the work of Sidorov et al. [10]., where for the first time, the concept of *universal GTM* (uGTM) has been introduced.

Here, the same protocol [10] has been applied to data extraction (ChEMBL 23) and standardization, as well as uGTM generation. A total number of 1.5M compounds with

known activities on 618 targets have been extracted. The subsets of the ligands of 236 targets, including GPCRs, kinases, nuclear receptors etc., have been used for 3-fold cross-validation, and 382 target-specific subsets have been used exclusively as external-validation sets. Directory of Useful Decoys (DUD) [72] has been used as a genuinely external-validation set. The DUD dataset has been standardized in the same way as previously extracted compounds from ChEMBL. The standardization has been followed by removal from the DUD dataset of the compounds that are already present in ChEMBL in order to create orthogonal external data sets. Most targets had a complete overlap of active compounds when they were simultaneously present in ChEMBL and DUD. In these cases, corresponding target-specific sets have been discarded, however, in nine cases the DUD database contained sufficiently numerous original actives, thus leading to 9 target-specific subsets.

Table 4-1:Description of target-specific subsets used for model training (ChEMBL) and VS (DUD).

ChEMBL ID	Target Name	DUD dataset		ChEMBL dataset	
		Active	Inactive	Active	Inactive
1827	Phosphodiesterase 5A	170	25334	691	1515
1952	Thymidylate synthase	63	6113	124	455
251	Adenosine A2a receptor	79	28001	1303	3618
260	MAP kinase p38 alpha	100	32925	1453	2567
279	Vascular endothelial growth factor receptor 2	94	22595	2047	4663
301	Cyclin-dependent kinase 2	189	25675	638	2305
4282	Serine/threonine-protein kinase AKT	52	14228	725	2619
4338	Purine nucleoside phosphorylase	102	6334	100	111
4439	TGF-beta receptor type I	82	8013	282	385

Eight uGTMs have been selected being ranked as “best maps” by genetic algorithm, with the scoring function being the average BA for all target-specific landscapes on 3-fold cross-validation. Each uGTM is based on different ISIDA descriptor space, encoding distinct structural features. Although the average BA of eight maps is roughly equivalent, the maps were showing different BAs on target-specific subsets. For instance, an individual map was showing high BA value for a given target-specific subset while having a relatively low value on another target-specific subset; moreover, another map could show the opposite behavior. This induced an in-depth analysis of 8 selected uGTMs.

4.2 Performance evaluation of universal maps

The uGTMs performance was evaluated using three scores: i) BA in 3-fold CV (using ChEMBL compounds) and in VS (using DUD); ii) Receiver Operating Characteristic Area Under Curve (ROC AUC) in VS; iii) Enrichment Factor (EF) in VS. BA has been mainly used during cross-validation. BA serves to assess the ability of landscapes to predict the correct activity class of candidates not used for landscape construction, i.e., both in “internal” cross-validation and “external” VS. Note that reported BA scores for individual maps – both in the CV and in VS applications – are always calculated on the entire target-specific sets. It includes *all* DUD compounds, even those projected onto empty map zones; hence these molecules are out of AD, and they are considered, by default, inactive.

However, ROC AUC is a more natural VS evaluation criterion than BA, since, in VS, the critical element is the relative ranking of candidates – a significant prioritization of the active compounds with respect to the inactive. The ranking was performed according to the GTM landscape-predicted probability of each compound to be active. The compounds falling outside the applicability domain were assigned zero probability of activity; thus, they were placed at the bottom of the ranking list. EF for the top 100 ranked molecules was calculated according to the equation below:

$$EF_{100} = \frac{Actives_{100}/100}{Actives_{total}/N_{total}}$$

Where $Actives_{100}$ is the number of true positives in the top 100 compounds, $Actives_{total}$ is the total number of active compounds in the dataset, N_{total} is the total number of compounds in the target-specific dataset.

Detailed description of the results is given in the article published in J. Chemical Informatics and Modeling, see below.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Cite This: J. Chem. Inf. Model. 2019, 59, 564−572
Article
pubs.acs.org/jcim

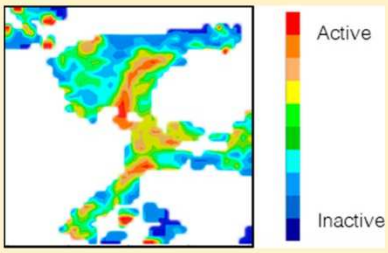
Virtual Screening with Generative Topographic Maps: How Many Maps Are Required?

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

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

Supporting Information

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



INTRODUCTION

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

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

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

Received: September 21, 2018
Published: December 19, 2018

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

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

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

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

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

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

METHODS

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

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

		DUD data set		ChEMBL data set	
ChEMBL ID	target name	active	inactive	active	inactive
1827	phosphodiesterase 5A	170	25 334	691	1515
1952	thymidylate synthase	63	6 113	124	455
251	adenosine A2a receptor	79	28 001	1303	3618
260	MAP kinase p38 alpha	100	32 925	1453	2567
279	vascular endothelial growth factor receptor 2	94	22 595	2047	4663
301	cyclin-dependent kinase 2	189	25 675	638	2305
4282	serine/threonine-protein kinase AKT	52	14 228	725	2619
4338	purine nucleoside phosphorylase	102	6 334	100	111
4439	TGF-beta receptor type I	82	8 013	282	385

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

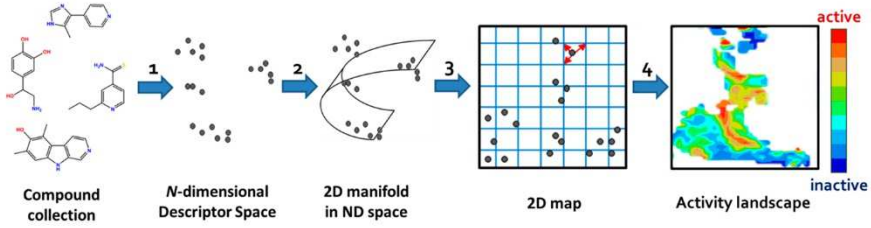


Figure 1. A frame set of compounds is represented in the *N*-dimensional descriptor space. A flexible 2D manifold, which is a square grid of nodes, is injected into that space and is fitted to the data. The molecules are nonlinearly projected onto it, and when the manifold is unbent, a 2D map is obtained. Each node can be colored according to the activities of molecules residing there, producing "activity landscapes", where red zones are populated only by active molecules and blue by inactive; all colors in between correspond to the regions occupied by compounds of both classes in different proportions. White zones are empty.

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

Workflow. The following workflow was applied:

(1) Standardization of ChEMBL and DUD data sets followed by descriptor generation;
(2) Coloring the manifolds of universal maps by each of nine target-specific class landscapes using ChEMBL subsets;
(3) 3-fold cross-validation of predictive landscapes within the ChEMBL data sets;
(4) Application of these landscapes for the VS of the corresponding DUD subsets

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

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

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

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

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

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

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

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

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

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

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

map	abbreviation	definition	descriptor space dimensionality
1	IA-FF-FC-AP-2-3	sequences of atoms with a length of 2−3 atoms labeled by force field types and formal charge status, using all paths.	5161
2	IIRAB-FF-1-2	atom-centered fragments of restricted atom and bonds of 1−2 atoms labeled by force field types	3172
3	IAB-PH-FC-AP-2-4	sequences of atoms and bonds of a length 2−4 atoms labeled by pharmacophoric atom types and formal charges using all paths	4245
4	IA-2-7	sequences of 2−7 atoms.	6520
5	IAB-FC-AP-FC-2-4	sequences of atoms and bonds of 2−4 atoms labeled by formal charge, using all paths	3437
6	IA-FF-P-2-6	sequences of atom pairs with a length of 2−6 intercalated bonds, labeled by Force Field type	2901
7	III-PH-3-6	atom triplets labeled by pharmacophoric atom types with topological distance from 3 to 6 bonds	4846
8	III-FF-3-4	atom triplets labeled by force field types, with topological distance from 3 to 4 bonds	8953

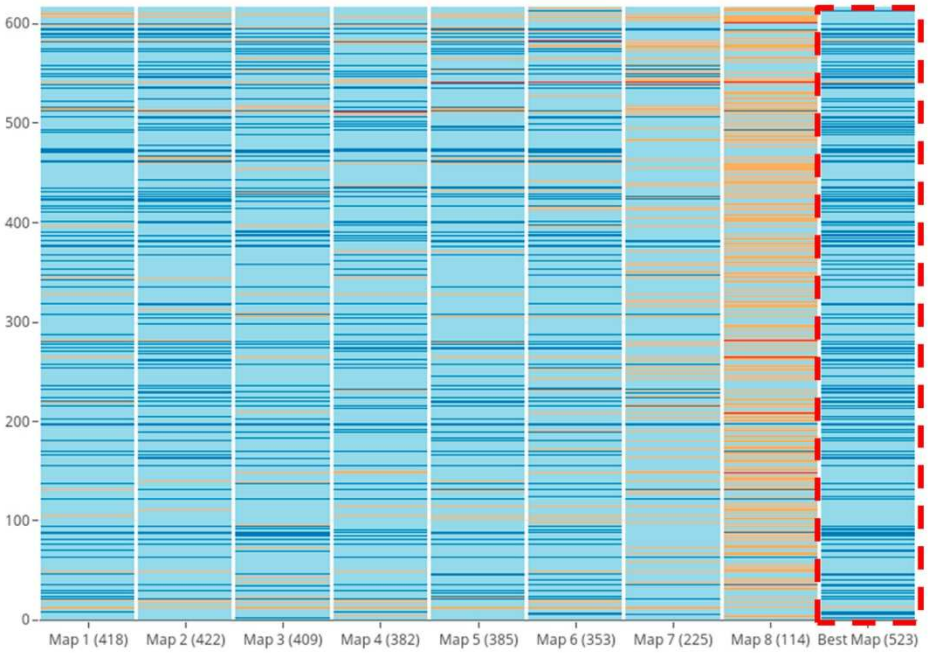


Figure 2. Heatmap showing the performance of universal maps on 618 selected series. Color-codes: dark blue, BA > 0.85; light blue, 0.65 < BA ≤ 0.85; orange, 0.5 < BA ≤ 0.65; and red, BA ≤ 0.5. Between parentheses is shown the number of target-specific classification problems for which a map scores BA > 0.75.

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

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

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

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

$$EF_{100} = \frac{Actives_{100}/100}{Actives_{total}/N_{total}}$$

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

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

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

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

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

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

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

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

RESULTS

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

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

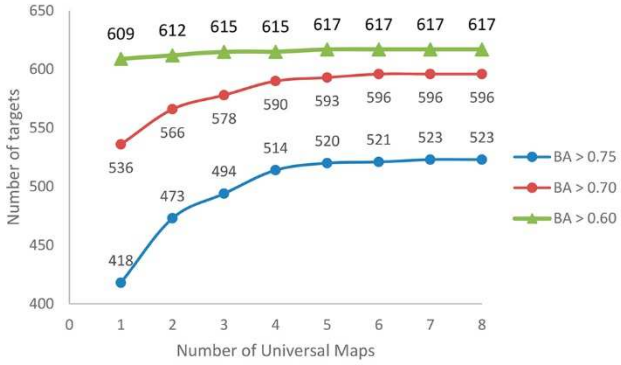


Figure 3. Cumulated performance of universal maps: number of predicted target-specific series vs number of used maps.

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

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

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

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

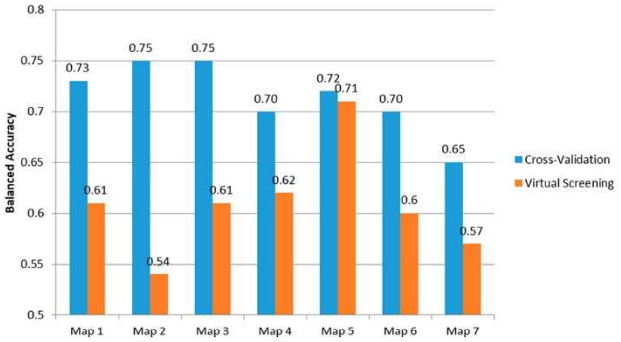


Figure 4. BA values obtained in CV and VS of the CHEMBL260 data set.

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

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

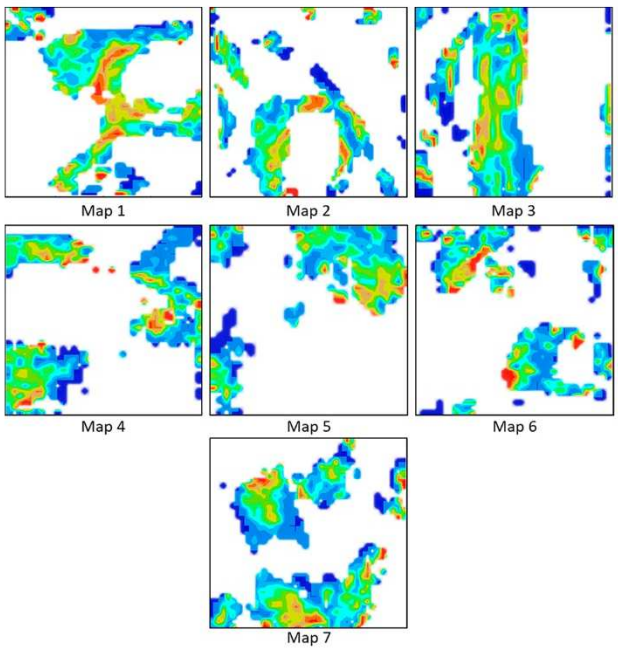


Figure 5. Activity class landscapes of the CHEMBL260 data set in seven universal maps numbered according to Table 2. Because the seven latent spaces are independent projections of distinct initial chemical spaces, these activity class landscapes cannot be "overlaid" to produce a "consensus" landscape.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

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

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

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

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

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

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

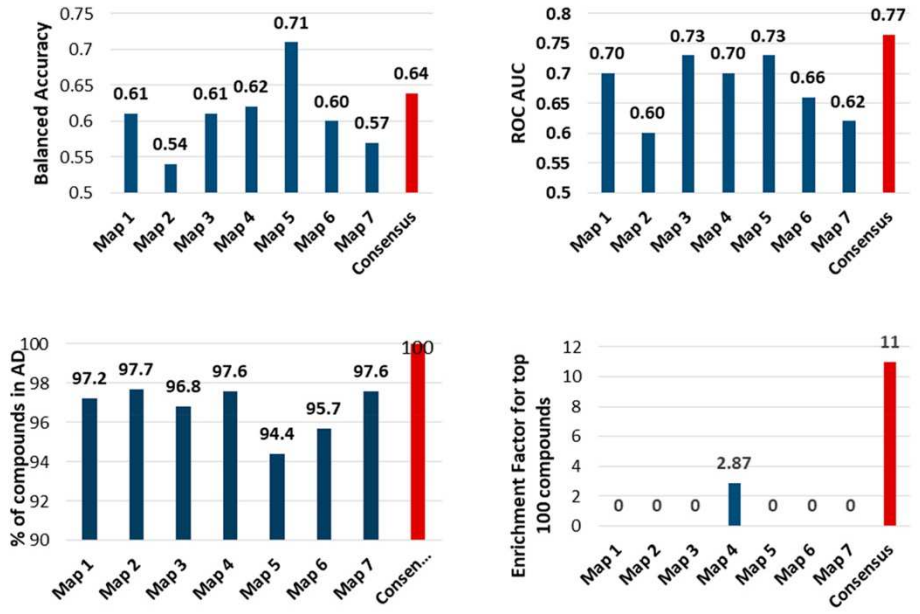


Figure 6. Performance of VS on DUD with the models developed for the CHEMBL260 data set assessed on the basis of BA (top left), ROC AUC (top right), data coverage (bottom left), and EF calculated for top 100 compounds (bottom right)

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

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

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

In light of the numerous factors impacting the predictive power of GTM landscapes, it may be very difficult to highlight a detailed explanation for the specific prediction successes and failures observed here. In the following, the predictive behavior for target CHEMBL4338 (purine nucleoside phosphorylase, the one exception for which no conclusive synergy effect of the individual maps was observed) has been analyzed in more detail. The herein used DUD set features 102 purine-like actives and 6334 decoy compounds.

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

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

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

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

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

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

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

Furthermore, Table 4 outlines a clear negative correlation between the number of wrongly predicted "active" phenyl sulfonamides and the ROC AUC score in VS. This is, of course, not only due to the cited compounds being misplaced on the ROC curves but illustrates the above-discussed effect of the different "perceptions" of neighborhood provided by each map. As mentioned, phenyl sulfonamides appear as clearly distinct from the CHEMBL purine-like reference compounds, actives, or inactives alike. From the medicinal chemist's point of view, these are expected to fall in blank zones of a landscape colored by the completely unrelated purines, hypoxantines, pyrrolopyrimidnes, etc. Maps 1, 5, and 6 fully comply with this point of view. Maps 2 and 4 demonstrate slightly "fuzzier" definitions of molecular similarity: a few phenyl sulfonamides are now being placed within the CHEMBL reference compounds, whereas map 7 based on scaffold-hop-supporting pharmacophore triplet counts actually assumes that most of the phenyl sulfonamides reside in the purine nucleoside phosphorylase-relevant chemical space zone. An overwhelming majority of these in-zone residing decoys are correctly recognized as inactives; however, even a "minority" of false positives may represent a very large number compared to the much rarer actives in the highly imbalanced DUD set. This is the reason for the predictive failure of map 7, which could not be understood in terms of its cross-validation results. When cross-validating, the map is exclusively confronted with purine nucleoside phosphorylase-relevant chemicals, where there are no "exotic" chemotypes to be spuriously co-opted into relevant chemical space by a—for this predictive challenge—"too permissive" perception of molecular similarity.

CONCLUSION

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

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

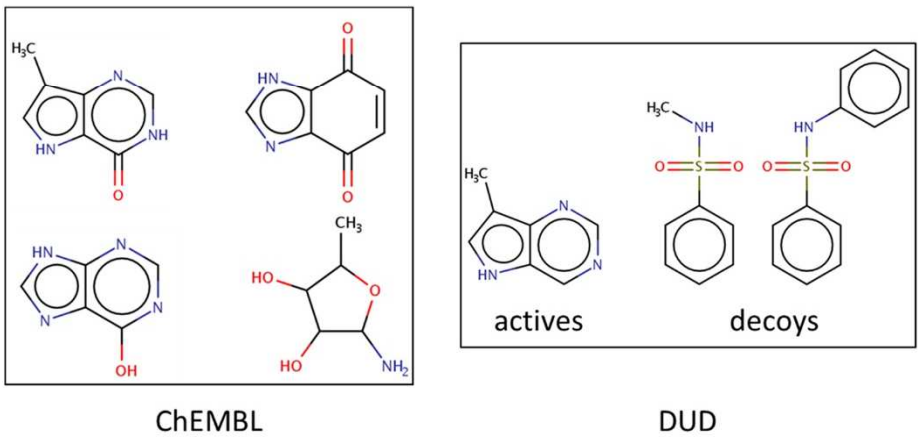


Figure 7. Representative substructures of compound subsets of the purine nucleoside phosphorylase receptor in the CHEMBL4338 data set and DUD.

DOI: 10.1021/acs.jcim.8b00650
J. Chem. Inf. Model. 2019, 59, 564−572

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

■ ASSOCIATED CONTENT

Supporting Information

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

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

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

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@unistra.fr.

ORCID

Dragos Horvath: 0000-0003-0173-5714

Jürgen Bajorath: 0000-0002-0557-5714

Alexandre Varnek: 0000-0003-1886-925X

Notes

The authors declare no competing financial interest.

ISIDA GTM software is developed by the Laboratoire de Chemoinformatique Strasbourg and can be obtained upon request (visit http://infochim.u-strasbg.fr/spip.php?rubrique41).

■ ACKNOWLEDGMENTS

I.C. thanks the Région Grand Est for a Ph.D. fellowship.

■ REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3 Conclusions

In this work, the predictive performance of eight newly constructed uGTM models in a "strict" 3-fold CV and VS of nine target-specific subsets of compounds extracted from DUD has been assessed. It has been shown that these maps can provide a relatively good separation (BA_{CV} >0.6) of active and inactive for the majority of 618 ChEMBL target-specific subsets, irrespective of whether these subsets have been used in model training or not. It has been found out that any individual map could not achieve consistently accurate predictions for each target-specific subset. However, it has been proven that these maps, which were each built on a different descriptor space, are highly complementary – the target-specific series of compounds that are being predicted poorly by one uGTM will be much better predicted by another. For 617 out of 618 activity classes, at least one uGTM provides a highly discriminatory activity landscape.

It was observed that there is no correlation between performance in the CV and external predictive power of individual activity landscapes. A solution has been found – a consensus approach. The most important advantages of this approach are 1) 100% data coverage in most of the cases; 2) a significant increase in EF for the 100 top-ranked compounds; 3) high performance of the consensus model compared to individual models based on ROC AUC. Last but not least, seven uGTMs have been proven to be sufficient to provide complementary views of biologically relevant chemical space that resulted in the enhancement of the performance in VS.

4.4 Supporting information

Supporting information includes a list of targets used for building of the universal maps (Table 4-2) and activity landscapes for 9 selected ChEMBL targets (Figures 4.1 -4.9).

Table 4-2: 618 ChEMBL (version 23) targets used for universal maps training and validation.

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

CHEMBL210	CHEMBL2326	CHEMBL2208	CHEMBL256	CHEMBL5491
CHEMBL2107	CHEMBL233	CHEMBL221	CHEMBL2563	CHEMBL5493
CHEMBL211	CHEMBL2334	CHEMBL2216739	CHEMBL2568	CHEMBL6101
CHEMBL2219	CHEMBL2337	CHEMBL2123	CHEMBL258	CHEMBL6115
CHEMBL222	CHEMBL2343	CHEMBL213	CHEMBL2581	CHEMBL6120
CHEMBL2231	CHEMBL2345	CHEMBL2146302	CHEMBL259	CHEMBL6136
CHEMBL2147	CHEMBL2349	CHEMBL248	CHEMBL2593	CHEMBL5818
CHEMBL2148	CHEMBL235	CHEMBL2487	CHEMBL2595	CHEMBL5819
CHEMBL215	CHEMBL236	CHEMBL2492	CHEMBL2598	CHEMBL5847
CHEMBL216	CHEMBL237	CHEMBL250	CHEMBL2599	CHEMBL5855
CHEMBL2163176	CHEMBL2373	CHEMBL2508	CHEMBL260	CHEMBL4900
CHEMBL2169736	CHEMBL238	CHEMBL251	CHEMBL261	CHEMBL4973
CHEMBL217	CHEMBL2386	CHEMBL2514	CHEMBL2611	CHEMBL4977
CHEMBL2179	CHEMBL239	CHEMBL2525	CHEMBL2617	CHEMBL5024
CHEMBL218	CHEMBL2390810	CHEMBL2527	CHEMBL262	CHEMBL5027
CHEMBL2185	CHEMBL240	CHEMBL253	CHEMBL2635	CHEMBL5028
CHEMBL2189110	CHEMBL241	CHEMBL2534	CHEMBL2637	CHEMBL5038
CHEMBL2424	CHEMBL2413	CHEMBL2535	CHEMBL2652	CHEMBL5073
CHEMBL2426	CHEMBL2414	CHEMBL2543	CHEMBL2664	CHEMBL5703
CHEMBL2431	CHEMBL242	CHEMBL255	CHEMBL267	CHEMBL5719
CHEMBL2434	CHEMBL268	CHEMBL2820	CHEMBL2996	CHEMBL5742
CHEMBL2439	CHEMBL2689	CHEMBL2828	CHEMBL3004	CHEMBL5747
CHEMBL2468	CHEMBL2693	CHEMBL283	CHEMBL3009	CHEMBL5203
CHEMBL2474	CHEMBL2695	CHEMBL2850	CHEMBL301	CHEMBL5247
CHEMBL3553	CHEMBL2716	CHEMBL288	CHEMBL3012	CHEMBL5251
CHEMBL3559	CHEMBL2717	CHEMBL2888	CHEMBL3023	CHEMBL5857
CHEMBL3568	CHEMBL2730	CHEMBL2889	CHEMBL3024	CHEMBL5879
CHEMBL2731	CHEMBL289	CHEMBL3025	CHEMBL3231	CHEMBL5896
CHEMBL2736	CHEMBL2896	CHEMBL3032	CHEMBL3234	CHEMBL5903
CHEMBL2742	CHEMBL290	CHEMBL3045	CHEMBL3238	CHEMBL5936
CHEMBL275	CHEMBL2903	CHEMBL3055	CHEMBL3243	CHEMBL5938
CHEMBL2778	CHEMBL2916	CHEMBL3060	CHEMBL325	CHEMBL5971
CHEMBL2781	CHEMBL2938	CHEMBL3070	CHEMBL3250	CHEMBL5979
CHEMBL2782	CHEMBL2939	CHEMBL308	CHEMBL3267	CHEMBL5366
CHEMBL2789	CHEMBL2955	CHEMBL3094	CHEMBL3268	CHEMBL5378
CHEMBL279	CHEMBL2959	CHEMBL3106	CHEMBL3272	CHEMBL5393
CHEMBL2793	CHEMBL2964	CHEMBL3116	CHEMBL3286	CHEMBL5407
CHEMBL2801	CHEMBL2971	CHEMBL3130	CHEMBL3308	CHEMBL5408
CHEMBL2803	CHEMBL2973	CHEMBL3142	CHEMBL331	CHEMBL6009
CHEMBL2808	CHEMBL298	CHEMBL3145	CHEMBL3310	CHEMBL6014
CHEMBL2815	CHEMBL299	CHEMBL3180	CHEMBL332	CHEMBL6030
CHEMBL3181	CHEMBL333	CHEMBL3522	CHEMBL3710	CHEMBL6032
CHEMBL3192	CHEMBL3338	CHEMBL3524	CHEMBL3714130	CHEMBL5518
CHEMBL3201	CHEMBL335	CHEMBL3529	CHEMBL3717	CHEMBL5522

CHEMBL3202	CHEMBL3351	CHEMBL3535	CHEMBL3721	CHEMBL5524
CHEMBL321	CHEMBL3356	CHEMBL3864	CHEMBL3729	CHEMBL5543
CHEMBL3227	CHEMBL3357	CHEMBL3869	CHEMBL3746	CHEMBL5545
CHEMBL3230	CHEMBL3359	CHEMBL3880	CHEMBL3759	CHEMBL5568
CHEMBL3385	CHEMBL3589	CHEMBL3764	CHEMBL3886	CHEMBL6003
CHEMBL3397	CHEMBL3590	CHEMBL3772	CHEMBL3890	CHEMBL6007
CHEMBL3399910	CHEMBL3616	CHEMBL3776	CHEMBL3891	CHEMBL6154
CHEMBL340	CHEMBL3622	CHEMBL3778	CHEMBL3892	CHEMBL4895
CHEMBL3401	CHEMBL3629	CHEMBL3785	CHEMBL3898	CHEMBL4896
CHEMBL3426	CHEMBL3636	CHEMBL3788	CHEMBL3902	CHEMBL4897
CHEMBL3437	CHEMBL3650	CHEMBL3795	CHEMBL3905	CHEMBL4898
CHEMBL3438	CHEMBL3663	CHEMBL3807	CHEMBL3906	CHEMBL4899
CHEMBL3468	CHEMBL3683	CHEMBL3816	CHEMBL3911	CHEMBL4444
CHEMBL3474	CHEMBL3687	CHEMBL3819	CHEMBL3913	CHEMBL4461
CHEMBL3475	CHEMBL3691	CHEMBL3820	CHEMBL3920	CHEMBL4462
CHEMBL3476	CHEMBL3961	CHEMBL3829	CHEMBL3922	CHEMBL4465
CHEMBL3510	CHEMBL3965	CHEMBL3831	CHEMBL3935	CHEMBL4478
CHEMBL3514	CHEMBL3969	CHEMBL3835	CHEMBL3959	CHEMBL4481
CHEMBL3836	CHEMBL3972	CHEMBL4051	CHEMBL4203	CHEMBL4482
CHEMBL3837	CHEMBL3973	CHEMBL4068	CHEMBL4204	CHEMBL4501
CHEMBL3861	CHEMBL3974	CHEMBL4071	CHEMBL4223	CHEMBL4506
CHEMBL3863	CHEMBL3975	CHEMBL4072	CHEMBL4224	CHEMBL4801
CHEMBL3572	CHEMBL3976	CHEMBL4073	CHEMBL4225	CHEMBL4803
CHEMBL3582	CHEMBL3979	CHEMBL4079	CHEMBL4227	CHEMBL4804
CHEMBL3587	CHEMBL3982	CHEMBL4080	CHEMBL4234	CHEMBL4816
CHEMBL3983	CHEMBL4081	CHEMBL4237	CHEMBL4422	CHEMBL4581
CHEMBL3991	CHEMBL4093	CHEMBL4247	CHEMBL4426	CHEMBL4599
CHEMBL4005	CHEMBL4101	CHEMBL4261	CHEMBL4427	CHEMBL4600
CHEMBL4015	CHEMBL4123	CHEMBL4270	CHEMBL4439	CHEMBL5261
CHEMBL4016	CHEMBL4128	CHEMBL4273	CHEMBL4441	CHEMBL5282
CHEMBL4018	CHEMBL4142	CHEMBL4282	CHEMBL4714	CHEMBL5285
CHEMBL4026	CHEMBL4145	CHEMBL4296	CHEMBL4718	CHEMBL5314
CHEMBL4029	CHEMBL4147	CHEMBL4302	CHEMBL4722	CHEMBL5330
CHEMBL4036	CHEMBL4158	CHEMBL4303	CHEMBL4761	CHEMBL5331
CHEMBL4040	CHEMBL4176	CHEMBL4306	CHEMBL4766	CHEMBL6164
CHEMBL4045	CHEMBL4179	CHEMBL4315	CHEMBL4608	CHEMBL6166
CHEMBL4374	CHEMBL4191	CHEMBL4338	CHEMBL4617	CHEMBL6175
CHEMBL4375	CHEMBL4198	CHEMBL4361	CHEMBL4618	CHEMBL4698
CHEMBL4376	CHEMBL4202	CHEMBL4367	CHEMBL4625	CHEMBL4699
CHEMBL4393	CHEMBL4508	CHEMBL4662	CHEMBL4630	CHEMBL4852
CHEMBL4394	CHEMBL4516	CHEMBL4674	CHEMBL4576	CHEMBL4829
CHEMBL4398	CHEMBL4523	CHEMBL4681	CHEMBL4578	CHEMBL4835
CHEMBL4408	CHEMBL4525	CHEMBL4683	CHEMBL4708	CHEMBL4601
CHEMBL4822	CHEMBL4575	CHEMBL4685		

Activity landscapes for nine studied subsets of ChEMBL targets selected from the DUD database are presented below. Red zones are exclusively populated by active molecules, blue populated by inactive molecules, whereas yellow and green colors characterize zones populated by both active and inactive compounds.

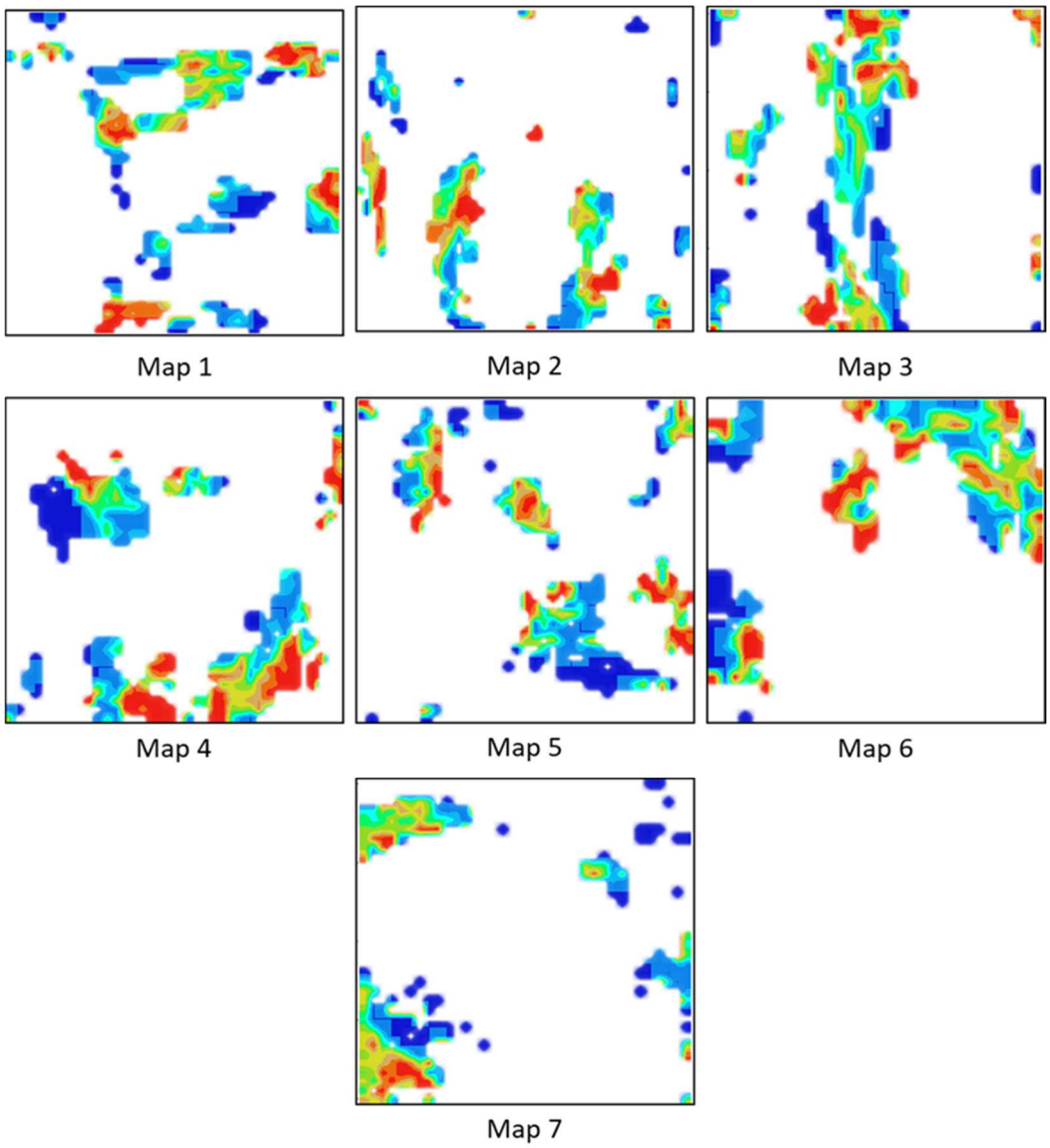


Figure 4.1: Activity landscapes for ChEMBL1827.

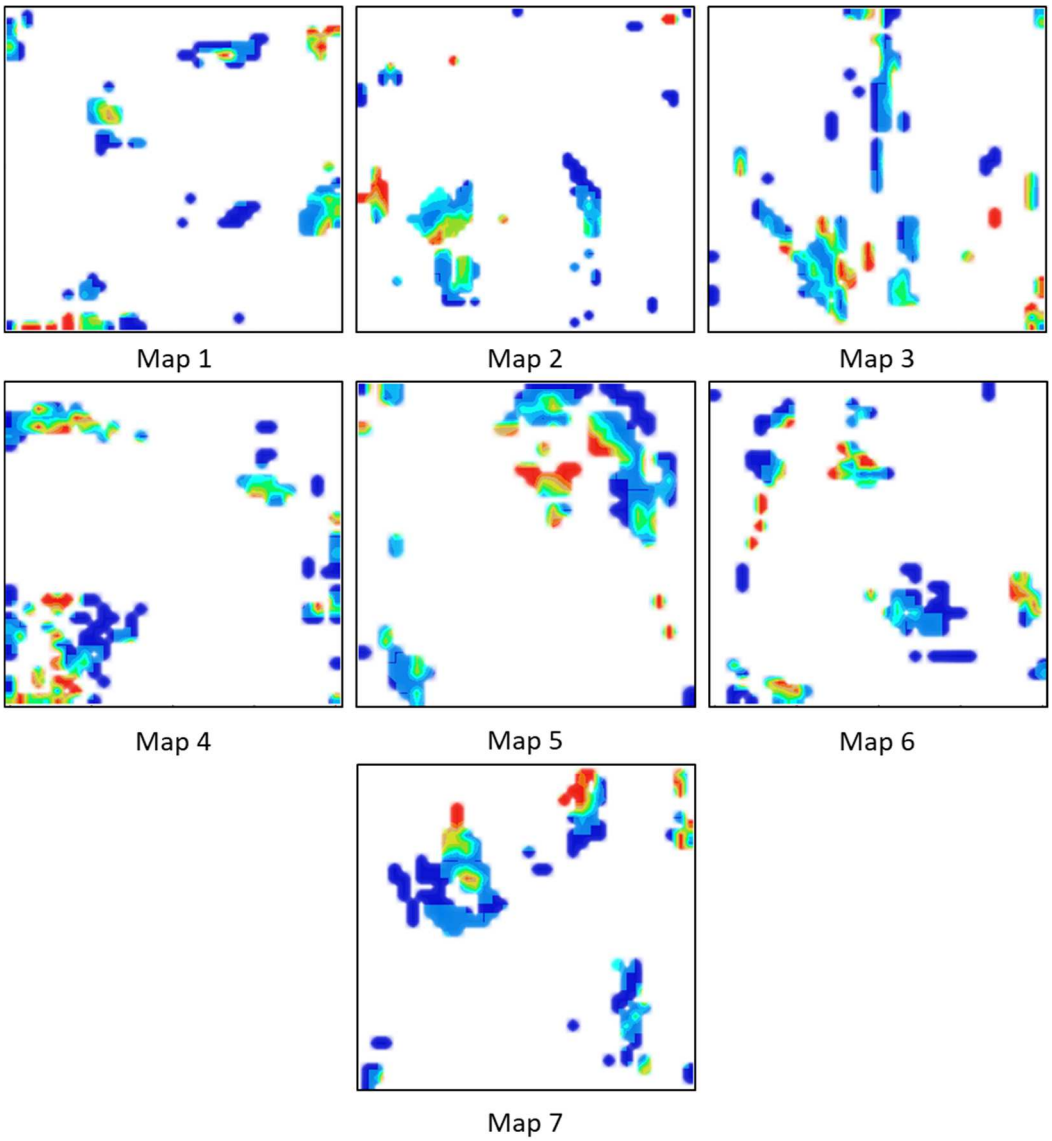


Figure 4.2: Activity landscapes for ChEMBL1952.

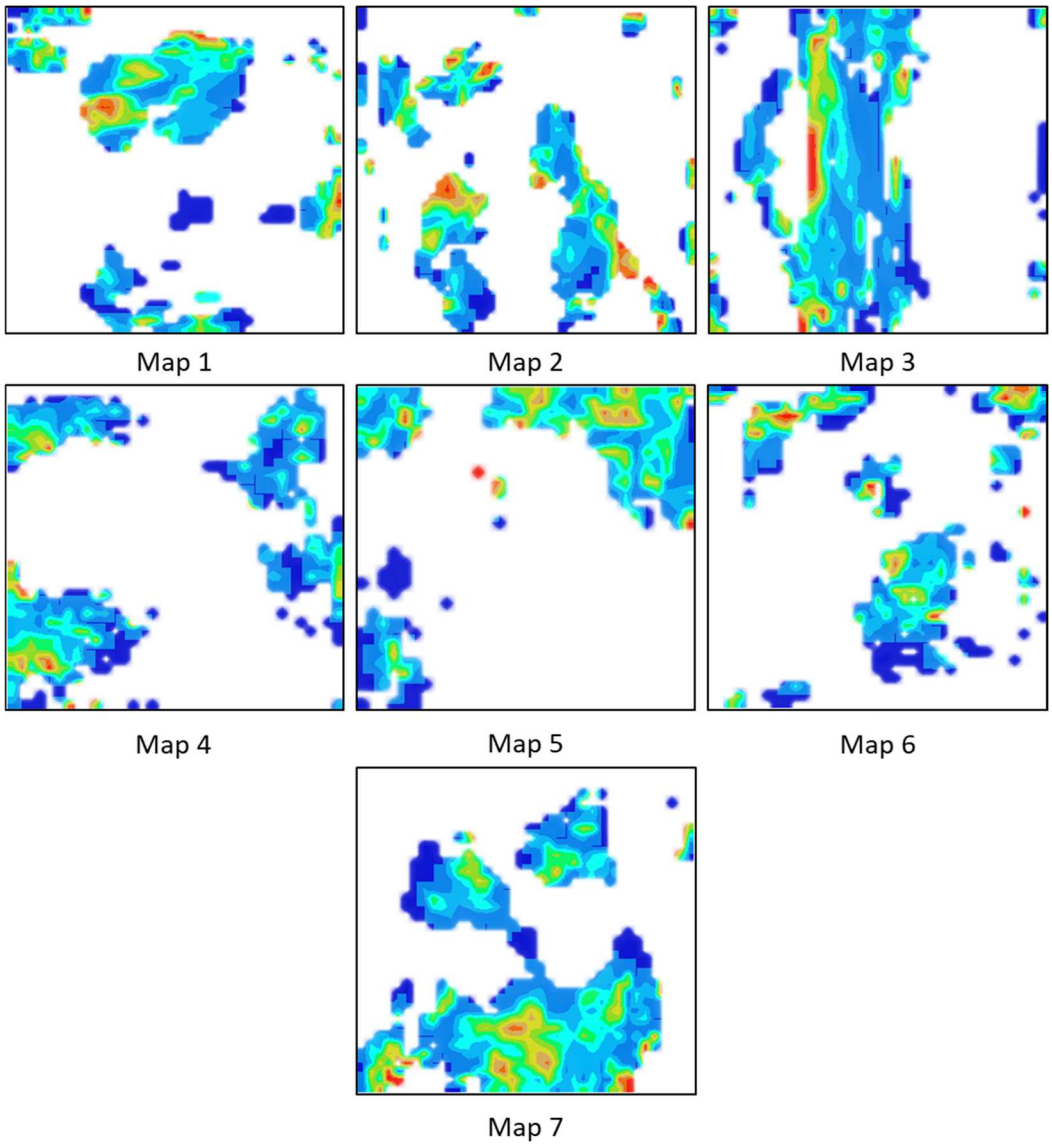


Figure 4.3: Activity landscapes for ChEMBL251.

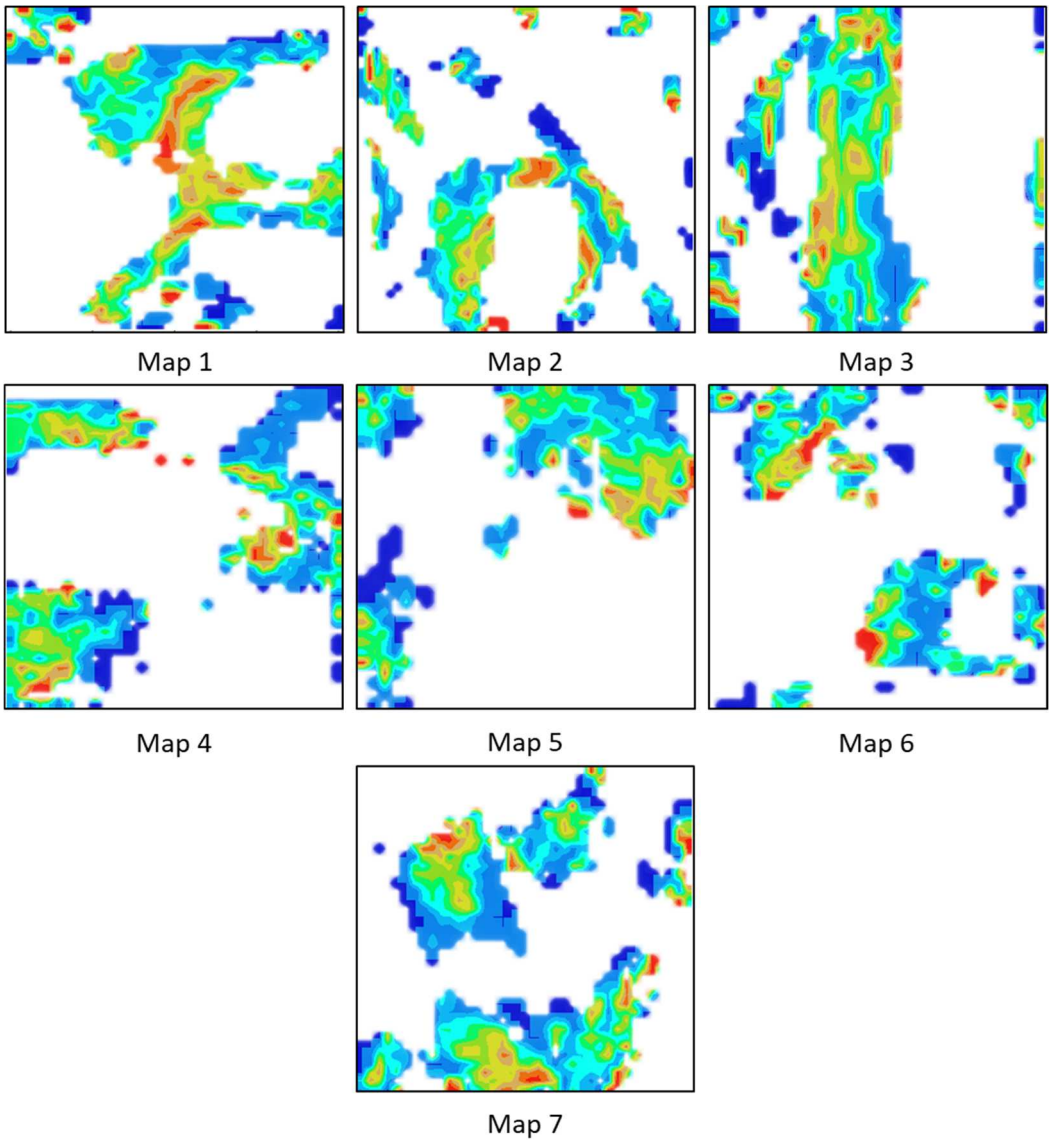


Figure 4.4: Activity landscapes for ChEMBL260.

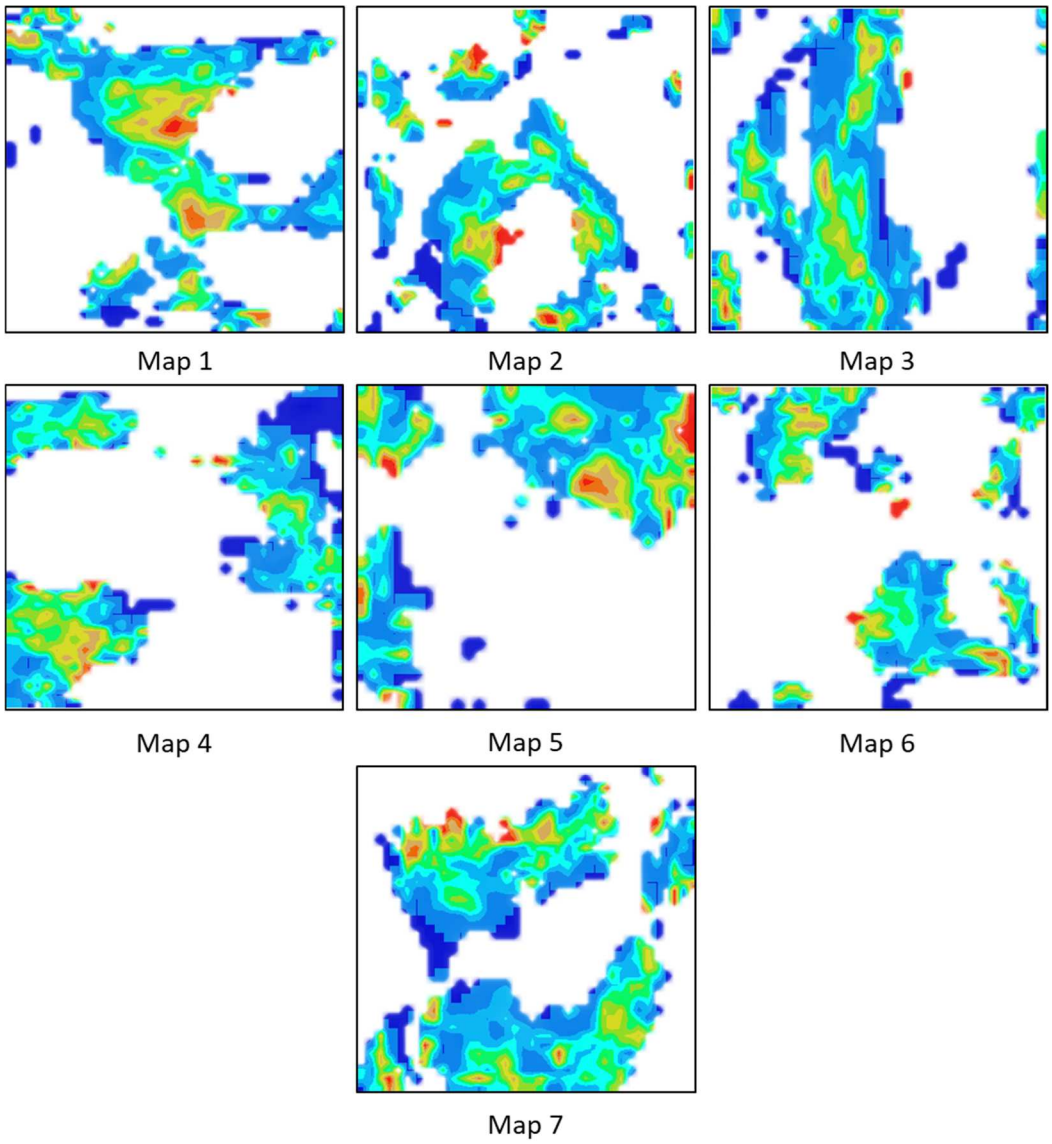


Figure 4.5: Activity landscapes for ChEMBL279.

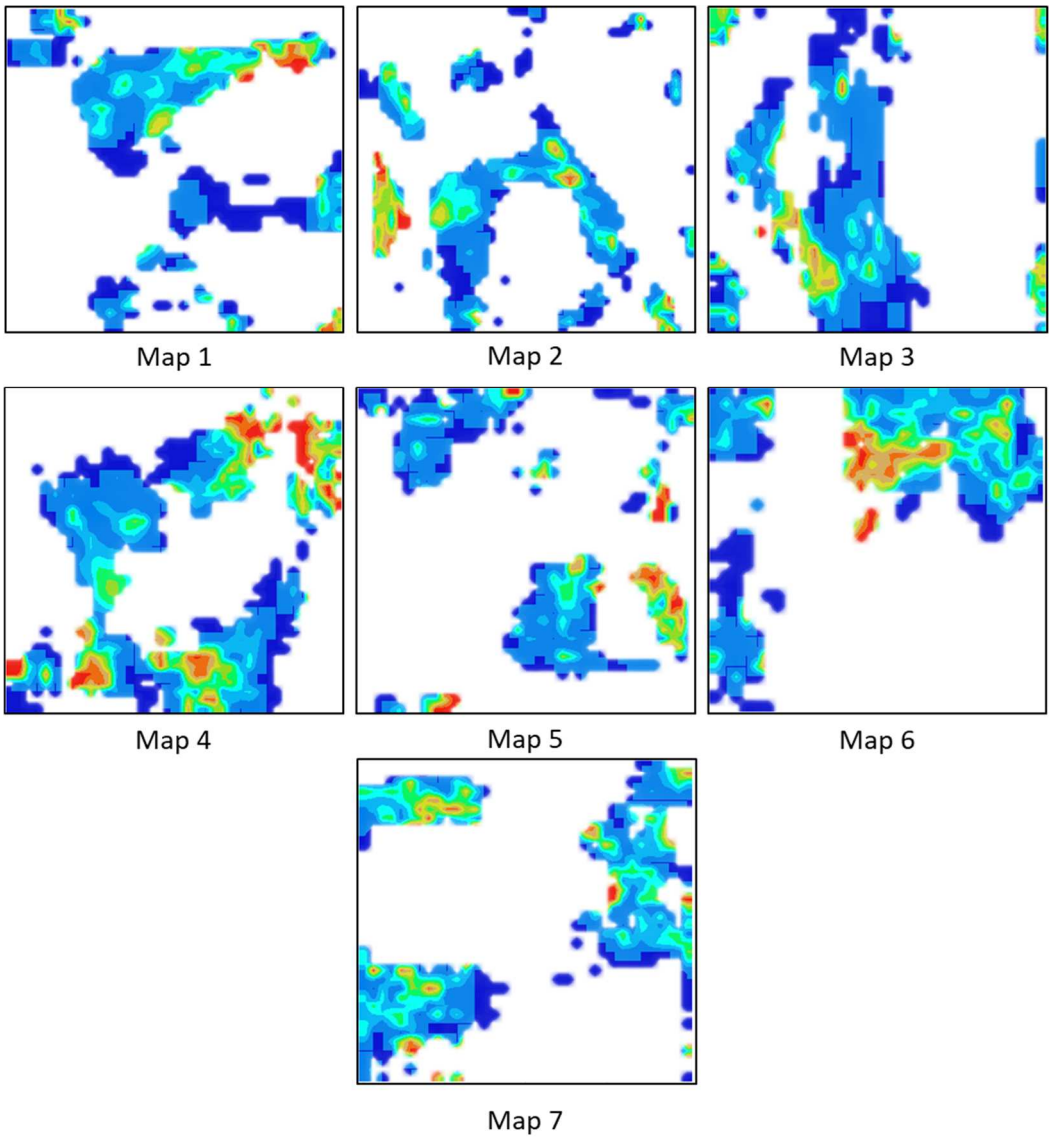


Figure 4.6: Activity landscape for ChEMBL301.

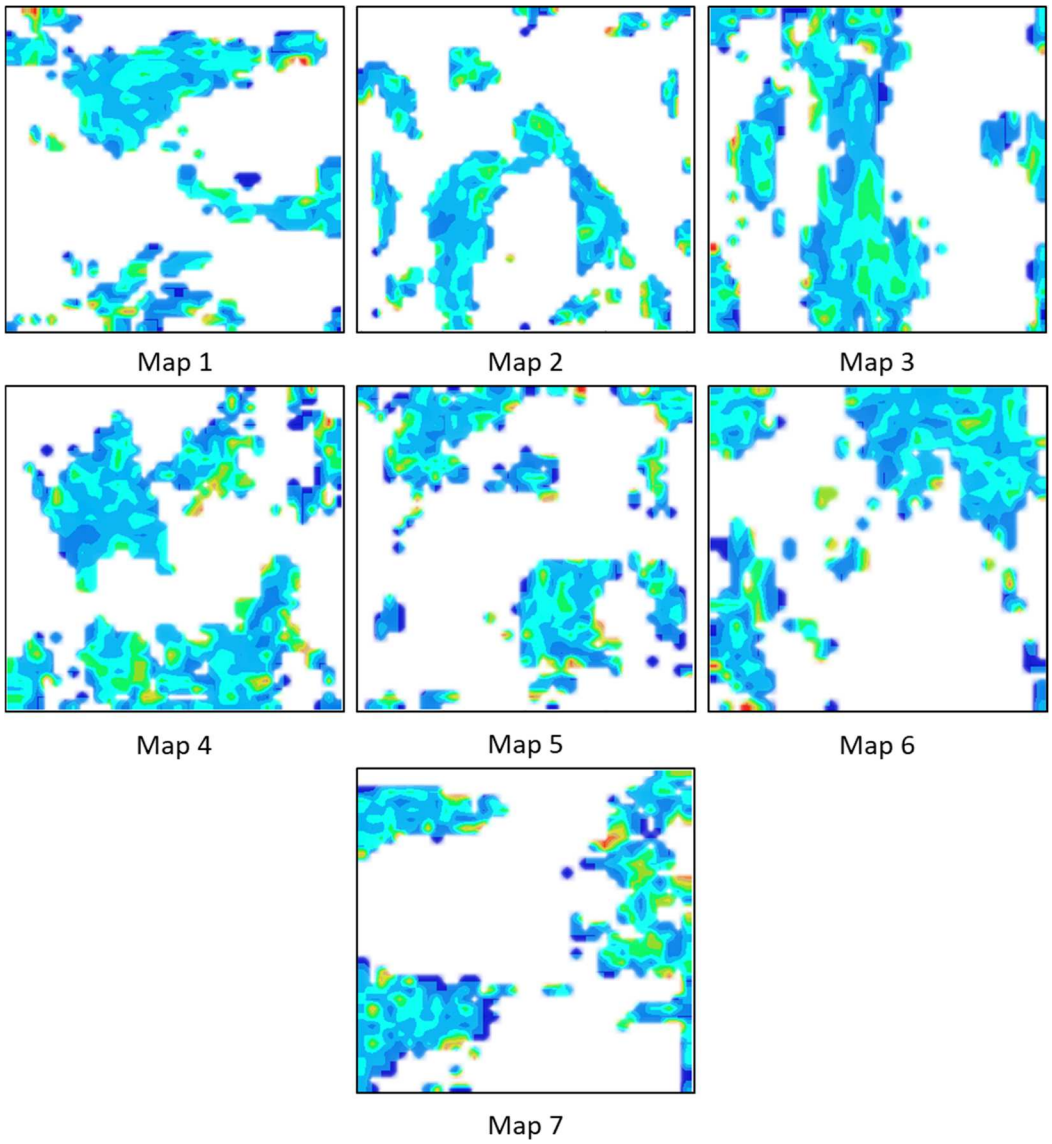


Figure 4.7: Activity landscapes for ChEMBL4282.

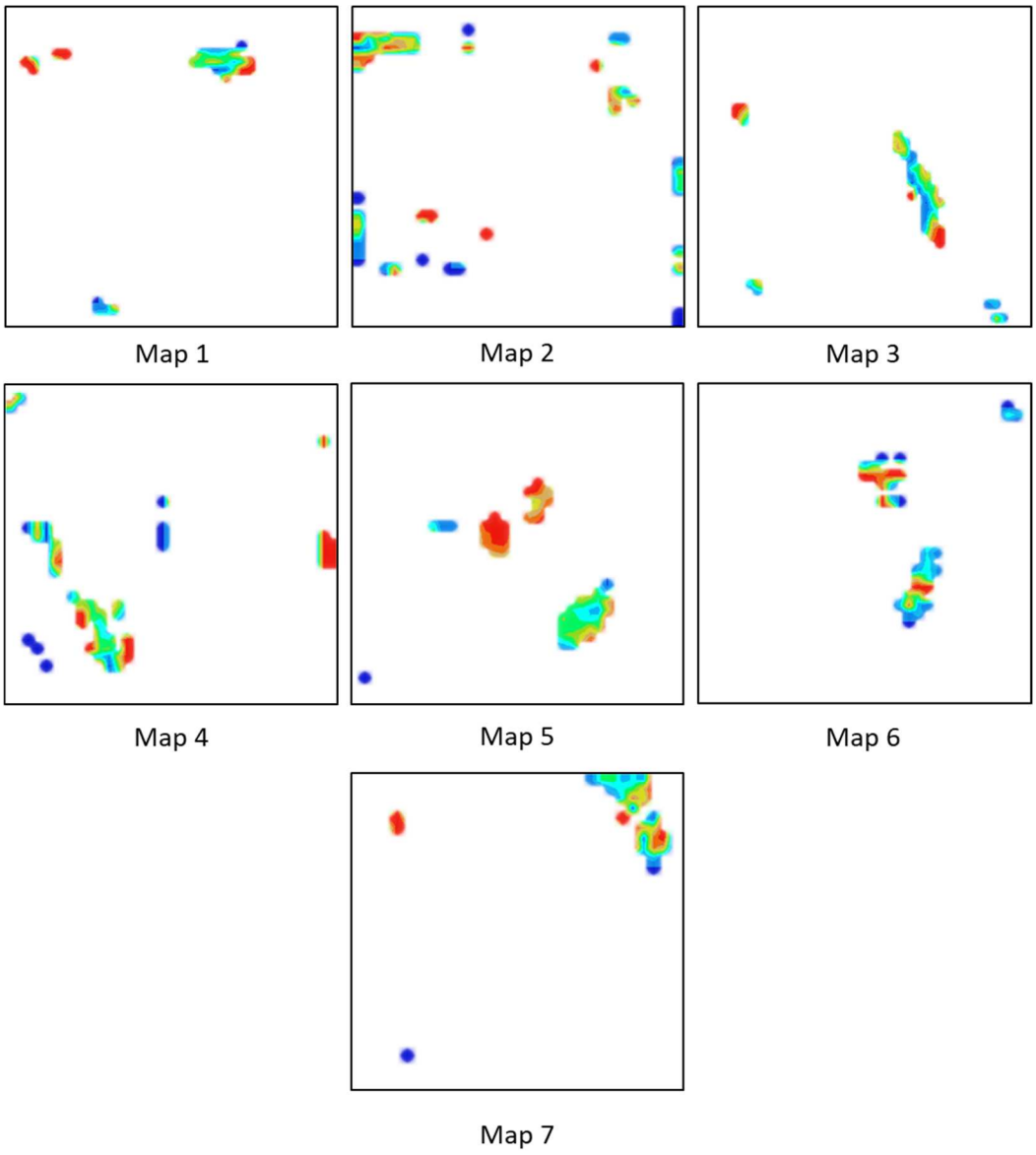


Figure 4.8: Activity landscapes for ChEMBL4338.

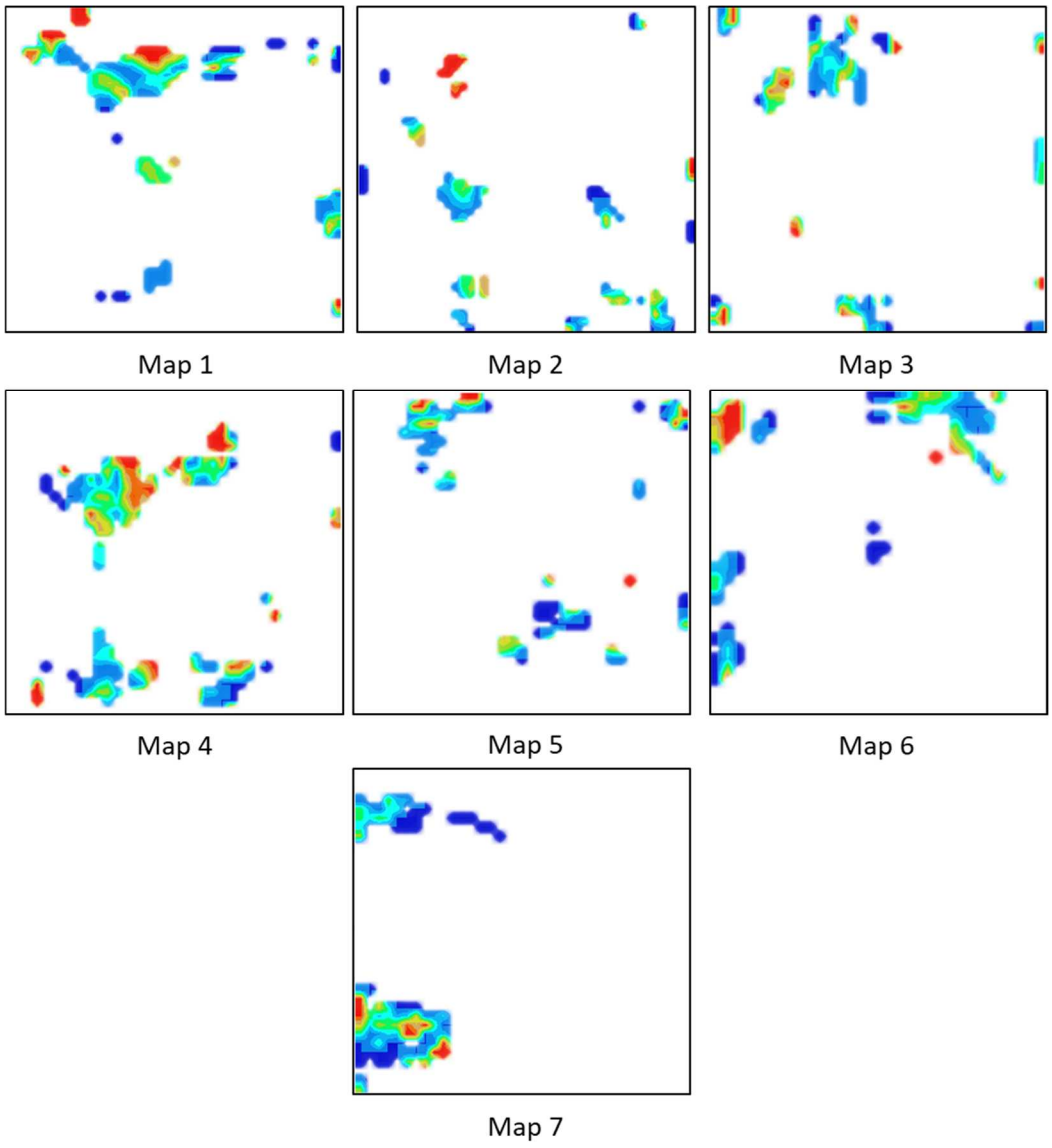


Figure 4.9: Activity landscapes for ChEMBL4439.

5 *In silico* mining for new Bromodomain inhibitors

5.1 Introduction

The goal of this project was to carry out a virtual screening (VS) of a dataset of 2M compounds provided by Enamine company [48], in order to find new inhibitors of Bromodomain 4 (BRD4) [47]. The provided dataset of 2M compounds corresponds to the compounds that are physically available in stocks, or that could be easily synthesized. It has been agreed to provide our Enamine collaborators with a dataset of 3000 compounds that would be tested. To get this dataset of 3000 compounds, a VS screening protocol has been developed, which included a consensus application of GTM and SVM classification models as well as ligand-based pharmacophore models.

5.2 Bromodomain 4

5.2.1 Biological role

Lysine acetylation of histone proteins is a fundamental post-translational modification regulating chromatin structure, and it plays a significant role in gene transcription [73]. Readers of post-translational modifications are structurally diverse proteins than contain one or more effector modules that recognize covalent modifications of proteins and DNA. The recognition of acetylation of lysine residues is primarily initiated by bromodomains [46]. Bromodomains are involved in the regulation of transcriptional programmers. They have been identified in oncogenic rearrangements [74] that lead to highly oncogenic fusion proteins, which play a crucial role in the development of several aggressive types of cancer [46, 75] (like NUT carcinoma, leukemia and lymphoma [76, 77]).

Recently it has been shown what role is playing BRD4 in NUT carcinoma. NUT carcinoma is a very aggressive and rare form of undifferentiated squamous-cells carcinoma. It is considered one of the most lethal solid tumors, which typically is non-responsive to chemotherapy or radiotherapy and an overall survival spanning from 6 to 9 months [78]. This disease is genetically defined by chromosomal rearrangements involving the NUT gene fused to the BRD4 [78]. This creates BRD4-NUT oncogene that is considered to be a main pathogenetic driver of cellular transformation. It has been found that the interception of the BRD4-NUT fusion gene results in the slowing of the differentiation and growth of NUT carcinoma cells [77, 79].

5.2.2 BRD4 as a therapeutic target

Bromodomain modules share a conserved fold that comprises a left-handed bundle of four α-helices (named αZ, αA, αB and αC) [80] that are linked by diverse loop regions of variable charge and length (known as ZA and BC loops) which surround a central acetylated lysine binding site (**Figure 5.1**). BRD4 can be considered a difficult target for virtual screening because of its flexible structure. Known inhibitors of BRD4 usually form 1 hydrogen bonds with the protein, and the rest of the protein-ligand interactions being hydrophobic [46].

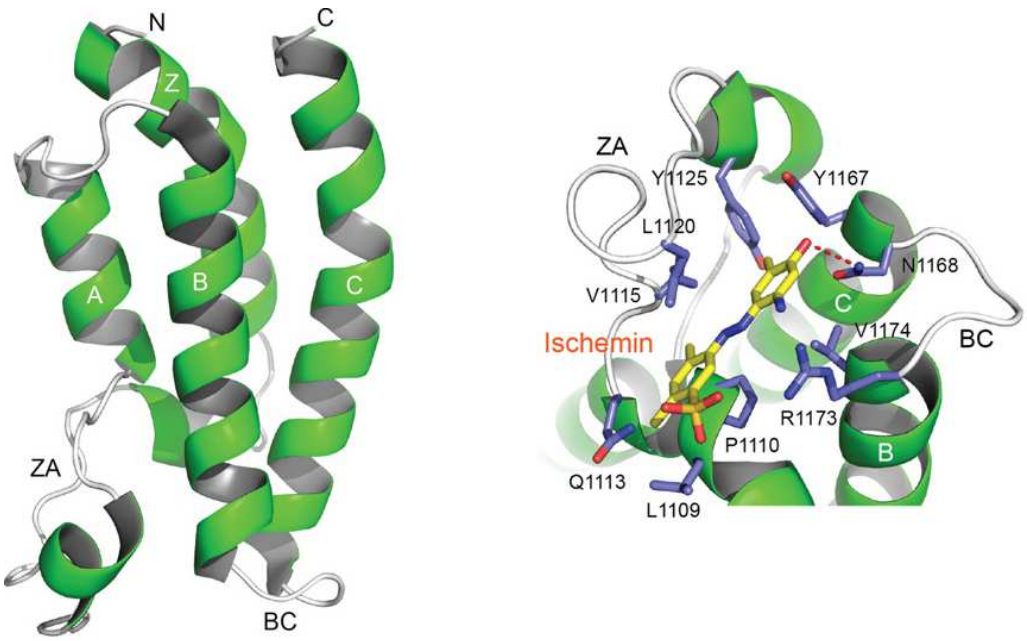


Figure 5.1: On the left – the structure of Bromodomain – 4 α-helices linked by two loops BC and ZA. On the right – BRD inhibitor (Ischemin) interactions with the protein. Binding site residues are shown in sticks. Note that only one hydrogen bond (red dotted line) is made with Aspargine1168 [46].

5.3 Methods

The virtual screening protocol (**Figure 5.2**) of this project involved several machine learning methods (GTM and SVM that have already been described in the Methods chapter), ligand-based pharmacophores and docking. SVM classification models, uGTM BRD4 landscapes, "local" GTM BRD4 landscapes were used in consensus with ligand-based pharmacophore models. Each model has treated the library of 2M compounds, and it ranked them by the likelihood of being active to BRD4, putting the most active compounds on the top of the list. Basing on the predictions of every individual model, 12000 compounds have been selected and sent to the docking procedure. A compound was selected *if* it was ranked among the top 10 by at least two models, *or* it was placed among the top 10000 compounds by > 50% of the models.

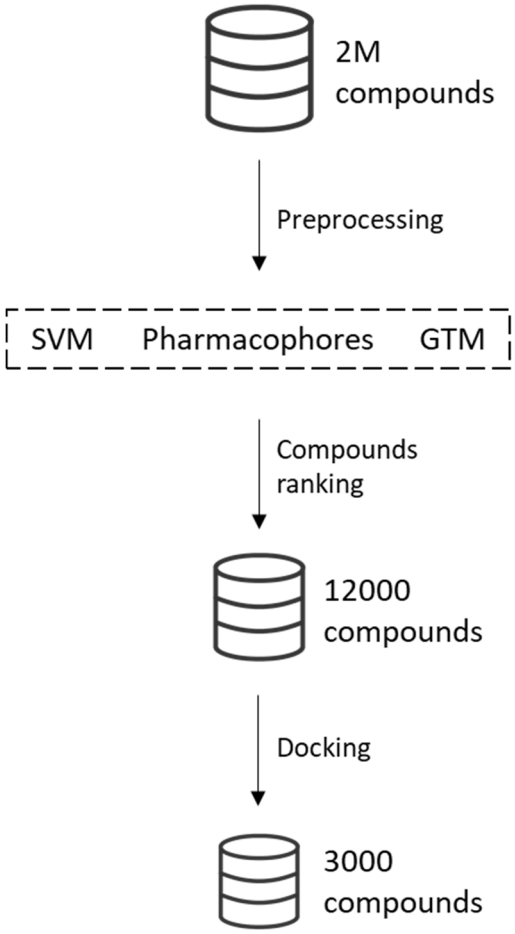


Figure 5.2: Applied virtual screening protocol.

5.3.1 Pharmacophore models

IUPAC's definition of the pharmacophore model is “an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” [81]. A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure. The pharmacophore can be considered as the largest common denominator shared by a set of active molecules. The

pharmacophore features are H-bond acceptors and donors, charged or ionizable groups, hydrophobic groups and aromatic rings.

The spatial relationships between the features in a 3D pharmacophore model can be specified as distances or distance ranges or by defining the (xyz) locations of the features together with some distance tolerance (typically as a spherical tolerance region). There are different possibilities to derive pharmacophore models: based on the three-dimensional structure of a ligand-protein complex (structure-based modeling) or based on the structural information of active compounds only (ligand-based modeling). In this project, ligand-based pharmacophore models have been developed.

Generally, a database is built in such a way that the molecules that it contains are usually (or at least it is expected) represented by a set of conformers that supposedly include the bioactive geometry adopted during the interaction with the target protein. All conformers of used compounds are superimposed, and the associated common pharmacophore features are generated. Then, it is up to the user to define the number and the types of needed pharmacophores that will form the model. Depending on the selectivity of the pharmacophore model, such a virtual screening of chemical databases consisting of millions of small molecules can result in tens to thousands of hits. For the compounds ranking and to model's quality determination, the matching between the pharmacophore model and each molecule of the virtual screening hit list, a score is calculated.

LigandScout [49, 50] was used in the current work. The main feature of LigandScout is the fast alignment algorithm due to the efficiency of the implementation and the advanced geometric similarity measure for the chemical features. In this algorithm, the first step concerns the generation of the 3D pharmacophore features (**Figure 5.3**) identified for each database conformer.

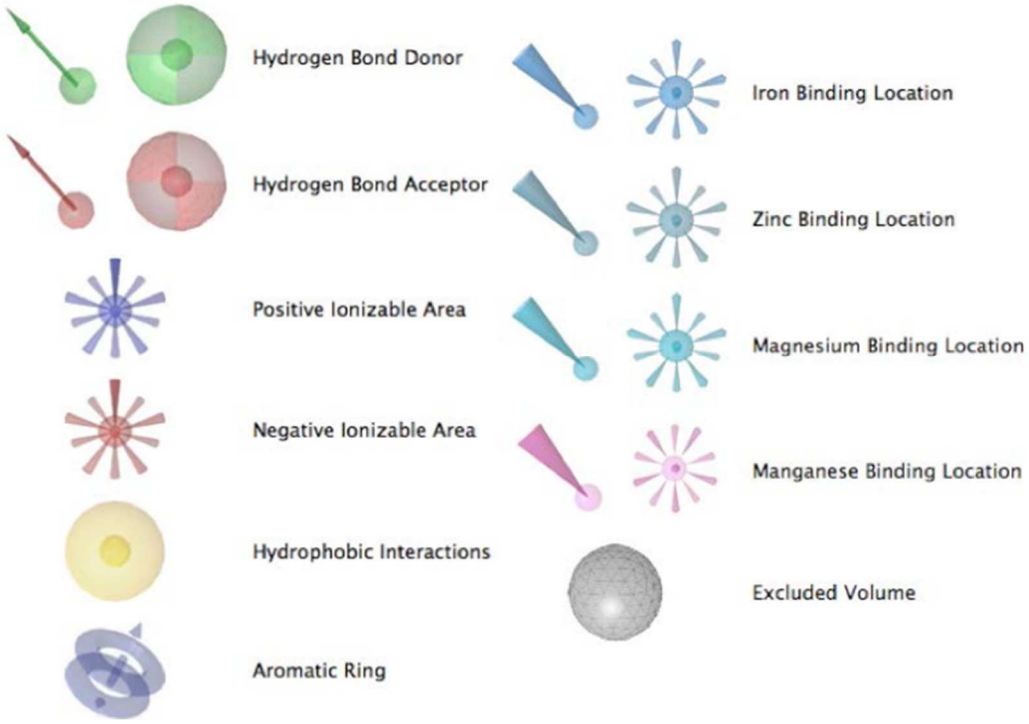


Figure 5.3: LigandScout pharmacophore features.

Then, the algorithm creates for each feature type a set of inter-feature distances. The distance sets created for the pharmacophore model, and the conformer pharmacophore features are then compared in a pairwise manner. In order to perform a pair assignment, the so-called Hungarian matching algorithm is executed. Finally, the feature distances between model and conformer are minimized using Kabsch alignment algorithm. For estimation of alignment quality, the pharmacophore fit score function has been used [49]:

$$S_{RMS} = 9 - 3 \times min(RMS_{FP}, 3)$$

$$S_{FCR} = c \times N_{MFP} + S_{RMS}$$

Where S_{RMS} is the matched feature pair Root Mean Squared Deviation (RMSD) score in range [0,9]; RMS_{FP} is the RMSD of the matched feature pair distances; S_{FCR} is the feature count/RMS distance score; c is a weighting factor for the number of matched feature pairs, and N_{MFP} is the number of geometrically matched feature pairs.

5.3.2 Docking

The docking part of the project was done using S4MPLE (Sampling For Multiple Protein-Ligand Entities) software [51]. It is based on a hybrid genetic algorithm, which allows the simulation of one molecule (conformer generation) or many molecules (docking). Energy calculations are done using the AMBER force field [38] for biological macromolecules and its generalized version - GAFF for ligands. The ability of S4MPLE to indiscriminately handle inter- and intramolecular degrees of freedom is achieved through the appropriate design of torsional angles, rotational and translational degrees of freedom. In S4MPLE, a genetic operator works on some randomly chosen covalently connected (or not) molecular substructure. If the structure is covalently connected, then the operator will affect the structure in such a way that bond length and valence angles will not be changed. If the structure is not covalently connected, then it might be, for example, one of the ligands competing for the binding site. In that case, the guidance role of missing covalent bond will be taken by a potentially favorable contact axis, which is randomly chosen as a pair of atoms (one atom belonging to the external partner and another to structure itself), that should be brought together in order to form a hydrogen bond or a hydrophobic interaction.

The following steps make the preparation of the active site: protein atoms have to be fixed by enumerating their sequence numbers. A predefined cutoff for non-bonded interactions was established on 12Å. Protein atoms that are too far from the active site in order to ever come within 12Å to any ligand atom would merely slow down calculations by requesting the regular update of their distances to ligand atoms. Therefore, docking was not run on the entire protein, but on the selection of relevant residues that have at least one atom at less than 10Å from any of the co-crystallized ligand, herewith used to define the active site region. Moreover, S4MPLE requires the user-specified input of "hot spots" – key solvent-accessible atoms, chosen preferentially at the bottom of the site cavity, which will be used for random prepositioning of the ligand into the active site. These may, but do not have to, include site atoms seen to make contacts to the co-crystallized PDB ligand.

Ligands, initially provided as standardized SMILES, preprocessed by the standardization tool of the Strasbourg virtual screening web server, undergo an automated conversion, using an in-house tool developed based on the ChemAxon API, to a fully protonated initial 3D structure. The tool relies on the tautomer and, respectively, pKa plugin to generate the most probable microspecies of the expected main tautomeric form. Users might request several tautomeric or protonation states to be generated, and each to be docked as an independent candidate – but the option was not used here. Explicit hydrogens are assigned, and the conformer plugin then generates a single conformer. Eventually, the charge plugin is used to assign Gasteiger charges [82] to this structure. Last but not least, the tool detects flexible rings and proposes, for each, the single bond to be formally "broken" in order to enable intra-cyclic torsional axes to be driven by S4MPLE.

S4MPLE docking begins by extracting all the data of a given ligand into a dedicated directory, then running a 200-generation evolutionary conformational search with S4MPLE, on the free ligand, at default settings. Next, active site data are added to this directory, and a brief fist simulation is run in order to calibrate the optimal cutoff for the interaction fingerprint dissimilarity value (*minfpdiff*), representing the threshold at which two conformers are considered as redundant, and thus pruned during the evolutionary process. The proper management of population diversity has been noted to be of paramount importance for ensuring the convergence/reproducibility of evolutionary simulations. As ligands vary in sizes, so does their interaction fingerprint, making it challenging to come up with a universally applicable threshold value – hence, the need to calibrate it for each system. The population initialization procedure, regularly serving as the first step for the evolutionary simulation, is called repeatedly (10 times). After each call, the interaction fingerprints of the randomly generated population members are compared to each other, generating the complete Hamming distance matrix for all pairs of conformers in the population. The lowest, mean and maximal Hamming distances for each population are

memorized. The *minfpdiff* threshold is defined as 90% of the average of the ten lowest intra-population Hamming distances.

Eventually, the main docking simulation is started with the above-determined minfpdiff value as a population diversity control parameter. Top poses are generated and stored together with their energy values $\langle E_i^{ligand@site} \rangle$. The docking index ΔE $for the current ligand can be directly estimated as $\langle E_i^{ligand@site} \rangle - \langle E_i^{ligand} \rangle$. After completion of docking calculations for all ligands, these can be ordered by increasing ΔE, and the final ROC curve can be generated in order to determine the area under it, as the final benchmarking criterion. The variation of the ROC AUC as a function of the performed number of generations may be informative about the minimal required computational effort needed in typical S4MPLE docking simulations

5.4 Results and discussion

The structures of 3000 selected compounds have been transmitted to our Enamine collaborators. They were able to test 2992 compounds, and the experiments confirmed 29 hits. While this result is objectively low, it is still 2.6 times better than the hit rate found in the random screening of 3200 compounds under identical conditions [83]. However, it has to be mentioned that the applied classification models have been trained using publically available SAR data on IC_{50} values. In contrast, our collaborators have measured the Thermal Shift Assay using Differential Scanning Fluorimetry (DSF). DSF is a biophysical method based on detecting the shift in protein denaturation temperature upon ligand binding, as reported by fluorescent dye interacting with the protein core exposed by heat denaturation. The method is a simple, label-free HTS technology applicable to most soluble proteins, irrespectively of their functions and activities. An in-depth analysis was performed in order to understand:

- How the public-data affinity values used for model building relate to the experimental hit detection criterion (ΔTm) used in DSF?

The original dose-response (such as IC_{50}) activity scores from public sources were shown to be *per se* rather poorly correlated to the hit selection criterion DSF-ΔTm. The fact that they were not used as such for model training, but first underwent conversion into a categorical variable has most likely had a negative impact on model performance.

- Which of the used models are better at selecting the 29 confirmed hits?

Seventeen hits have been ranked #1 by at least one of the GTM models, while two were ranked #1 by SVM. The other ten hits were selected because of "broader" consensual selection by multiple models that ranked them within the top of the list.

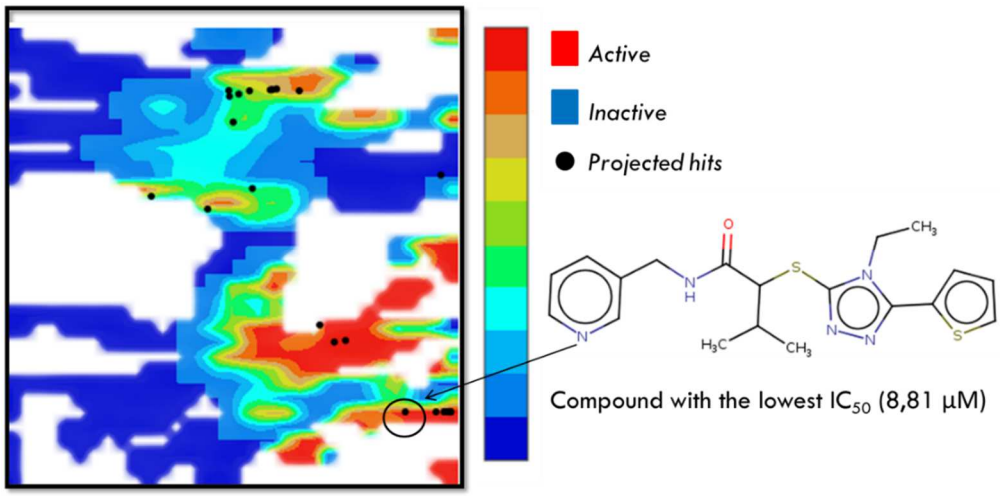


Figure 5.4: Confirmed hits projected on one of the used GTM landscapes. The red and blue zones of the map are populated by, respectively, active and inactive compounds. The regions of the map colored in "intermediate" colors are populated by the compounds of both classes.

More detailed description of the obtained results is given in our article published in Eur.J.Med.Chem., see below.

European Journal of Medicinal Chemistry 165 (2019) 258–272

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Research paper

Pros and cons of virtual screening based on public "Big Data": In silico mining for new bromodomain inhibitors

Iuri Casciuc [a], Dragos Horvath [a], Anastasiia Gryniukova [b], Kateryna A. Tolmachova [c, d], Oleksandr V. Vasylchenko [c], Petro Borysko [b], Yurii S. Moroz [e, f], Jürgen Bajorath [g], Alexandre Varnek [a, *]

[a] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal str, 67081, Strasbourg, France
[b] Bienta/Enamine Ltd, Chervonotkatska Street 78, Kyiv, 02094, Ukraine
[c] Enamine Ltd, Chervonotkatska Street 78, Kyiv, 02094, Ukraine
[d] Institute of Bioorganic Chemistry & Petrochemistry, NAS of Ukraine, Murmanska Street 1, Kyiv, 02660, Ukraine
[e] National Taras Shevchenko University of Kyiv, Volodymyrska Street 60, Kyiv, 01601, Ukraine
[f] Chemspace, ilukstes iela 38-5, Riga, LV, 1082, Latvia
[g] B-IT, Limes, Unit Chem. Biol. & Med. Chem, University of Bonn, Germany

ARTICLE INFO

Article history:
Received 31 October 2018
Received in revised form
24 December 2018
Accepted 5 January 2019
Available online 9 January 2019

Keywords:
Bromodomain BRD4 binders
Generative topographic mapping
Virtual screening
Classification models
Ligand-based pharmacophores
Docking

ABSTRACT

The Virtual Screening (VS) study described herein aimed at detecting novel Bromodomain BRD4 binders and relied on knowledge from public databases (ChEMBL, REAXYS) to establish a battery of predictive models of BRD activity for in silico selection of putative ligands. Beyond the actual discovery of new BRD ligands, this represented an opportunity to practically estimate the actual usefulness of public domain "Big Data" for robust predictive model building. Obtained models were used to virtually screen a collection of 2 million compounds from the Enamine company collection. This industrial partner then experimentally screened a subset of 2992 molecules selected by the VS procedure for their high likelihood to be active. Twenty nine confirmed hits were detected after experimental testing, representing 1% of the selected candidates. As a general conclusion, this study emphasizes once more that public structure-activity databases are nowadays key assets in drug discovery. Their usefulness is however limited by the state-of-the-art knowledge harvested so far by published studies. Target-specific structure-activity information is rarely rich enough, and its heterogeneity makes it extremely difficult to exploit in rational drug design. Furthermore, published affinity measures serving to build models selecting compounds to be experimentally screened may not be well correlated with the experimental hit selection criterion (in practice, often imposed by equipment constraints). Nevertheless, a robust 2.6-fold increase in hit rate with respect to an equivalent, random screening campaign showed that machine learning is able to extract some real knowledge in spite of all the noise in structure-activity data.

1. Introduction

The exponential accumulation of structure-activity data in public databases, representing the advent of Big Data in medicinal chemistry is expected to lead to the development of robust and potent *in Silico* Quantitative Structure-Activity Relationships (QSAR), mathematical models able to serve for Virtual Screening (VS) of compound databases, *i.e.* detect and prioritize novel active structures therein and herewith accelerate drug discovery. Both predictive accuracy and Applicability Domain (AD) of QSAR models are expected to increase significantly with the size and chemical diversity of training sets, while machine learning has already provided Big Data-compatible tools for the fitting of such models. Methods like Support Vector Machines (SVM) [1] and Generative Topographic Mapping (GTM) [2] routinely provide QSAR models based on tens of thousands of compounds. GTM – essentially a fuzzy-logic-based variant of popular Self-Organizing (Kohonen) Maps [3] – has no upper limit on training set size, as GTM-driven predictive models consist of property landscapes that are "colored" (created) by projecting known reference actives and inactives on the map, and attributing to every map point a property value equaling the mean of therein residing compounds. For

* Corresponding author. .
E-mail address: varnek@unistra.fr (A. Varnek).
URL: http://www.chem-space.com

https://doi.org/10.1016/j.ejmech.2019.01.010
0223-5234/

prediction, candidate compounds are also projected on the map, and are assigned the property value of their residence spot or declared "out of AD" if they fall into blank spots, where no reference compounds are residing. GTM [4–7] is a multivalent "Swiss-army-knife"-like tool of chemoinformatics, as one of the rare tools competent for both chemical space visualization (with particular interest in library comparison) and predictive modeling, including implicit AD assessment. Albeit it is not primarily designed for VS, the latter abilities nevertheless qualify GTM as a robust VS methodology, whilst its visualization support may be useful to graphically compare relevant compound sets (here, reference "actives" from various sources, *vide infra*). By contrast, SVM is one of the most powerful QSAR predictors. Both of these methods are fast since they can operate on 2D molecular descriptors avoiding the need of a costly conformational sampling step. Pharmacophore models and, eventually, docking, can be used in conjunction to the fast 2D ligand prioritization tool in a VS funnel, to gradually focus in on candidates with maximum probability to be active.

In practice, however, medicinal chemistry data is rather heterogeneous. The nearly two million compounds in ChEMBL are often associated with reliable IC_{50}/K_i measures, but these concern a plethora of different targets. Thus, compound sets associated with a given target are more often likely to represent classical QSAR sets of hundreds of compounds, rather than Big Data sets. Moreover, dose-response activity measures are typically reported by different groups and may follow distinct protocols, which raises the question whether they are comparable. The key point here is that in "classical" medicinal chemistry, the expert is closely following the work of colleagues/competitors and is familiar with all those distinct protocols, knowing what is comparable. Or, in the Big Data era, the information is allegedly too rich to be trackable by a human expert. Structure-activity sets should be algorithmically extracted, standardized and processed into training sets – with no human intervention. Is this a realistic scenario, or would data heterogeneity eventually outweigh the benefits of information richness provided by on-line public databases? This is a central question addressed in this work, which reports a "Big Data" VS search, followed by experimental validation, for novel BRD4 inhibitors. A database of 2 million available compounds from Enamine (enamine.net) was virtually screened using a hierarchy of 2D QSAR methods coupled to pharmacophore screening based on docking and publicly available structure-activity data from REAXYS [8] and ChEMBL [9] databases for automated model training.

Readers of post-translational modifications are structurally diverse proteins than contain one or more effector modules that recognize (that is, read) covalent modifications of proteins and DNA. The recognition of ε-N-acetylation of lysine residues is primarily initiated by bromodomains, a family of evolutionarily conserved protein interaction modules that were identified in the early 1990s in the brahma gene from *Drosophila melanogaster* [10]. The human genome encodes 61 bromodomains present in 46 different proteins [11,12], where differences in the amino acid residues around the acetyl-lysine binding site impart ligand specificity. Proteins that contain bromodomains are involved in the regulation of transcriptional programs and have been identified in oncogenic rearrangements that lead to highly oncogenic fusion proteins, which have a key role in the development of several aggressive types of cancer. They are also implicated in the replication of viral genomes and regulate the transcription of some viral proteins.

Bromodomain modules share a conserved fold that comprises a left-handed bundle of four α-helices (named αZ,αA, αB and αC) that are linked by diverse loop regions of variable charge and length (known as ZA and BC loops) which surround a central acetylated lysine binding site. Structural data have established that acetylated lysine is recognized in a central hydrophobic pocket, where it is anchored to a conserved asparagine residue. More recently, it has been demonstrated that BRD4 bind to two acetylated lysine histone marks that are simultaneously recognized by the same bromodomain module [13]. This property is shared by all members of the Bromodomain Extra-Terminal (BET) subclass of BRDs. High-resolution crystal structures showed that the first acetylated lysine mark of histone H4 docks directly onto the conserved asparagine (Asn140 in the first bromodomain of BRD4). Simultaneously, a network of hydrogen bonds, formed via conserved water molecules found in the bromodomain cavity, link to the second acetylated lysine mark, thus stabilizing the peptide complex.

BRD inhibitors are reported in several public databases – ChEMBL and REAXYS were the ones exploited here and reported inhibition strength stem from various methods such as DSF or FRET experiments. The only way to cope with data heterogeneity was to base the VS protocol on categorical models, returning an estimate of the likelihood of a candidate compound to be "active". Training of categorical models however implies an upstream classification of so-far tested compounds into "actives" and "inactives". The choice of these examples of actives and inactives used in the machine learning process is empirical, as it implies setting arbitrary thresholds in terms of the available affinity scores. These thresholds might not only be activity score-specific but would also depend on the stage of the hit or lead discovery process. Whilst at primary screening stage a 10 μM affinity level might count as "active", this will no longer be the case in the more ambitious hit-to-lead development stage. As no obvious consensus in designing the "active" BRD training set could be reached, several distinct training sets were employed in parallel, featuring various working hypotheses concerning the "actives".

A battery of SVM and GTM models, combining above-mentioned training set choices and various methodological strategies were built. In parallel, structure-based pharmacophore models were derived from BRD4-ligand crystal structures. All these were used to screen the 2 million compound library of Enamine, and 12000 structures were selected on the basis of a consensus scheme. Experimental screening of a fixed-size pool of 3000 candidates has been carried out by Enamine, representing a VS-driven alternative to a similar screen done on a randomly picked compound set of same size.[1] Docking with S4MPLE [14,15] was used to further reduce the primary 12K compounds to the final pool of 3000 molecules submitted to testing. Experimental DSF retrieved 29 hits (1%) in the 3K VS-based library – roughly three times more than the base hit rate in the above-mentioned random screening experiment [16]. This is a significant, yet slightly disappointing enrichment factor. As a consequence, more effort has been allotted to better understand the discrepancies in affinity measurements introduced by different methods. On one hand, some ChEMBL training set compounds with reported IC_{50}/K_i values had their melting temperature shift (ΔT_m) measured by DSF under the same conditions as the herein retrieved hits. Alternatively, IC_{50} values for some of the newly discovered hits were also experimentally determined. The weak to moderate correlation between the actual hit detection criterion (ΔT_m) and the dose-dependent public data affinity scores (on which model training was based) has significantly and negatively impacted the success rate of this large-scale VS experiment. An *a posteriori* analysis of individual models, aimed to verify how well each one performed in ranking the 29 hits within the 3K selection, showed that hit rates for the most successful individual models could have been as high as 10%. However, for every successful model, alternative models of the same category – differing only with respect to the choice of the "actives" defined in the training set, and the added ChEMBL "decoy" molecules – were found to be low performers. This is clear evidence that the problem stems from data variance or noise and not from rigorously

cross-validated models.

2. Materials and methods

2.1. Training sets

2.1.1. Two sources were used in this project

- The REAXYS set contains 75 strong actives (compounds having $IC_{50} \leq 100$ nM), 404 moderate actives (compounds having IC_{50} between 100 nM and 10 μM) and 742 inactive ($IC_{50} > 10$ μM) molecules. In order to remain within the two-class classification strategy, this set was duplicated into two "clone" sets differing only with respect to the assignment of the class labels. The Strict set considers only the 75 strong as "active", and all others are "inactive". The Soft set counts both the 75 strong and the 404 intermediates as "actives".
- The ChEMBL data, where active versus inactive BRD compounds were extracted automatically, as part of a data curation procedure internal to the Laboratory of Chemoinformatics [6] (120 "active" and 554 tested "inactive"). This set is marginally overlapping with the REAXYS data, sometimes with conflicting activity class assignment.

STRICT, SOFT and ChEMBL were thus considered as three independent training sets, and used for model calibration and/or class landscape coloring, in conjunction with random decoy compounds, assumed as inactive and randomly picked among the non-BRD molecules in ChEMBL.

2.1.2. Screening set

The provided screening set contained 2 million compounds (synthesized at Enamine) encoded in SMILES format.

2.1.3. Virtual screening protocol

In this project the VS protocol included following steps, as detailed below:

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

2.2. Compound standardization and description

Compound standardization followed the default protocol installed on our public web server (infochim.u-strasbg.fr/webserv/VSEngine.html), powered by ChemAxon [17] tools. It includes:

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

The descriptors used here were ISIDA descriptors computed by ISIDA Fragmentor [18,19]. More than 100 different types of descriptors sets were generated. They include sequences, atom pairs, circular fragments and triplet counts of different length, colored by formal charges, pharmacophore features or force field types.

3. Modeling methodology

3.1. Support vector machine

The Support Vector Machine (SVM) is a machine learning method developed by Vapnik [1]. The input variables are mapped into a higher dimensional feature space using a kernel function, and then a linear model is built on this new feature space. The most common kernel functions include linear, polynomial and radial basis functions. The performance of this method depends on type of kernel and a number of parameters. SVM performs classification by finding the hyperplane that maximizes the margin between the two classes.

SVM models were validated by means of a 3 fold Cross-Validation (CV) procedure repeated 12 times. These were built on the basis of STRICT and respectively SOFT training sets, each randomly completed with a number of 1221 decoys (i.e. as many decoys as total BRD compounds). Model fitting was performed using the libSVM model optimizer [20], and resulting consensus were posted on the Strasbourg web server. There were thus two (STRICT and SOFT) SVM consensus models serving for selection.

3.2. Generative Topographic Mapping

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

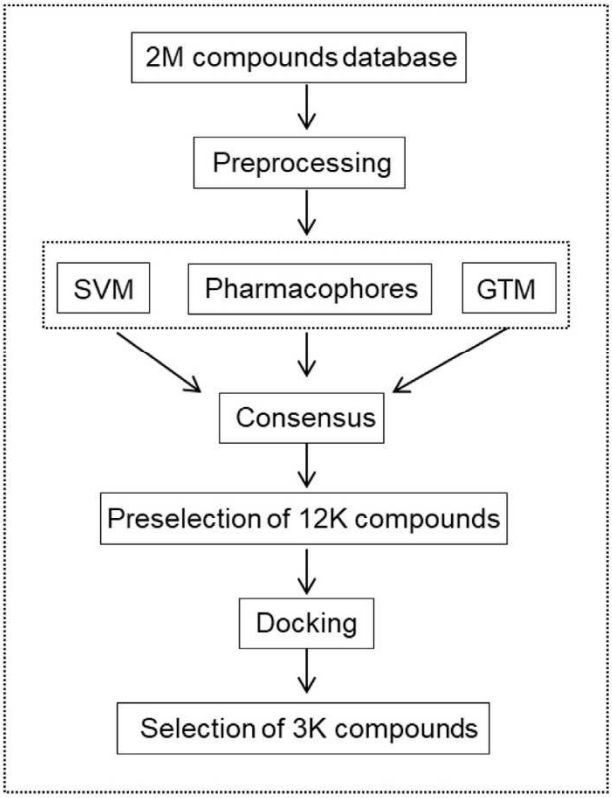


Fig. 1. Applied Virtual Screening protocol.

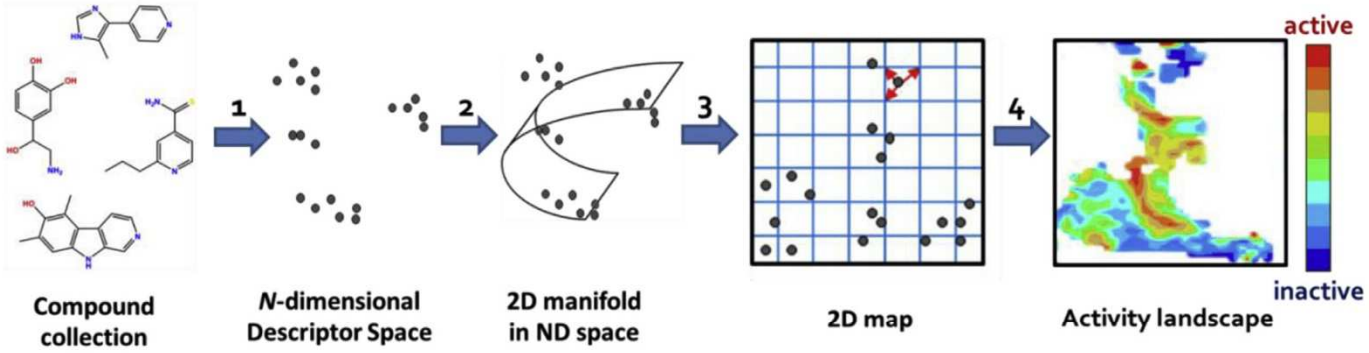


Fig. 2. Generative topographic mapping.

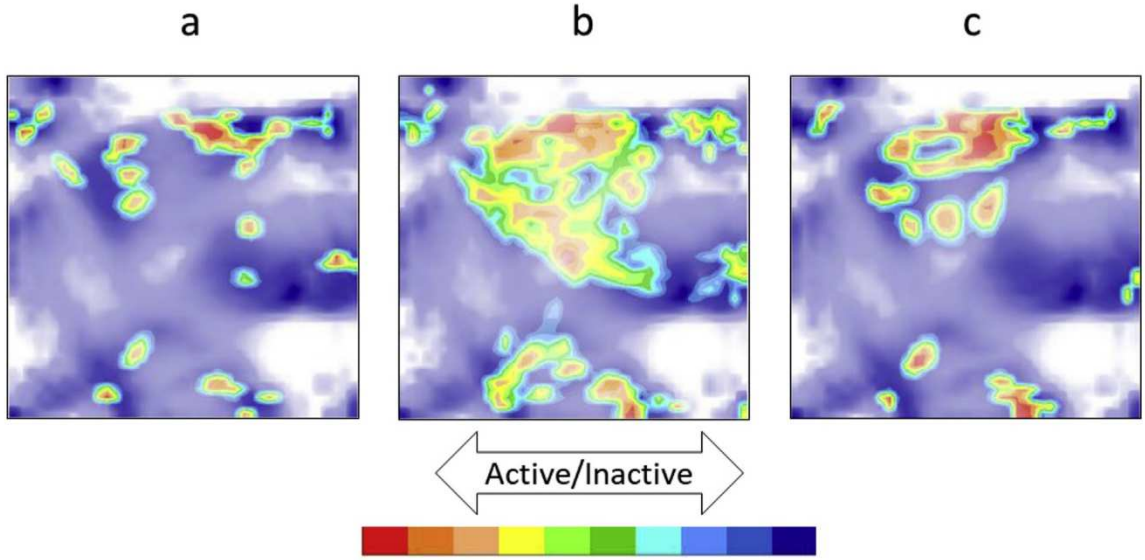


Fig. 3. Fuzzy classification landscapes on UGTM1, highlighting zones populates by the actives of the – a) CHEMBL, b) SOFT and c) STRICT training sets, against a common background of the ~1.5M ChEMBL v.23 compounds that were not tested on BRD. These plots apply Bayesian normalization in order to compensate for the extreme imbalance between active and inactive set sizes.

represent unpopulated areas.

GTM activity class landscapes are obtained after the "transfer" of the knowledge about the most likely class to be encountered in a given chemical space neighborhood onto the latent grid nodes that represent this neighborhood. The prediction implies locating the candidate into one of these neighborhoods represented by the population of the nodes, therefrom learning the class to which it should be assigned. GTM-driven predictors typically behave like Nearest-Neighbors-based predictors, which includes the support of identification of candidates outside of its applicability domain, i.e. compounds which do not sufficiently resemble to any of the reference compounds in order to allow an extrapolation of their properties in virtue of the similarity principle.

Eight (4 universal [6] and 4 local) maps based on eight distinct ISIDA fragment descriptor spaces were used in this study, as described below:

Note that some of the maps are built on hand of descriptors (detailed atom-centered fragments) capturing connectivity information, whilst other rely on fuzzier atom pair counts, while still others rely on topological pharmacophore descriptors. If the virtually screened library contains compounds from the same chemical series of reference actives (featuring a roughly same scaffold and/or pharmacophore pattern), these will be consensually selected by all the maps and models, irrespective of underlying descriptor space. However, virtual hits may be only partially related to reference actives, so that only the maps able to recognize the specific underlying similarity will be able to retrieve these compounds At one extreme, candidates may be scaffold-hopping analogues of reference compounds, typically not perceived as similar by the human eye. In this case, maps focusing on connectivity-based similarity criteria might exclude such candidates from their AD. Pharmacophore descriptor-based maps will, by contrast, successfully recognize their "matching" pharmacophore patterns. Last but not least, it is important to highlight that similar activity of two compounds does not imply any underlying structural similarity: two actives may have both distinct topologies and distinct pharmacophores, because they bind to different (sub)pockets of the active site. No machine learning technique could infer the activity of the one based on the example of the other – only docking could in principle predict that both are interacting favorably with the site.

3.3. Class landscapes based on universal generative topographic maps

Universal GTMs were built independently of this work, as "best compromise" maps, able to properly accommodate a maximum of classification landscapes for very diverse biological properties. These GTMs were proven to successfully serve as hosts for 618

classification landscapes associated to the respective target-specific structure-activity ChEMBL compound series and providing significant separation of actives from inactives. Note that the herein employed, automatically extracted CHEMBL BRD4 training set is one of the above-mentioned 618 targets. More specifically, it did not serve at map building stage, but was one of the external validation sets in that study. Each of the four maps was used to "color" a BRD class landscape according to each of the 3 sets (STRICT, SOFT, CHEMBL) which were supplemented with ChEMBL decoy compounds. For each set, two distinct landscapes were obtained by toggling the Bayesian normalization option on/off (this latter serves to "enhance" the impact of rare actives in the "ocean" of inactives on the landscape). With Bayesian normalization on, 5% of non-BRD ChEMBL molecules were randomly added as decoys (note – different 5% being used for each landscape). Without normalization, only 1% of the non-BRD ChEMBL compounds were added as decoys. Thus, the combination of 4 maps × 3 sets x 2 normalization options produced 24 distinct "Universal" BRD class landscapes. In order to keep track of individual models, we propose, for each such landscape, the nomenclature scheme UGTM(map number, 1–4)-(BRD training set: SOFT, STRICT, CHEMBL)-DEC (decoy set ID) –BN(Bayesian normalization toggle on or off)". For example, UGTM2-SOFT-DEC0-BNon is the landscape based on universal manifold #2 (as labeled in the article describing it), considering the SOFT BRD training set completed with the pool DEC0 of random ChEMBL decoy compounds, and using Bayesian normalization.

4. Class landscapes based on dedicated generative topographic maps

Dedicated (or local) GTMs were built with the goal to specifically achieve optimal separation of BRD actives from inactives. For this purpose, the evolutionary map builder procedure used for universal map generation was employed with the key restriction of using as "selection sets" the decoy-enhanced STRICT, SOFT and CHEMBL training sets. Note that while the four universal GTMs work each in a given ISIDA descriptor space – selected independently of this BRD4-related project, the evolutionary optimizer of dedicated maps is free to pick, out of the 100 distinct fragmentation schemes considered, the ISIDA descriptor space(s) that specifically maximize separation of BRD4 actives and inactives. The BRD data sets were each "triplicated" (STRICT1, STRICT2, STRICT3, SOFT1, etc.) by addition of different pools of ~5000 decoys. Four DGTMs with top separating propensities for BRD compounds were retained. For each of the 4 DGTMs, BRD landscapes were created by coloring with each of the decoy-enhanced triplicates of the three sets, again with and without Bayesian normalization – this gives 4×3 x $2 = 72$ landscapes based on the dedicated GTMs. The same model nomenclature introduced for universal maps will be used, however using the "DGTM" label for these BRD-dedicated maps.

4.1. Ligand-based pharmacophores

LigandScout [21] was used in the current work.

The models were obtained on the basis of STRICT dataset. The procedure is the following:

1. Generation of conformers of each molecule, with an RMS threshold of 0.5 and energy window 15 kcal/mol, having as maximal number of possible conformers set to 25.
2. Clustering the ligand sets according to the geometry of the 3D pharmacophoric features. Here Pharmacophore radial distribution function was used for similarity calculations. Cluster distance was set to 0.45.
3. Five different pharmacophore hypotheses capable to accommodate actives and discard inactives were considered (see Supporting Information).

4.2. Virtual screening using QSAR and pharmacophore models

The Enamine collection of 2M compounds was first submitted to standardization, according to the internal procedure of the Strasbourg web server. The molecular descriptors (ISIDA fragment counts) required for the predictive models were generated. Alternatively, stable conformers were enumerated for the compounds, and submitted to the pharmacophore matching procedure of LigandScout, which allowed ranking of all the 2M candidates by their quality of fit into each of the five pharmacophore models.

Each GTM landscape is a predictive model, since projecting a candidate compound onto it allows to "read" its propensity to be active. Furthermore, GTM projection may explicitly assess the pertinence of each prediction, which is trustworthy if (a) the projected candidate compound is close to the GTM manifold in original descriptor space (it has a "LogLikelihood" criterion similar to the frame compounds used to build the manifold), and (b) if it resides in an area of the map which hosts many compounds from the set used to color the landscape. Both aspects (a) and (b) were used, for each landscape-based prediction, to discard candidates not fulfilling the conditions (technically, they were "ranked" at the bottom of the preference list). The ranking of the other candidates was done according to the propensity to belong to the active class, as read from the landscape.

Similarly, the consensus SVM models also predict the propensity to belong to the active class, and also provide various measures for assessing the applicability of the model to each candidate. Likewise, candidates within the AD were ranked, for each model, according to predicted propensity to be active, whilst the ones out of AD were ranked as lowest priority.

Thus, each of the 2 SVM models +24 UGTM landscapes +72 DGTM landscapes +5 pharmacophore models proposed their own ranked list of candidates, for the 2 million Enamine compounds. No single molecule was systematically ranked number one by all the approaches. Therefore, a "frequency@TopN" (f@N) empirical scale was established for the final selection: selected compounds are asked to achieve some empirically established minimal frequency of presence within the TopN of some methods, where N was varied. f@N represents the number of models that have simultaneously ranked compound C among the most promising top N. The lower the chosen N, the lower will be f@N. In other words, the notation f@50, for example, means how many models have ranked the compound C in top50. At low N – in particular, for $N = 1$, the event of retrieving the same molecule ranked #1 by many independent models is quite rare. Being ranked #1 by only a few of the different models is a good enough reason to be kept for the final selection. By contrast, being a member of the much broader Top1000 is less "prestigious" – as compensation, membership in Top1000 must be achieved with a significantly higher frequency in order to justify the selection of the compounds.

Selection follows thus a "Pareto" philosophy – some compounds are selected if some few models give them an excellent ranking, while other are coopted because very many models give them an acceptable ranking. By empirically choosing minimally required thresholds for frequency@TopN values, a pool of 12K compounds was preselected.

This 12K preselection was submitted to docking into the BRD receptor, using the S4MPLE program. This provided an estimation of their binding energy as the final selection criterion. The 3000 best dockers (with lowest calculated binding energies) were

communicated to the Enamine team, in view of experimental assessment of their BRD4 affinity.

4.3. Docking

Docking was performed with S4MPLE, (Sampler For Multiple Protein or Ligand Entities) a conformational sampling tool [14,15] based on a hybrid genetic algorithm, which allows the simulation of one molecule (conformer generation) or many molecules (docking). Energy calculations were carried out using AMBER force field for biological macromolecules and its generalized version - GAFF for ligands. Here, S4MPLE was used for standard, rigid docking into the active site of the BRD4 structure (PDB code 3MXF), which was assigned standard protonation states for amino acid side chains and then truncated to a sphere of residues with at least one atom within 12 Å from the co-crystallized ligand. Site atoms directly interacting with the ligand were set as "hot spots" for the initial position of ligands by S4MPLE. Ligand processing and docking followed the standard S4MPLE procedure previously described [22] and, like in the cited protocol, the binding energy difference served as final docking score. Before applying S4MPLE to select the 3000 best docking candidates of the 12K pool preselected by the SVM/GTM QSAR models, it was first challenged to dock the REAXYS training set of 1221 actives and inactives, completed with 1221 randomly picked ChEMBL decoy compounds, assumed BRD4-inactives. Fig. 5 shows that the BRD4 cavity is mainly hydrophobic formed at one end of the BRD α-helix and the residues of the αZ-αA and αB-αC loops, thus leading to the fact that the nature of binding pocket allows various possible interactions with the ligands.

4.4. Experimental testing protocol

Compounds were experimentally tested using DSF, which detects the shift in protein denaturation temperature upon ligand binding as reported by fluorescent dye interacting with protein core exposed by heat denaturation. DSF is a simple, label-free HTS technology applicable to most soluble proteins, irrespectively of their functions and activities. BRD4 sequence fragment corresponds to sequence entry O60885.1 in UniProtKB Database [24]. Represents domain 1 (44–168 AA), contains N-terminus His6-tag and 16-amino acid linker. For details, please refer to previous publications already reporting the use of this experimental protocol [16].

5. Results

Out of the 3000 selected compounds, 2992 were actually tested and 29 were found to be active. Molecular structure of the hit with the lowest IC50 value is shown on Fig. 7. A recently screened random selection of 3200 compounds tested with the same technique gave 0.375% of hits [16], *e.g.* achieved a 2.6 times lower hit rate. VS has thus clearly enhanced the hit rate, albeit a higher enrichment score was expected (see Table 1).

5.1. Structural novelty of discovered hits

In order to assess the originality of the novel hits, they were encoded as ISIDA fragment descriptors, using the three fragmentation schemes that were selected by the DGTM models (see Table 2). Pairwise Soergel distances (1-Tanimoto similarity) were calculated, in each descriptor space, between the 29 hits and all the active BRD4 compound present both in SOFT and CHEMBL training sets. Considering the lowest distance in either of the descriptor spaces, 18 of the 29 hits were found to display at least one of the training actives within a neighborhood radius of 0.2. Fig. 8 displays the hits closest to training active in the 4th DGTM map descriptor space systematically returning lowest Soergel distances.

The close relationship to known actives is visible – scaffold being often shared, but not always. Even within the eight closest pairs, three examples of "scaffold hopping" are present.

Understanding the reasons of this modest success is a direct opportunity to investigate the strength and pitfalls of this VS strategy based on public data for model training. The two following paragraphs address two key questions:

- How do public-data affinity values that served to build the model relate to the experimental hit detection criterion ΔT_m?
- Which of the specific models were better at prioritizing the 29 discovered hits, and why?

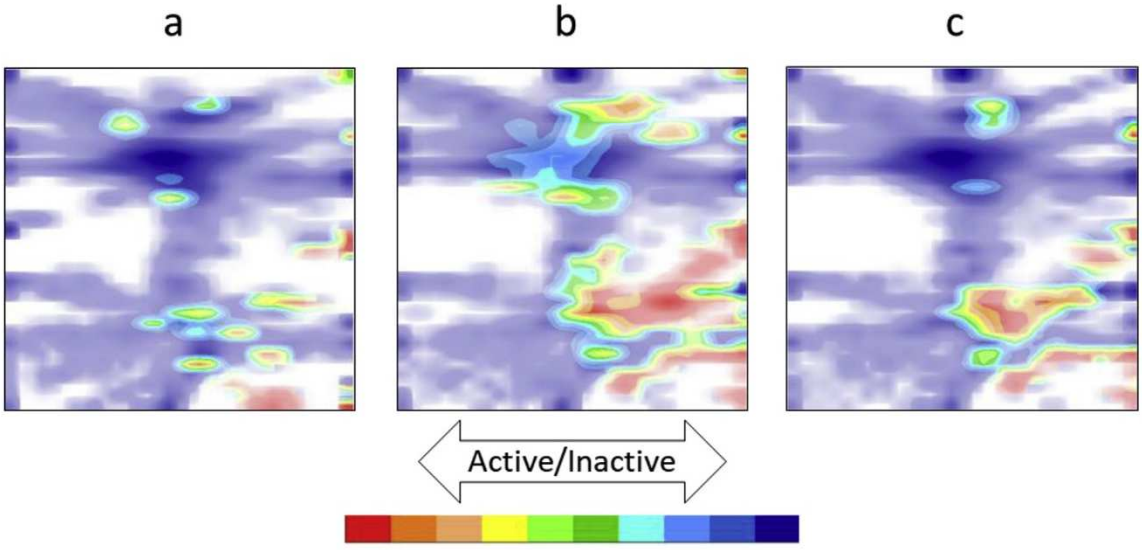


Fig. 4. Fuzzy classification landscapes on DGTM2, highlighting zones populates by the actives of the – a) CHEMBL, b) SOFT and c) STRICT training sets, against a common background all the inactive BRD compounds.

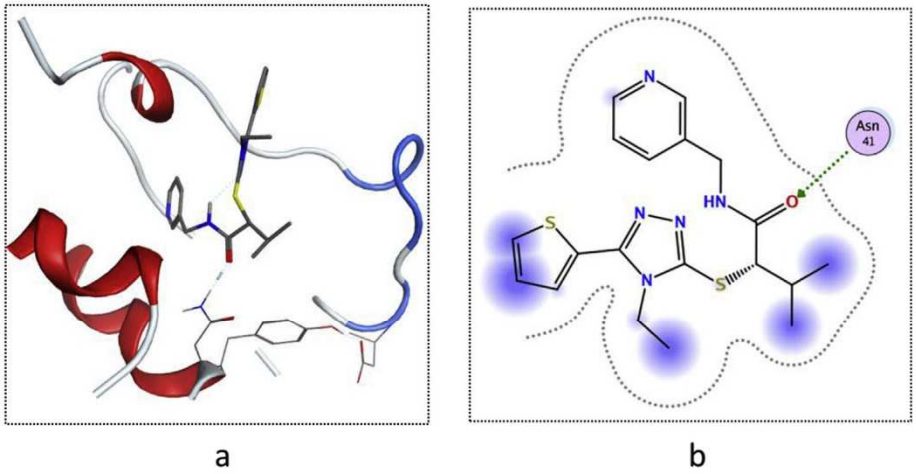


Fig. 5. Docking pose (a) of the hit with the lowest IC_{50} value, and associated 2D interaction map according to the MOE [23] software (b).

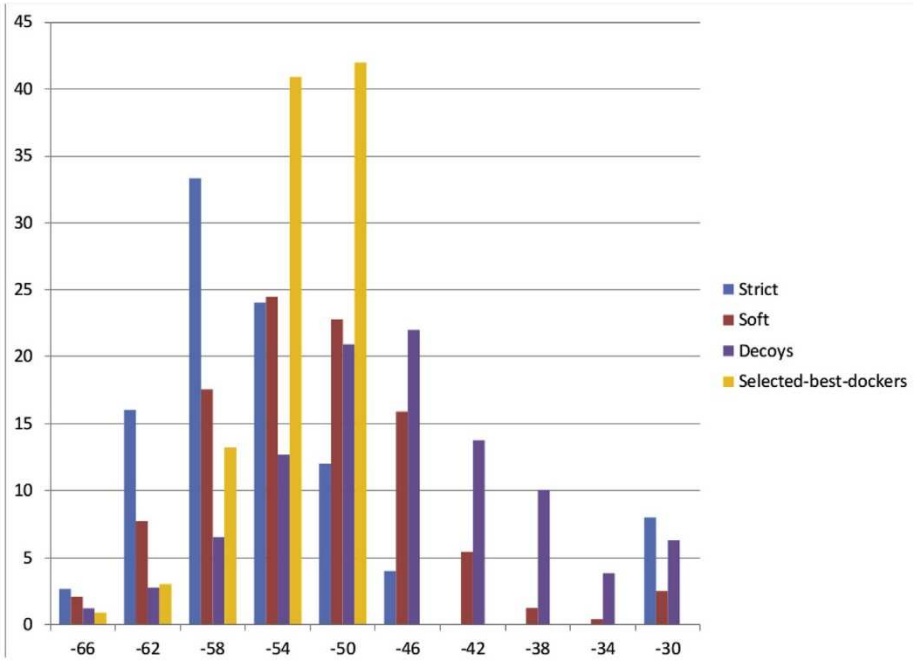


Fig. 6. Distribution of docked compounds by binding energy. X-axis: S4MPLE binding energy bins, Y-axis: the percentage of compounds of a given set found to score the given energy.

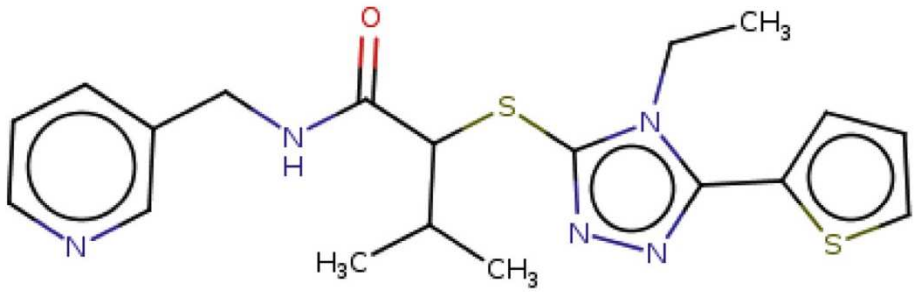


Fig. 7. Structure of the hit having the lowest IC_{50} value. Structures and activities of all 29 hits are given in Supplementary Materials.

Table 1
Detailed description of individual SVM models included in the "SOFT", respectively "STRICT" consensus predictors. The table reports the characteristics of descriptor types involved and the model performances (Balanced Accuracy) in cross-validation.

Map number	Fragments topology	Informational content	Min/max number of atoms	Atoms labeling	BA_{CV}
SOFT					
1	Sequences	Atoms only	2/8	Force Field (FF)	0.87
2	Sequence	Atoms only	2/7	FF and Formal Charge (FC)	0.87
3	Sequences	Atoms and bonds	2/6	FF	0.87
4	Sequences	Atoms and bonds	2/5	FF and FC	0.86
STRICT					
1	Sequences	Atoms only	2/7	FF and FC	0.73
2	Sequences	Atoms and bonds	2/4	FC	0.75
3	Sequences	Atoms and bonds	2/6	FF	0.74

Table 2
Description of eight maps (4 universal and 4 local), their descriptor types and cross-validated predictive propensity (Balanced Accuracy BA_{CV}) of the two-class classification landscapes colored according to the CHEMBL (universal maps) and SOFT (local maps) activity labels. Some typical fuzzy classification landscapes are illustrated in Fig. 3 and Fig. 4, respectively.

Map number	Fragments topology	Informational content	Min/max number of atoms	Atoms labeling	BA_{CV}
Universal maps					
1	Sequences	Atoms only	2/3	Force Field (FF) and Formal Charge (FC)	0.83
2	Atom-centered	Atoms and bonds	1/2	FF	0.82
3	Sequences	Atoms and bonds	2/4	Pharmacophore (Ph) and FC	0.80
4	Sequences	Atoms only	2/7	None	0.84
Local maps					
1	Sequences	Atoms only	2/4	FF	0.87
2	Atom centered	Atoms and bonds	1/3	FC	0.88
3	Normalized Atom centered	Atoms and bonds	1/3	FC	0.86
4	Sequences	Atoms and bonds	2/3	FF and FC	0.84

5.2. Is DSF-ΔT_m correlated with dose-response affinity measures?

Due to objective constraints, the experimental testing of the selected 3K compounds followed a protocol other than the ones used to characterize the affinity of the training set compounds from the public databases. This requires a better understanding of the degree of correlation. In the absence of strong correlation the training data used may not be relevant with respect to the measured property used to select hits. To this purpose, 39 BRD4-associated compounds in ChEMBL and are furthermore found among the compounds in stock at Enamine were also subjected to DSF measurement of ΔT_m at three different concentrations (10, 20 and 40 μM, respectively). 22 of these compounds were present in the CHEMBL training set and have reported IC_{50} or K_i values (the negative log of ChEMBL dose-response affinity value will further on generically be referred to as "pX"). However, none of them qualified for the "active" class as assigned by the automated procedure used to extract ChEMBL structure-activity class sets. For the remaining 17, the ChEMBL records could not be interpreted by the algorithm, so they were not included in the CHEMBL training set at all. Seven of the 22 were also present in the REAXYS set – five of which were assigned as inactives, and two as moderate actives by the human expert.

For twelve compounds, ChEMBL actually reports *both* IC_{50} and ΔT_m from measurements by their initial discoverers were reported. The magnitudes are weakly correlated, at $R^2 = 0.5$ (see Fig. 9 a).

If the analysis is extended to the herein measured ΔT_m values *versus* ChEMBL-reported pX data for 22 compounds (Fig. 9 b), the strongest correlation is obtained with the ΔT_m values at 10 μM concentration. Note, furthermore, that the average melting temperature shift for the 22 compounds with reported pX values was of $0.37 \pm 0.33°$, whereas the 17 ChEMBL compounds which had no associated pX values were all inactive with respect to ΔT_m: their average shift was of $0.08 \pm 0.15°$. They would have represented valuable true negatives but were not considered due to the intrinsic limitations of the ChEMBL activity series extraction protocol.

Last but not least, the correlation (Fig. 10 a) between ChEMBL ΔT_m and Enamine ΔT_m has been determined for 13 compounds and it turns out that the Enamine measure at 40 μM is the one best correlating with the ChEMBL data. The Enamine measure at 10 μM, the one that best correlated the ChEMBL-pX, is significantly less well related to ChEMBL ΔT_m values (R^2~0.4).

Eventually, for nine of the herein obtained hits, a FRET-based estimation of their IC_{50} values was experimentally undertaken. As seen in Fig. 10 b, these results are completely uncorrelated with the reported ΔT_m values.

The above discussion shows clearly that the exploitation of public databases obliges the user to face a wide spectrum of heterogeneous activity indices, which may or may not be "compatible" with the setups of the in-house experimental protocols for hit discovery. Clearly, the ChEMBL text mining protocol that assigned active/inactive labels to the BRD-associated compounds was successfully used for hundreds of other targets and returned modelable structure-activity sets. However, it specifically focused on molecules with reported dose-response activity measures (all while ignoring their exact nature – no distinction between K_i and IC_{50} values was made). It was not considering DSF-ΔT_m values. Activity classification based on such values is heavily target-specific, thus there is no simple threshold to be provided to a general data mining protocol. Text mining algorithms cannot cope with the subtleties of biological testing. The alternative of considering separate compound sets published by a same source using a same testing protocol does not solve the problem, but simply produces many disjoint small series, of no use in QSAR training. Also note that falling back to binary classification models – be it either by automated, algorithmic, or by expert hand-made choice of the activity threshold – is *per se* a source of information loss. The original dose-response activity scores from public sources were shown to be *per se* rather poorly correlated to the hit selection criterion DSF-ΔT_m. The fact that they were not used as such for model training, but first underwent conversion into a categorical variable has most likely had a negative impact on model

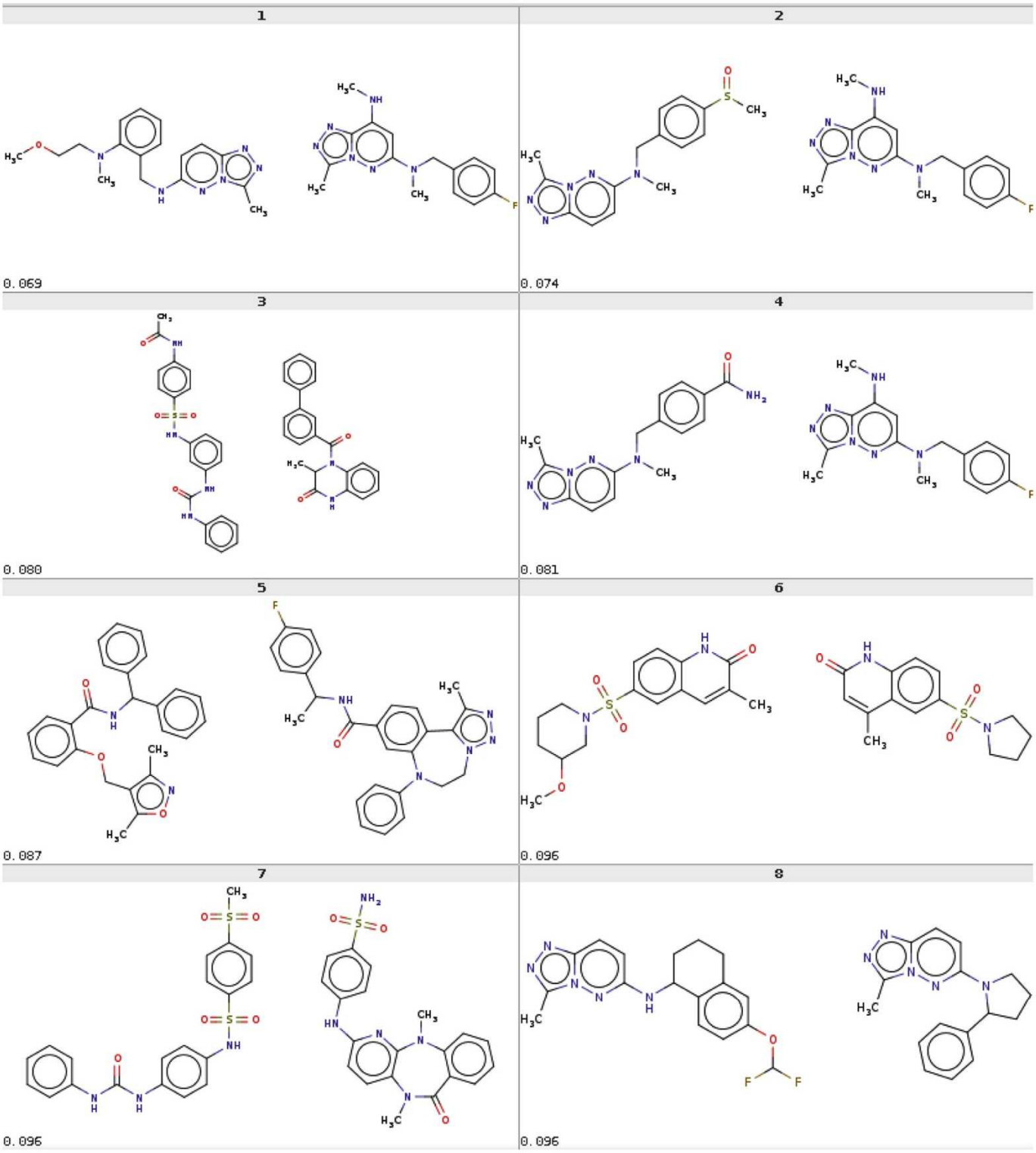


Fig. 8. The eight hits (left) closest to training set actives (right), with Soergel distance in the descriptor space of 4th DGTM given below each pair.

performance as well.

5.3. *A posteriori* analysis of the ability of individual models to prioritize discovered hits

Seventeen of the 29 hits have been ranked #1 by at least one of the dedicated GTM landscapes (notably DGTM2, but also DGTM1 and DGTM4) while two were ranked first by SVM. Note that "ranked #1" practically means that these compounds were given the maximum likelihood to be active, *ex aequo* with (often numerous) other candidates. None of the hits was seen to clearly outperform all the other 2M candidates of the Enamine library in terms of predicted likelihood of activity. More precisely, no model selected a single compound ranked #1. The other twelve hits

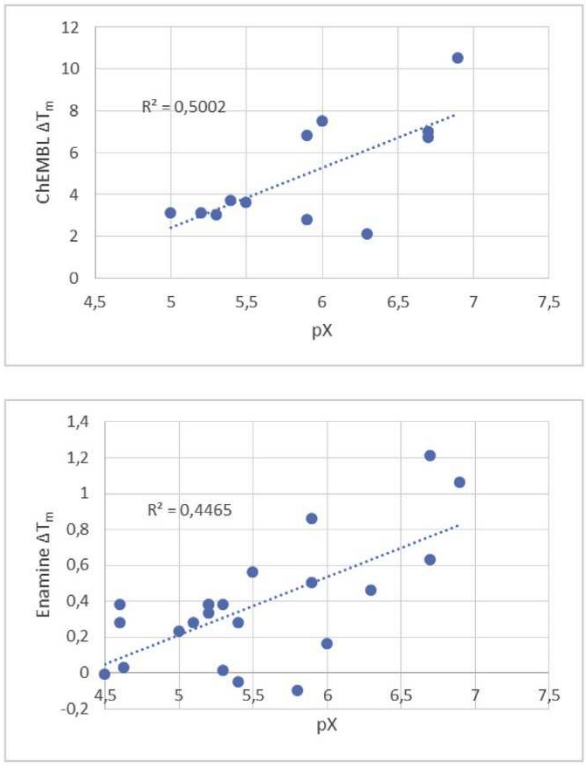


Fig. 9. a) Correlation between ChEMBL-reported ΔT_m and ChEMBL-reported dose-response-based affinity measure (pX) for 12 BRD4 inhibitors. b) Correlation between the ΔT_m values measured according to the current experimental screening protocol at Enamine, and ChEMBL-reported dose-response affinity measure (pX) for 22 BRD4 inhibitors from ChEMBL.

entered selection because of consensual ranking within top IM by several models.

In order to gain a better insight of the individual models that would have preferentially top ranked the 29 confirmed hits within the pool of selected 3K compounds, the plot (Fig. 11 a) was realized by monitoring, for each model, the minimal number of better-or-equally ranked compounds it would have had to select in order to "discover" *H* hits. Let *M(H)* denote the minimal size of the subset of top-ranked compounds by the model *M* that include $H = 1 \ldots 29$ of the confirmed hits. The percentage of hits in such selection, $H/M(H) \times 100$ has been plotted against *H*, for all considered QSAR, pharmacophore and docking models. For each *H* value ($H \geq 10$ shown in Fig. 11), there will be one "winning" model which managed to regroup *H* hits within the smallest *M(H)*, *i.e.* provided the best possible ranking for the *H* hits. A Pareto front of "dominating" configurations (hit number, hit percentage) can be drawn, all methods confounded. This Pareto front is mainly contributed by two methods: DGTM2-SOFT-DEC1-BNoff (Dedicated GTM2 landscape "colored" by the SOFT training set completed with decoy pool 1, without Bayesian normalization) and the SVM model trained on SOFT. Out of the tested 3K library, eleven of the 29 hits are found within the 106 top-ranked compounds DGTM2-SOFT-DEC1-BNoff, which represents the highest hit density (10.38%) that was achieved by any of the methods, all while retrieving a significant number of discovered hits. This represents the ten-fold of the hit rate of the current experiment, and the 26-fold of the one achieved in random screening [16]. The SVM model is an equivalent potent solution, favoring however the retrieval of more hits at a lesser hit rate (see Fig. 12).

It is instructive to construct such Pareto fronts not only for the entire battery of models, but also for specific subsets of models. The Pareto front of all the DGTM2-based models would consist of the best (hit number, hit rate) combinations scored by either of the landscapes based on the DGTM2-manifold, irrespective of training set, added decoys or Bayesian normalization strategy. The following plot illustrates the fronts constructed for the best performers amongst the QSAR models, which were able to reach or exceed 2% of hit rate (2-fold enrichment over the global hit density, and 5.2-fold enrichment over random VS).

These also include, in addition to already highlighted DGTM2 and SVM-based models, another dedicated GTM (DGTM3) and a Universal map (UGTM2). All classes of 2D-QSAR models have at least one representative that would have been in principle able to significantly enhance the hit rate beyond the achieved 1% – but not pharmacophore models, nor docking. In terms of docking scores these 29 compounds do not stand out in any way, compared to the rest of the tested 3K library – the mean of their docking scores perfectly matches the mean over the 3K set. Nevertheless, S4MPLE was proven to effortlessly discriminate between the BRD actives *versus* inactives and decoys of the STRICT set, with a ROC AUC of 0.77. When the compounds of intermediate potency are counted as active, in the SOFT training set, the S4MPLE ROC AUC value decreases to 0.66, *i.e.* remains well above random selection level. The clear separation of strong, medium actives and respectively inactives in terms of S4MPLE binding energy differences is visible in the histogram (Fig. 6).

Independently of this, S4MPLE has been successfully used in fragment-based drug design of novel BRD4 inhibitors [25]. However, the weak correlation between DSF-ΔT_m and actual IC_{50} values is certainly a significant reason for which the 29 selected hits are not "special" in terms of docking scores: the seven hits for which FRET-based IC_{50} values were actually measured are μM at best and all would have qualified as "inactive" according to the STRICT class assignment criteria, while some would have been qualified "inactive" even by the more lenient SOFT criteria. Clearly, the DSF-ΔT_m measurement protocol is useful for primary screening and is meant to select hits that are just potent enough to serve as a departure point in hit to lead optimization. It would not specifically single out very potent (but very rare) nM binders, which are unlikely to be discovered as such in primary screens. Or, discrimination by docking between weak binders and non-binders is notoriously difficult. Using the STRICT set for machine learning in general turned out to be a poor working hypothesis: the remarkable SVM model Pareto front is completely contributed by the SOFT set-trained SVM model, while the STRICT alternative provides no prioritization at all for the 29 hits. The same applies for the CHEMBL set, which also featured mostly sub-μM compounds as actives, its usage in both DGTM and UGTM landscape coloring would not have led to preferential selection of the 29 hits. By contrast to the highly diverse set of medium-potency inhibitors, the highly active BRD4 compounds in STRICT are fundamentally based on three key scaffolds: a seven-membered N heterocycle fused to phenyl, a pyrrole/imidazole ring fused to a saturated 5-membered lactam ring, and respectively benzimidazoles and derivatives featuring additional heterocyclic N atoms (Fig. 13 ab and c respectively). Other molecules (e.g., a macrocycle, spiro derivatives, or a diazo derivative) are basically singletons, *i.e.* too rare to allow machine-learning of their underlying structural patterns. Apparently, the dominating patterns are not well represented in the Enamine candidate collection submitted to this VS experiment.

Fig. 14 displays the specific behavior of DGTM2 landscapes based on the SOFT training set and without Bayesian normalization, as a function of the added pools of decoys. Interestingly, the maximal hit enrichment is seen to substantially depend on the randomly included decoys (here, 1% of ChEMBL compounds, excluding the BRD-tested structures). Decoy pools are thus rather

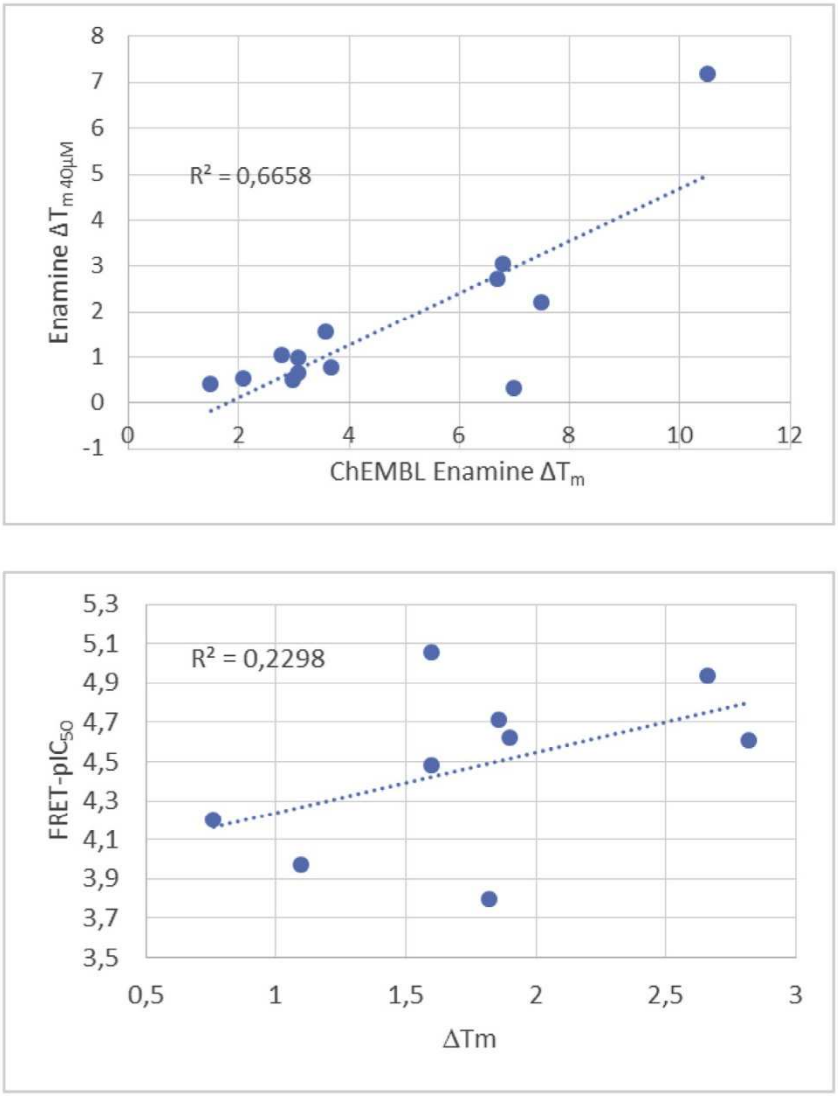


Fig. 10. a) Correlation between ChEMBL-reported and Enamine re-measured ΔT_m values, for ChEMBL BRD4-compounds. b) Correlation between FRET-based IC_{50} values estimated at Enamine for 7 of the newly discovered hits and their ΔT_m values in primary screening.

large compound collections, of sizes around 15K. However, the obtained class landscapes might behave significantly different with respect to the relative ranking of the 29 hits (the size of the compound set needed to encompass 11 hits actually doubles when the decoy pool 0 is used instead of pool 1). "Noise" from the decoy pools notwithstanding, these landscapes remain amongst the models that are significantly prioritizing the discovered hits.

No meaningful selection of 3K compounds could be achieved, making the addition of random ChEMBL decoys a necessity. In absence of decoys both GTM and SVM models displayed unexpectedly high propensities to rank the Enamine candidates as "active" – a tendency most marked with the SOFT training set, and also with the slightly more specific CHEMBL. The underlying reasons can be understood by inspecting fuzzy class landscapes – Fig. 15 shows the separation of BRD inactives (red) from actives (blue) in the (decoy-free) SOFT set, on UGTM1 and UGTM2, respectively. The (moderately) actives reported are – as expected from public databases compiling many sources – structurally quite diverse. Irrespective of the underlying descriptor space (ISIDA Force-field-colored atom sequence counts for UGTM1 *versus* force-field-colored circular atom counts for UGTM2), the SOFT "actives" cover a very large area of the SOFT-populated GTM landscapes (the majority of it, in case of UGTM1).

According to the perception of chemical diversity supported by UGTM1 descriptors, the SOFT subset of actives is actually more diverse than the inactive – and this perception of chemical diversity cannot be dismissed as irrelevant, because it supports robust separation of actives from inactives for >600 target-specific compound series, including BRD4 (for the SOFT set, the cross-validated balanced accuracy of separation is of 0.78).

Note that "visual" monitoring of diversity by the areas covered on the map, as illustrated above, may appear less rigorous than some quantitative measure – like the count of "clusters" that may be obtained by a classical algorithm. In practice, this is not the case – first, because the outcome of such a clustering algorithm may widely fluctuate in response of chosen descriptor set, dissimilarity metric, clustering algorithm and (algorithm-dependent) clustering thresholds, *etc.* Again, such hyperparameters were, in case of herein used GTMs, chosen as a result of a quantitative optimization of some predictive power propensity of the model. The density patterns seen on the map are thus representative of chemical diversity of a compound library and may even be quantitatively expressed –

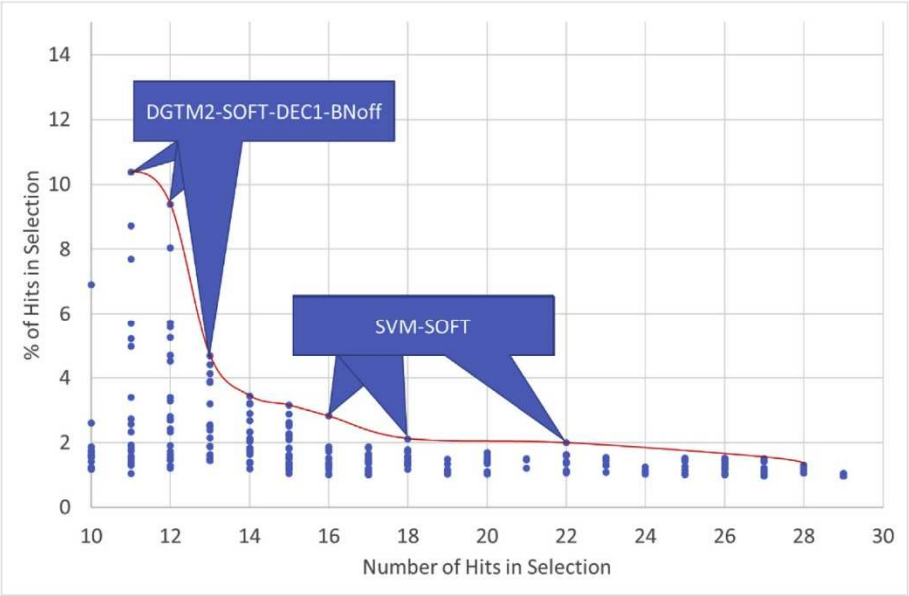


Fig. 11. The number of hits H (out of the 29 discovered) found within a minimal subset of M(H) compounds top-ranked by a model M, *versus* the percentage they represent within this subset. The red "Pareto front" regroups models returning the most hit-rich subsets containing H of the 29 hits. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

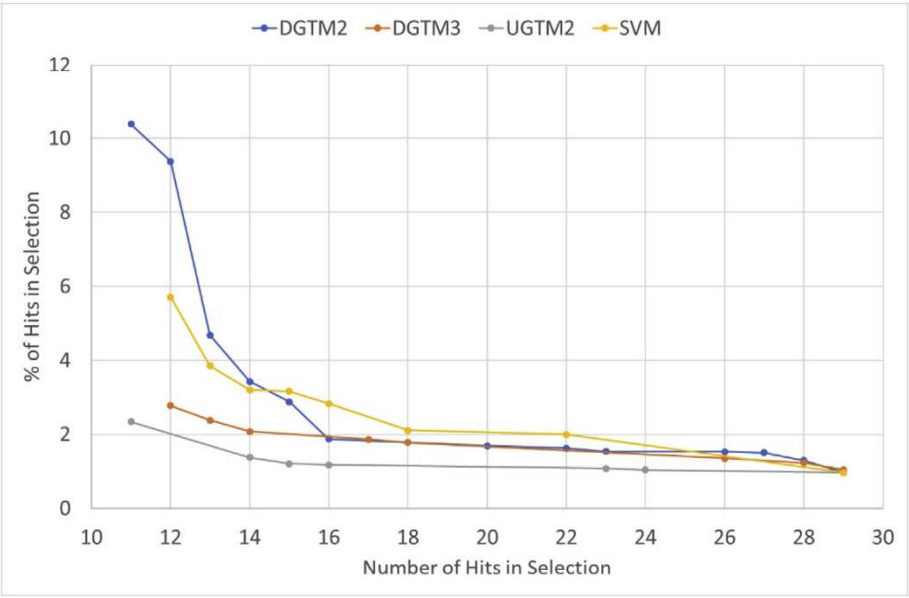


Fig. 12. The most significant (hit number, hit rate) Pareto fronts associated with the four QSAR approaches DGTM2, SVM, DGTM3, UGTM2.

by the entropy score [26]. Such values are however of relevance for large library comparison and will not be reported here.

The addition of decoys is indeed arguable – some of these decoys might actually be yet untested actives – but is nevertheless needed to "reclaim" some of the chemical space dominated by SOFT actives due to the fact that the training set as such fails to include such examples.

6. Conclusions

While the "Big Data" label may apply to public structure-activity databases as a whole, the specific target-related data needed for predictive model building in view of virtual screening of electronic compound databases is unfortunately rather sparse and heterogeneous. The VS study presented herein aimed at detecting novel BRD4 binders and relied on knowledge from public databases (ChEMBL, REAXYS) to establish a battery of predictive models used to virtually screen a collection of 2M compounds from Enamine. This industrial partner then experimentally screened a subset of 3K (2992) molecules selected by the VS procedure for their high likelihood to be active. Previous work at Enamine – random selection and screening, by the strictly identical Differential Scanning Fluorimetry protocol – of an equal-sized library drawn out of the same initial collection [16] – presented an excellent reference to estimate the benefit of VS in terms of hit rate enhancement. Twenty nine confirmed hits were detected alter experimental testing,

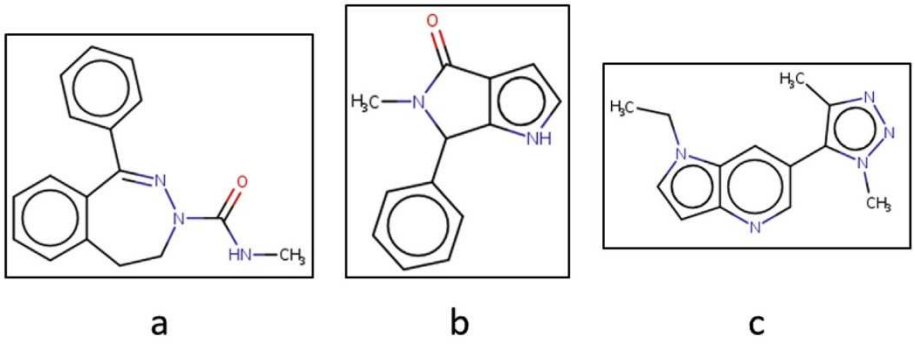


Fig. 13. Three key scaffolds of STRICT dataset.

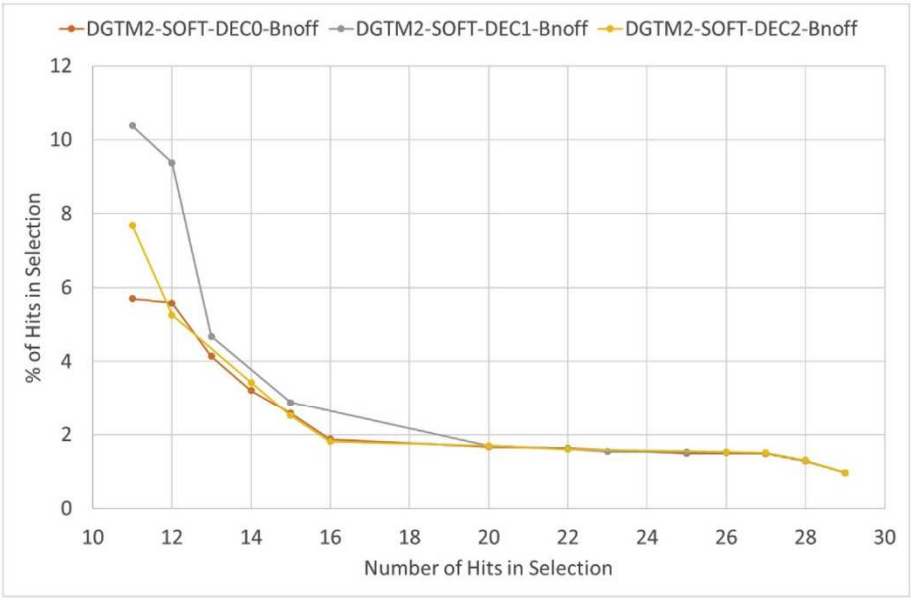


Fig. 14. (Hit number, Hit percentage) Pareto fronts of DTGTM2 landscapes as a function of the used random ChEMBL decoy set.

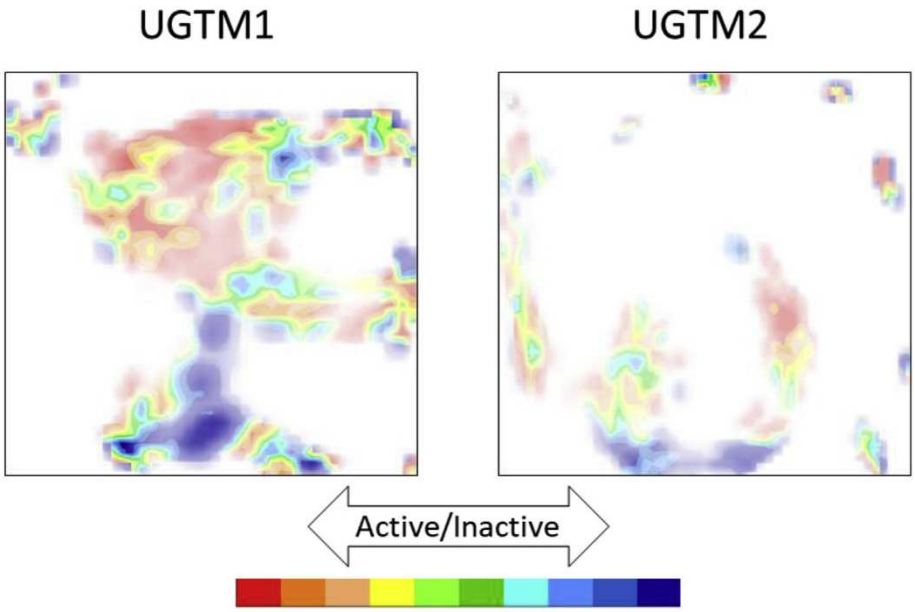


Fig. 15. Fuzzy classification landscapes showing the separation of actives and inactives as defined in the SOFT training set, on two Universal maps. These landscapes do not use Bayesian normalization – a red color means that SOFT actives are the absolute majority of residents in those areas. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

representing 1% of the 3K selected candidates. While this hit rate is a robust 2.6 times superior to the hit rate found in random screening under identical conditions, it is, on the absolute, rather disappointing for a "Big Data"-driven VS experiment, in which every single model (including docking) has been thoroughly (cross-)validated with respect to public BRD4 data. This prompted us to an in-depth investigation of the quality of training data, and its compatibility with the in-house hit selection criterion, DSF-ΔT_m. On one hand, it was shown that the heterogeneous public data cannot be fused into a single, rigorously defined training set. Specific dose-response-based activity values reported by authors contributing to the public databases cannot be merged and are only weakly correlated with DSF-ΔT_m assays. Therefore, active/inactive classification models were the only option, and attribution of the "active" label to public database ligands is a highly empirical, arguable undertaking. On one hand, ChEMBL compounds were extracted and classified by a simple text mining procedure developed for previous studies. By contrast, REAXYS-extracted compounds were classified according to IC_{50} thresholds and, as it was impossible to know beforehand what threshold will lead to the most predictive models, two distinct hypotheses were pursued in parallel, leading to the alternative STRICT and SOFT training sets.

Retrospectively, the SOFT training set produced models that provided the best rankings for the discovered hits. The highly active compounds exclusively labeled as "active" in the STRICT set mostly represent congeneric series based on common scaffolds completed with several singletons that cannot be exploited by machine learning. In the SOFT set, adding the moderately actives to the "active" class leads to the opposite scenario where "actives" seem to dominate a significant (even majority) of the training chemical space. In absence of decoys – random ChEMBL compounds that were never associated to the BRD receptors – SOFT-based models tend to overestimate the likelihood to be active. Adding 1%–5% of ChEMBL compounds as decoys counterbalances the artificial dominance of SOFT actives in the chemical space. Given the overall low to moderate correlation between dose-response activity values that are at the basis of training set definition and the DSF-ΔT_m criterion used to select hits, the decoy-enhanced SOFT training set worked better in conjunction with a screening method focused on discovery of moderate actives, *i.e.* typical primary hits. ChEMBL is a good source of decoys – within the intrinsic limitations of this approach. The random-drawn decoy subsets were seen to have a visible impact on the rankings returned by the GTM landscapes. STRICT and CHEMBL sets put more weight on the high potency of active examples, herewith limiting the number and diversity of actives to be "learned" by models.

VS selection was based on ranking each of the 2M candidates by their likelihood to be active, according to each model. These included SOFT, STRICT and CHEMBL-trained SVM, GTM (featuring both Universal and BRD4-Dedicated manifolds) and LigandScout pharmacophore models. A consensus scheme was employed to select 12K candidates – either top-ranked by at least one model, or ranked within the top N candidates by at least M models (several empirical N,M pairs were used – with more models M required as the required top rank N is relaxed). Docking with S4MPLE, which showed robust ROC AUC separation of training set actives, and independently served for successful fragment-based design of BRD inhibitors – was used to pick the 3K best-docked of the 12K selected candidates. These were experimentally screened by the Enamine partner, with already mentioned results.

As a general conclusion, this study emphasizes that public structure-activity databases are nowadays key players in drug discovery. Their limits are set by the state-of-the-art knowledge harvested so far by published studies, and these limits can be very stringent. Data heterogeneity makes it extremely difficult to exploit in rational drug design. Furthermore, published activity values may not be easily comparable or may not be correlated with values from other assays. In spite of this, a robust 2.6-fold increase of the hit rate with respect to an equivalent, random screening campaign showed that machine learning is able to enrich active compounds in selection sets.

Acknowledgment

Iuri Casciuc thanks the Région Grand Est for a PhD fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmech.2019.01.010.

References

[1] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[2] C.M. Bishop, M. Svensén, C.K.I. Williams, GTM, The generative topographic mapping, Neural Comput. 10 (1998) 215–234.
[3] T. Kohonen, The self-organizing map, Proc. IEEE 78 (1990) 1464–1480.
[4] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, A. Varnek, Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison, Mol. Inform. 31 (2012) 301–312.
[5] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, A. Varnek, GTM-based QSAR models and their applicability domains, Mol. Inform. 34 (2015) 348–356.
[6] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds, J. Comput. Aided Mol. Des. 29 (2015) 1087–1108.
[7] P. Sidorov, E. Davioud-Charvet, G. Marcou, D. Horvath, A. Varnek, AntiMalarial mode of action (AMMA) database: data selection, verification and chemical space analysis, Mol. Inform. 37 (2018) 1800021.
[8] Reaxys Database, 2017. https://www.reaxys.com.
[9] 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, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2011) D1100–D1107.
[10] J.W. Tamkun, R. Deuring, M.P. Scott, M. Kissinger, A.M. Pattatucci, T.C. Kaufman, J.A. Kennison, brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2SWI2, Cell 68 (1992) 561–572.
[11] S.-Y. Wu, C.-M. Chiang, The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation, J. Biol. Chem. 282 (2007) 13141–13145.
[12] V. Brès, S.M. Yoh, K.A. Jones, The multi-tasking P-TEFb complex, Curr. Opin. Cell Biol. 20 (2008) 334–340.
[13] J. Morinière, S. Rousseaux, U. Steuerwald, M. Soler-López, S. Curtet, A.-L. Vitte, J. Govin, J. Gaucher, K. Sadoul, D.J. Hart, others, Cooperative binding of two acetylation marks on a histone tail by a single bromodomain, Nature 461 (2009) 664.
[14] L. Hoffer, D. Horvath, S4MPLE–Sampler for Multiple Protein–Ligand Entities: simultaneous docking of several entities, J. Chem. Inf. Model. 53 (2012) 88–102.
[15] L. Hoffer, C. Chira, G. Marcou, A. Varnek, D. Horvath, S4MPLE-sampler for multiple protein-ligand entities: methodology and rigid-site docking benchmarking, Molecules 20 (2015) 8997–9028.
[16] P. Borysko, Y.S. Moroz, O. V Vasylchenko, V. V Hurmach, A. Starodubtseva, N. Stefanishena, K. Nesteruk, S. Zozulya, I.S. Kondratov, O.O. Grygorenko, Straightforward hit identification approach in fragment-based discovery of bromodomain-containing protein 4 (BRD4) inhibitors, Bioorg. Med. Chem. 26 (2018) 3399–3405.
[17] Standardizer ChemAxon, (2012) version 5.12.
[18] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, ISIDA property-labelled fragment descriptors, Mol. Inform. 29 (2010) 855–868.
[19] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V Tetko, G. Marcou, ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors, Curr. Comput. Aided Drug Des. 4 (2008) 191.
[20] D. Horvath, J.B. Brown, G. Marcou, A. Varnek, An evolutionary optimizer of libsvm models, Challenges 5 (2014) 450–472.
[21] Gerhard Wolber and Inte:Ligand GmbH, LigandScout 4.1, 2017. http://www.inteligand.com/ligandscout/.
[22] M. Zhenin, M.S. Bahia, G. Marcou, A. Varnek, H. Senderowitz, D. Horvath, Rescoring of docking poses under Occam's Razor: are there simpler solutions? J. Comput. Aided Mol. Des. (2018) 1–12.
[23] H.2R7 Chemical Computing Group Inc, 1010 Sherbrooke St. West, Suite #910,

vol. 08, Molecular Operating Environment (MOE), Montreal, QC, Canada, 2016 (2016).
[24] T.U. Consortium, UniProt: the universal protein knowledge base, Nucleic Acids Res. 45 (2017) D158–D169, https://doi.org/10.1093/nar/gkw1099.
[25] L. Hoffer, C. Muller, P. Roche, X. Morelli, Chemistry-driven hit-to-lead optimization guided by structure-based approaches, Mol. Inform. (2018), https://doi.org/10.1002/minf.201800059.
[26] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, A. Varnek, Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge, J. Chem. Inf. Model. 55 (2014) 84–94.

5.5 Conclusions

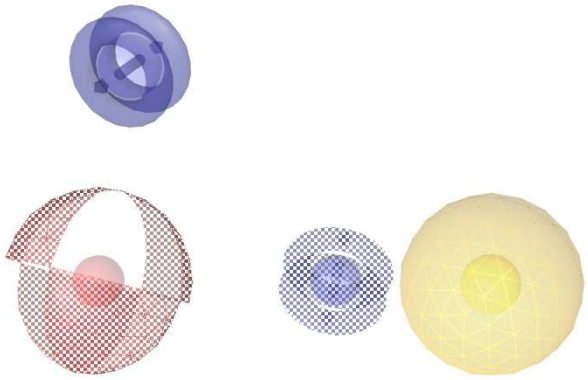
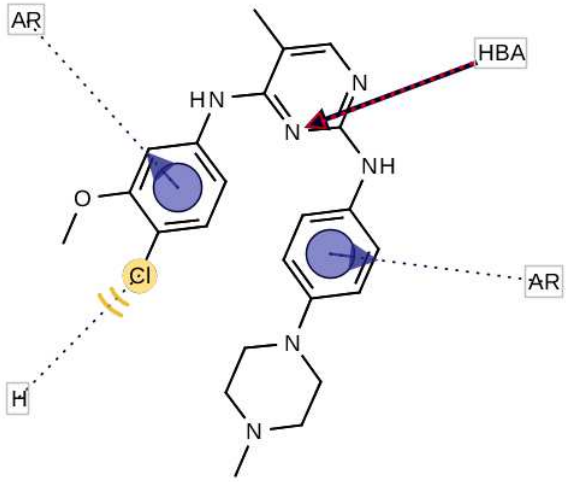
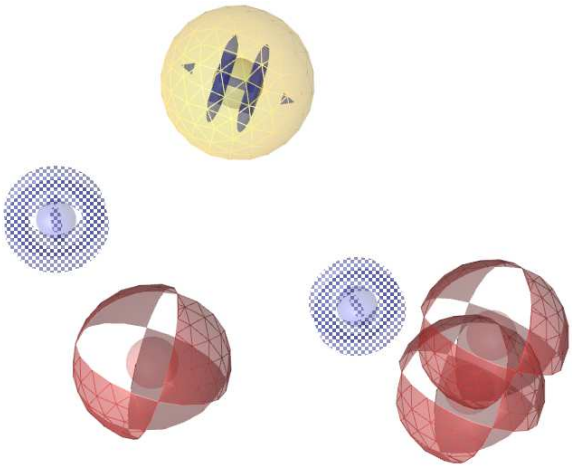
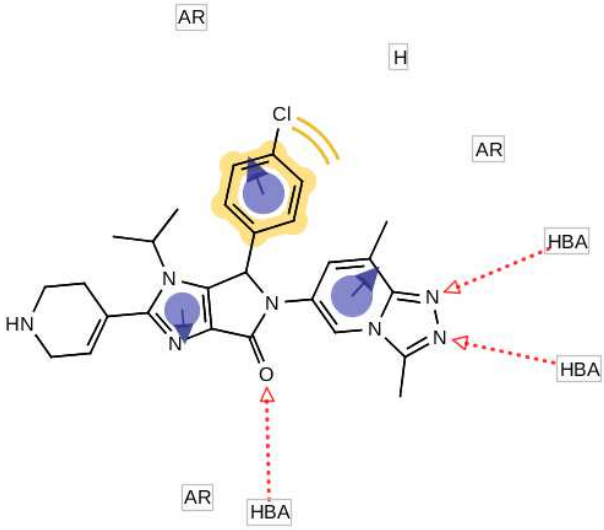
A collection of 2M compounds have been subjected to a virtual screening funnel involving classification SVM and GTM models as well as ligand-based pharmacophores trained on publicly-accessible SAR data on BRD4 IC_{50}/pK_i from Reaxis and ChEMBL. Each model has provided a ranked list of 2M candidates according to their likelihood to be active. The consensus application of these approaches has been used to obtain a subset of 12k compounds that have been submitted to a docking procedure. The docking has been used for a further selection of the "best" 3k compounds out of the previously selected pool of candidates. These 3k compounds have been experimentally screened by the Enamine partner using the Thermal Shift Assay method.

Twenty-nine confirmed hits had been detected, which represents 1% of the 3k selected candidates. While the obtained hit rate is still 2.6 times better than the hit rate found in random screening under identical conditions, it is still objectively low. An in-depth analysis of the quality of the used data for the models' training has been performed, as well as the correlation between IC_{50}/pK_i and DSF-ΔT_m. First of all, it has been shown that public data from different sources cannot be fused into a single and rigorously defined dataset adapted for QSAR modeling. Moreover, it has been shown that the dose-response activity values reported in publicly available databases are weekly correlated with DSF-ΔT_m. Last but not least, the retrospective hit analysis has shown that GTM models have outperformed SVM and ligand-based pharmacophores in terms of hits identification.

5.6 Supporting information

Supporting Information includes information about (*i*) developed pharmacophore models (**Table 5-1**) and (*ii*) structures and activities of 29 hits validated experimentally (**Table 5-2**)

Table 5-1: Used pharmacophore models depictions and the associated binding mode 2D maps.

Model	Model depiction	Associated binding mode 2D map
1		
2		

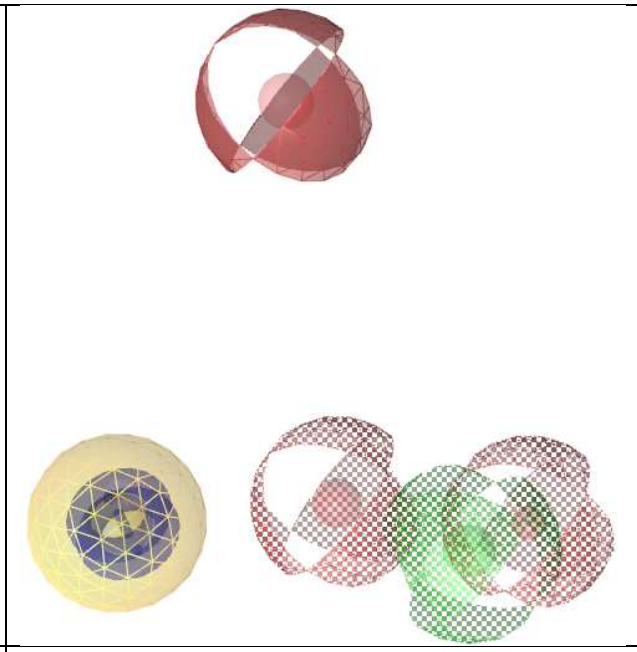
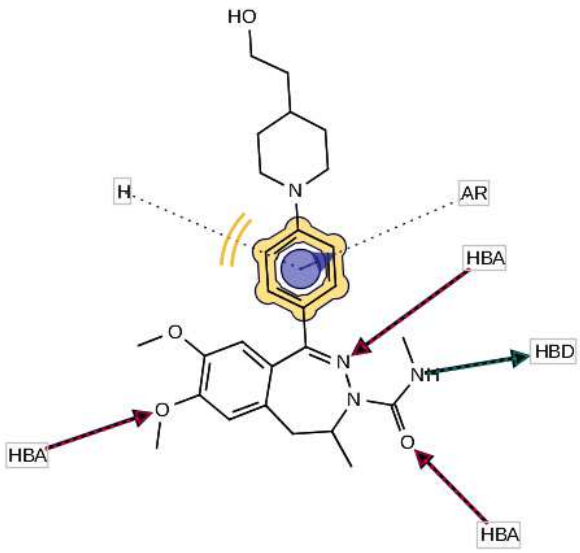
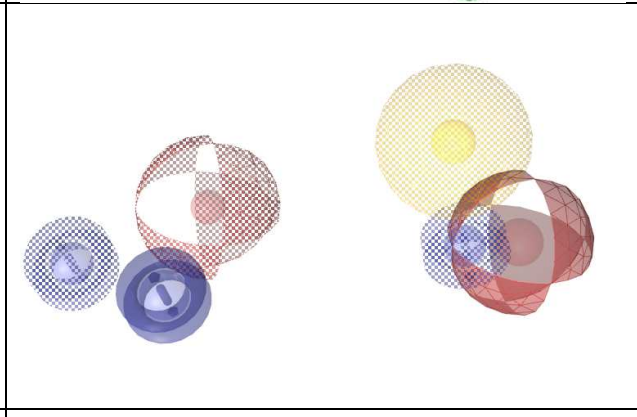
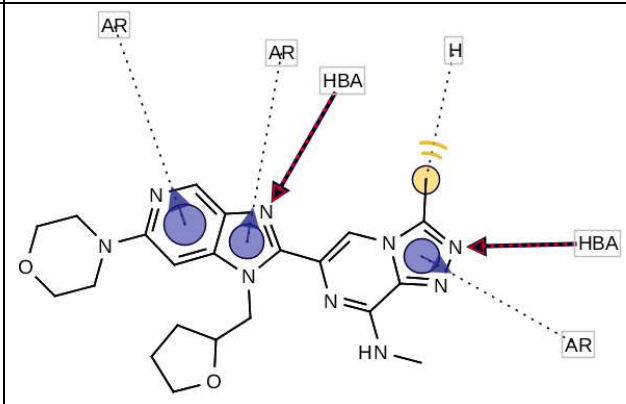
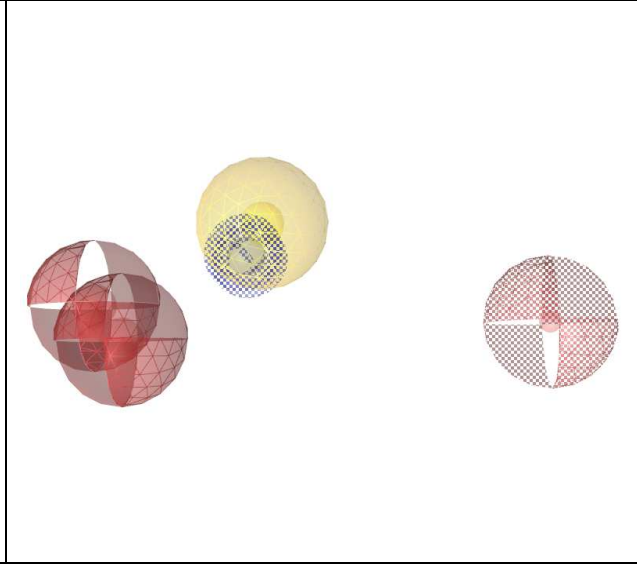
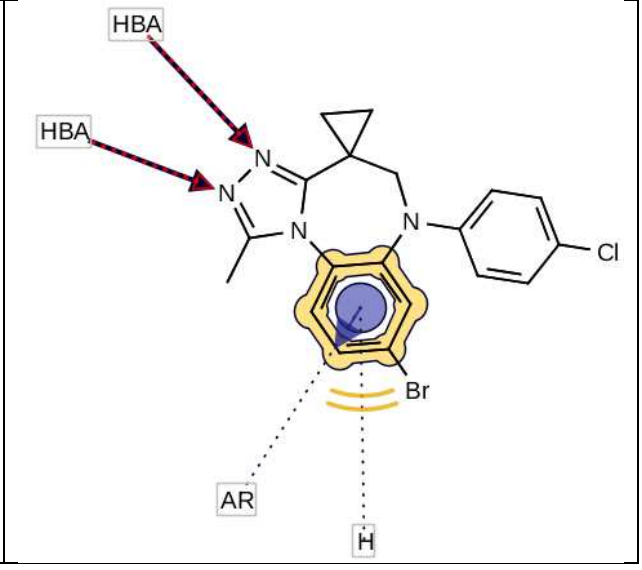
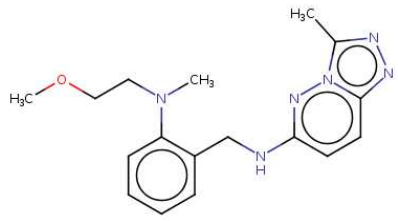
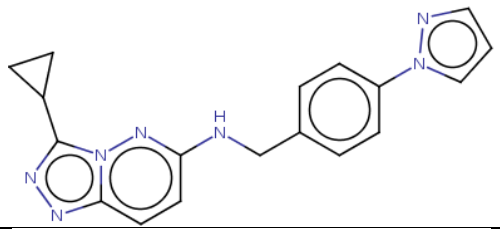
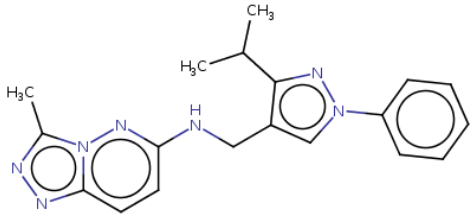
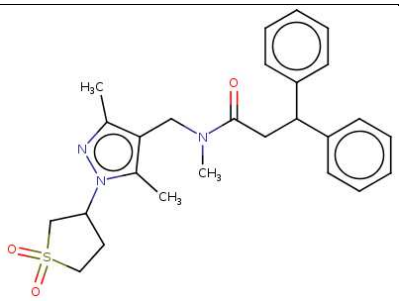
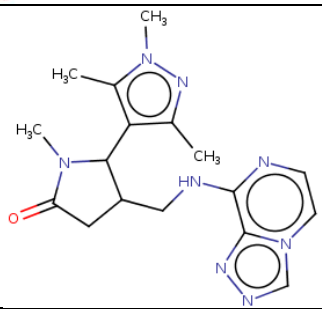
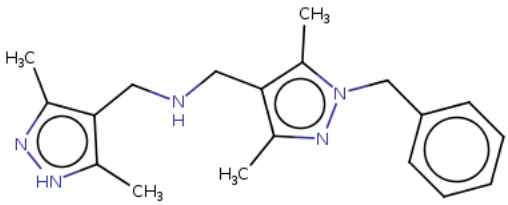
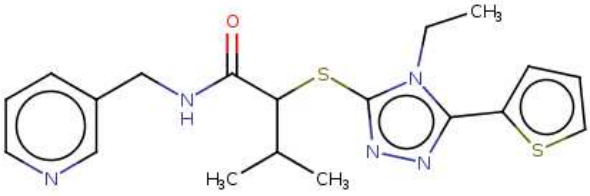
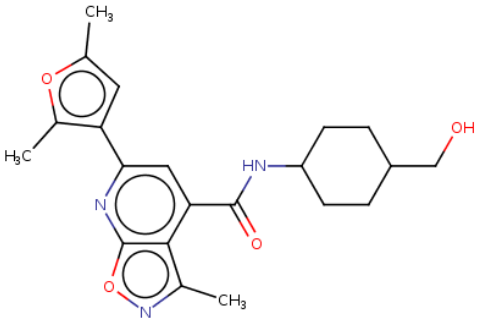
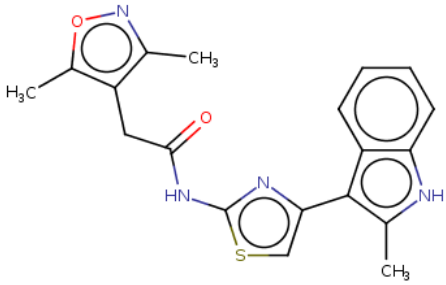
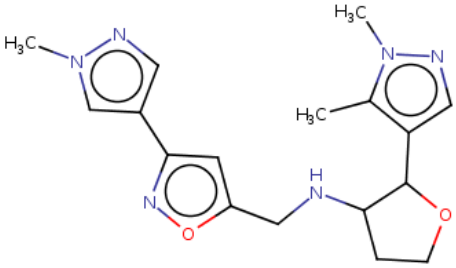
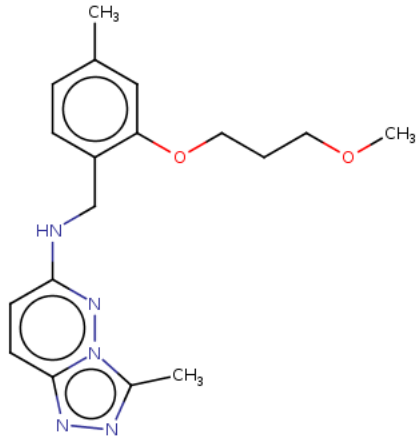
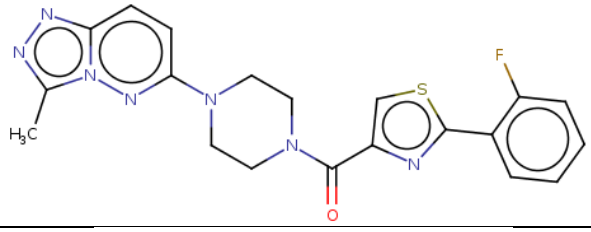
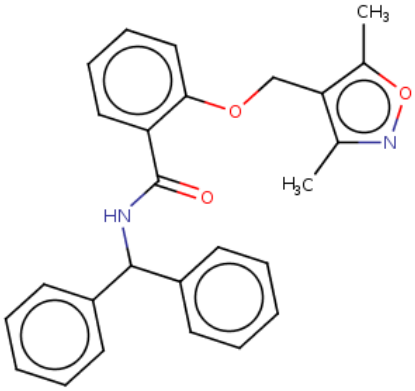
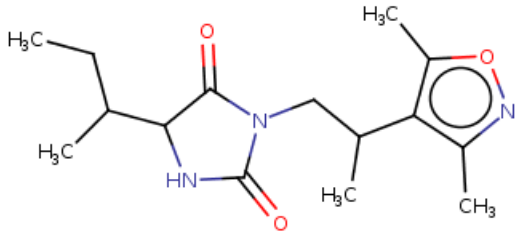
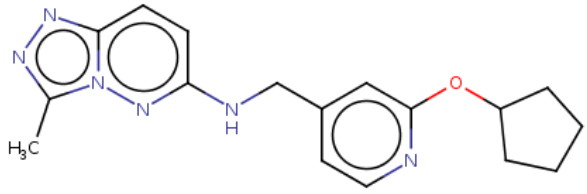
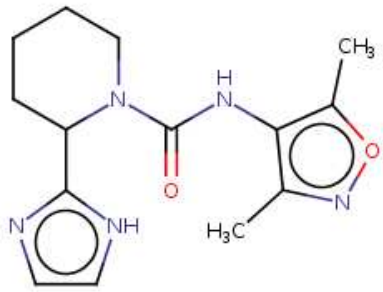
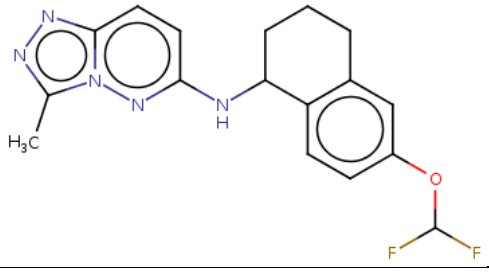
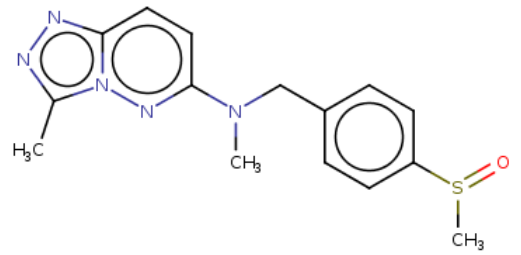
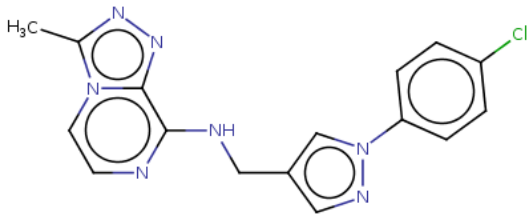
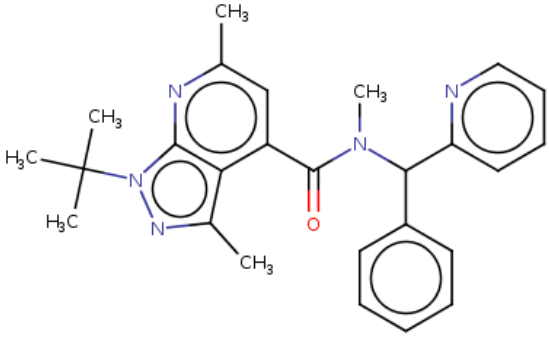
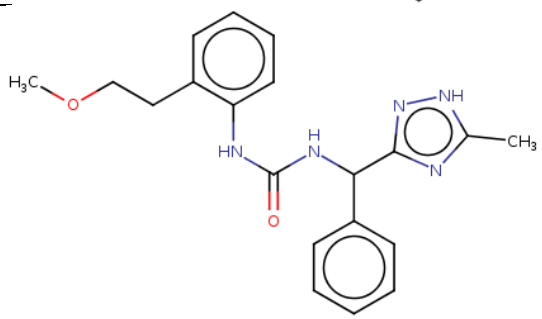
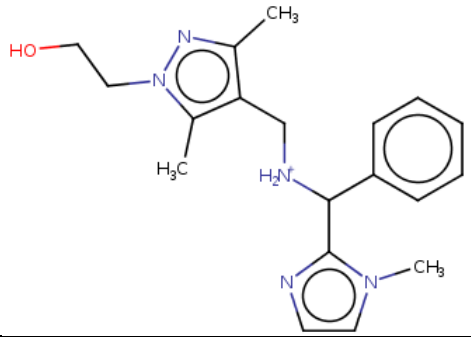
3		 HO N H AR HBA O N NH HBD N O HBA O HBA
4		 AR AR HBA H N N O N N N N N HBA HN AR O
5		 HBA HBA N N N N Cl Br AR H

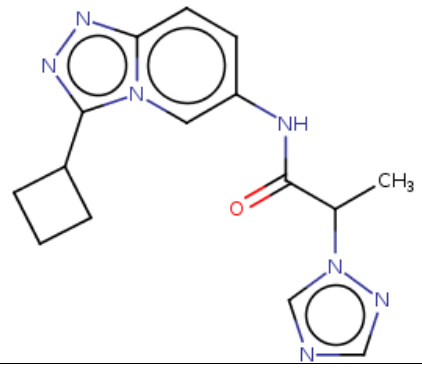
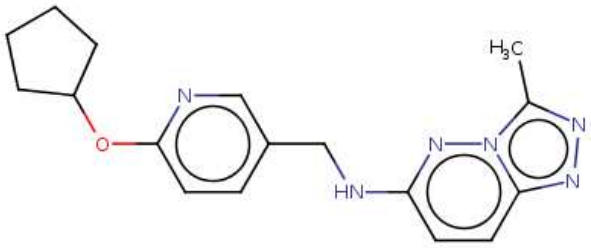
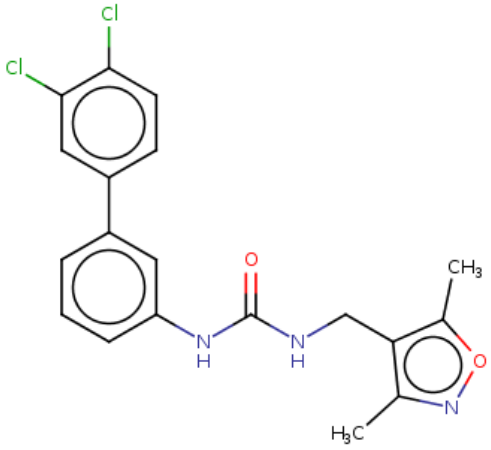
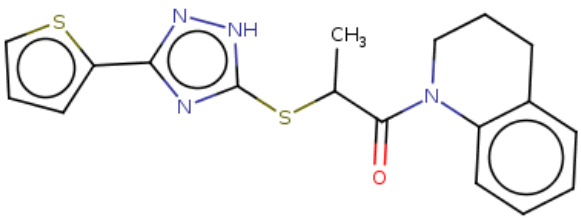
Table 5-2: 29 confirmed hit structures, experimental values of ΔTm and IC_{50} (if available)

Structure	ΔTm(pos)	IC_{50} (μM)
	1.1	107.41
	0.76	63.28
	2.66	11.72
	1.82	160.03
	1.90	23.87
	1.60	33.16

	1.60	8.81
	2.82	24.91
	1.86	19.53
	2.02	NA
	1.56	NA

	1.72	NA
	0.93	NA
	2.12	NA
	1.69	NA
	1.76	NA
	1.70	NA

	1.80	NA
	1.86	NA
	1.49	NA
	1.26	NA
	1.09	NA

	1.49	NA
	1.89	NA
	1.56	NA
	1.21	NA
	1.78	NA

	1.08	NA
	1.36	NA

6 *In silico* speciation assessment of Dynamic Combinatorial Libraries of imines

The reversible combination of molecular building blocks via covalent or non-covalent bonds is a cornerstone of the Dynamic Combinatorial Chemistry [7]. The reversible nature of the reactions between the building blocks leads to the fact that their thermodynamic stability dictates product distribution in the mixture. Once a Dynamic Combinatorial Library (DCL) is exposed to an external effector (like a biological target), it might happen that the latter selectively binds to one (or several) members of the DCL; thus, the equilibrium is shifted according to Le Chatelier principle, which leads to global change of the solution composition. The nature of the effector is not limited to the biological target. It could be physical (like a change of temperature [84]), or chemical (introduction of an "alien" species to the solution, like a metal ion [85], a protein/enzyme [8]; change of solvent/pH [86]).

For example, in a hypothetical library, only containing two pairs of building blocks, the outcome is easily predictable –the solution composition will reflect the relative stability of the reactions' products. Once an effector is added, the library composition changes as a function of products affinity to the effector. In the simplest case, only one reaction product selective binds the given effector, leading to the shift of all equilibria in DCL in favor of the formation of the complexed species (**Figure 6.1**).

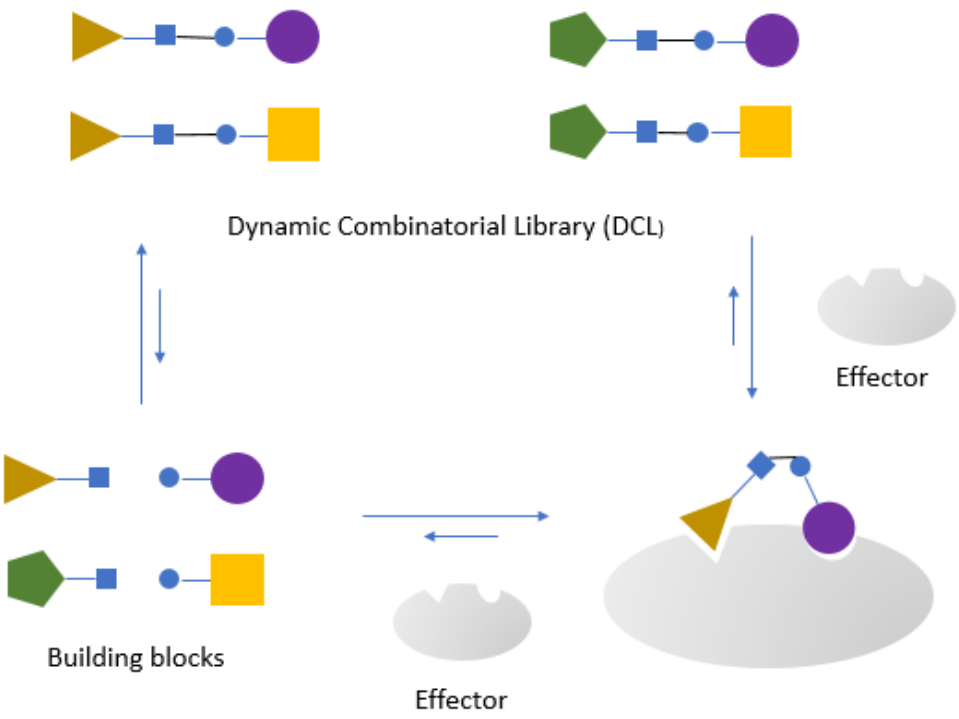


Figure 6.1: Principle of work of a hypothetical DCL made of 2 pairs of building blocks in the presence of effector. Small squares and circles represent complementary chemical functions.

A typical example of DCL is imine formation from aldehyde and amine (**Figure 6.2**). Since this reaction is reversible, mixing *m* aldehydes with *n* amines resulting in the formation of *mxn* imines with different combinations of R_1 and R_2.

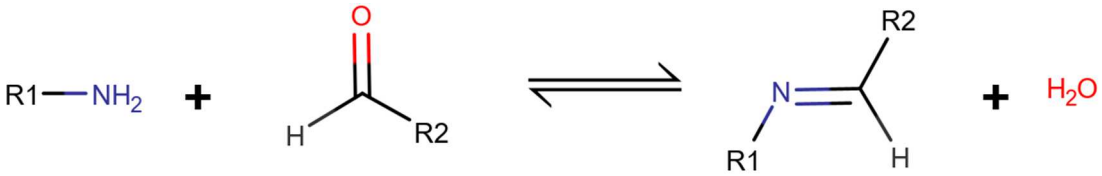


Figure 6.2: Reaction of imine formation from amine and aldehyde.

In DCL formed by two aldehydes (A1 and A2) and two amines (B1 and B2), four products are expected: A1B1, A2B1, A1B2 and A2B2. As one may see from **Figure 6.3**, two pairs of "opposite products," A1B1/A2B2 and A2B1/A1B2, are in an *agonistic relationship* (green lines), whereas "adjacent products" are in an *antagonistic relationship* (red lines). The products A1B1/A2B2 in the agonistic relationship favor the formation of

each other, more A1B1 is formed, less non-reacted A1 and B1 building blocks remain in solution, therefore the concentration of A2B2 increases. When it comes to the products that are in an antagonistic relationship, it is the opposite. More A1B1 is formed, less A1B2 and A2B1 are present in the solution because of the lack of needed building blocks.

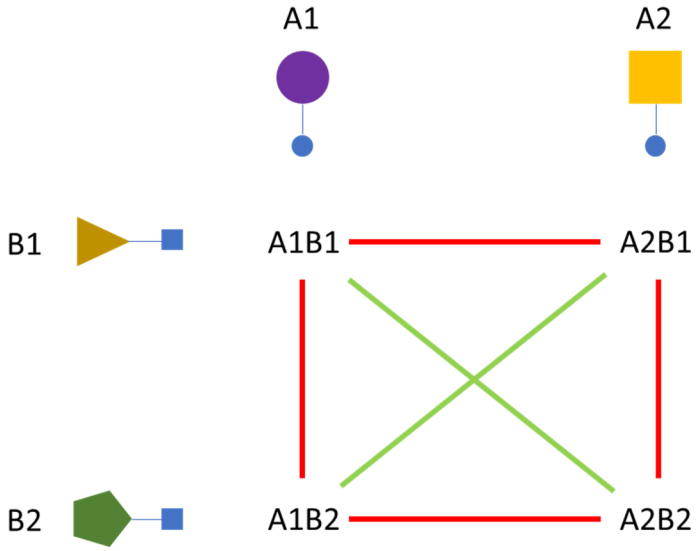


Figure 6.3: A DCL composed of 2 pairs of building blocks. The pairs in an *agonistic relationship* are shown with green lines. The compounds being in an *antagonistic relationship* are shown with red lines.

Although the number of known reversible reactions is relatively high, not all of them can be carried out in aqueous media, which prevents their usage in protein-templated DCL. Moreover, the functional groups of building blocks are not supposed to react with the target themself. The non-exhaustive list of reversible and biocompatible reactions includes imine, hydrazine and acylhydrazone formation; alkene cross-metathesis; disulfide, thioether and boronate ester formation. Imine formation reaction was the first reaction applied to a DCL in the presence of bovine carbonic anhydrase II [8] as a receptor.

Usually, when one wants to use a DCL for the identification of a “best binder,” the reaction products are expected to be in almost equal concentrations, since in the case of the biased library, where one or a few constituents would be highly favored, the preferred interaction of a minor constituent with the target may not be strong enough to overturn the

equilibrium situation. In some cases, an experienced chemist could guesstimate what reactants should be chosen for the given DCL; however, this approach might be inefficient. The usage of the software which can estimate *speciation*, i.e., equilibrium concentrations of all species in solution. This requires a knowledge of equilibrium constants, which can be problematic because of the thermodynamic data availability.

Chemoinformatics models predicting equilibrium constants could be a reasonable solution to assess speciation for any DCL with or without effector. In this work, DCL based on the reaction of imine formation is modeled with and without an external chemical effector. As a tribute to the seminal work [4], human carbonic anhydrase II (CA II) was chosen as an effector to model the adaptive behavior of the imine-containing DCL. The project workflow (**Figure 6.4**) involves several steps. At the first step, a predictive model for the logarithm of imine formation constant (logK) as a function of the structure is built using experimental data measured in chloroform solution. In the second step, a model for the logarithm of the binding constant (pK_i) of organic molecules to human CA II should be prepared on experimental data extracted from the ChEMBL database. Since the latter were measured in water, one needs to scale the imine formation constant determined in chloroform to those in water solution. Once both types of models are available, predicted stability and binding constants obtained could be used as input to a speciation software. One of the important tasks was to select a representative training set for the imine formation constants modeling. This work is described in section 6.1, followed by the modeling part described in section 6.2.

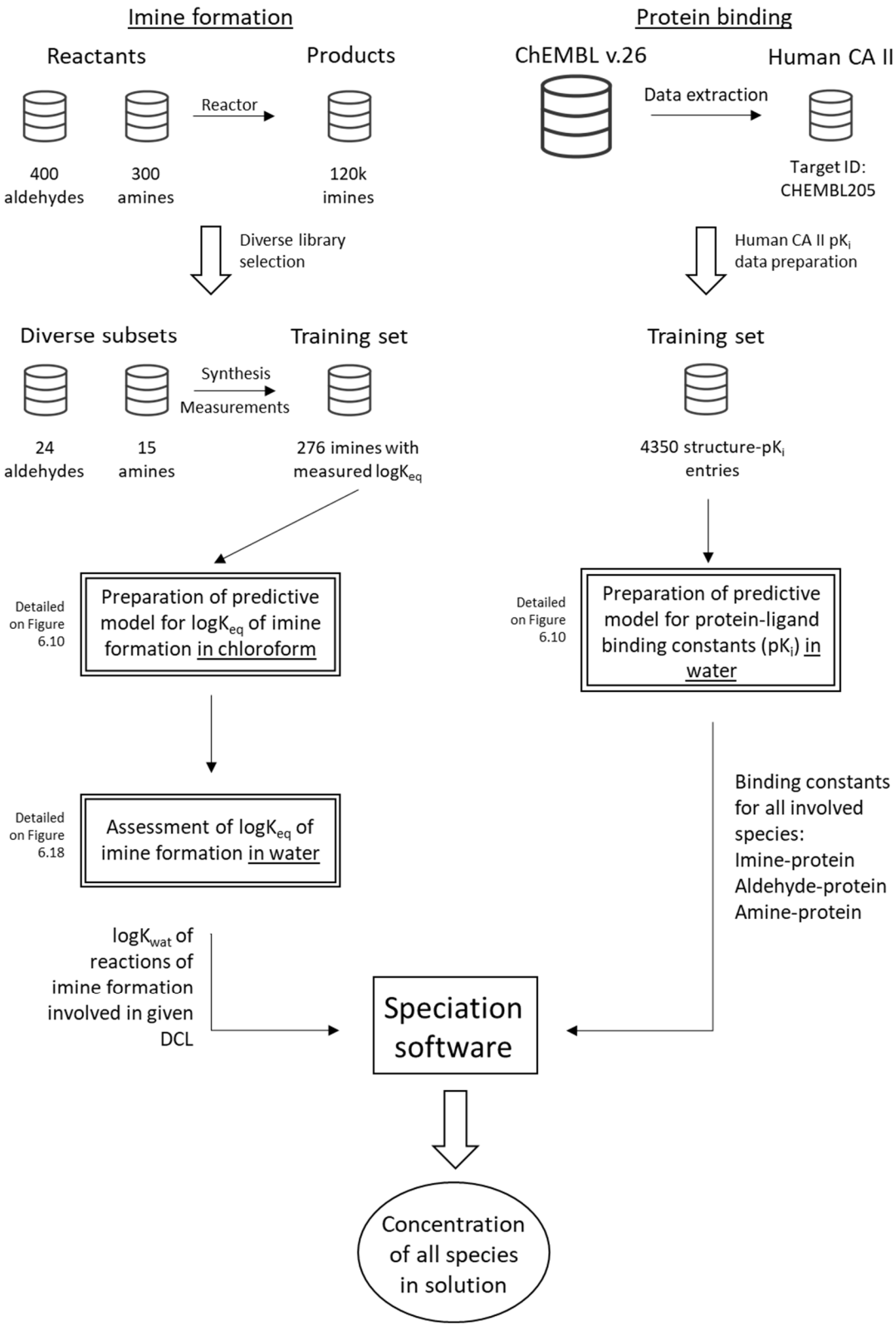


Figure 6.4: Workflow for *in silico* DCL speciation.

Reported case studies concerned empirically designed DCLs involving a relatively small number of reactants. Our goal is to develop a theoretical (*in silico*) approach allowing one to predict the species concentration for DCL of any size in the presence or in the absence of effector. The workflow of such *in silico* speciation of DCL is given on **Figure 6.4**. It involves two essential steps: (*i)* selection of a diverse library of imines which serves as a training set in model building, and (*ii)* preparation of statistical models able to predict equilibrium constants of imines formation and binding constants of protein-ligand complexes in the solution used, in turn, as an input in a speciation software. These two steps of the project are described in two separate sections below.

6.1 Application of GTM for diverse library selection

6.1.1 Introduction

The identification of representative and diverse subsets in large libraries of compounds is crucial to medicinal chemistry since a diverse subset of compounds provides more chances to contain a compound with the required type of activity during screening tests. When one or several compounds from a *diverse* subset have been proven to show a certain level of activity, then a *focused* library of compounds is being selected. Focused library design implies the selection of similar compounds to the known "hits". Traditionally *diverse* subsets have been created by having a medicinal chemist select compounds manually based on a series of 2D structures [87]. Although this approach could be successfully used on relatively small datasets, it would become extremely difficult (or probably impossible) when the datasets have thousands and hundreds of thousands of compounds. Moreover, the level of "representativity and diversity" of the manually selected subset could vary from chemist to chemist. The selection of a diverse library from a large dataset of compounds can be approached in several ways: using clustering, using dissimilarity-based methods or using cell-based methods. Clustering implies that the compounds are grouped according to some similarity measure. Then from each cluster, a

random compound is selected, thus leading to a library containing a representative of each cluster. Dissimilarity-based methods are relying on the calculation of pairwise distances for every compound of the initial dataset, which is followed by a one-by-one selection of compounds in the diverse library according to a pre-defined rule. The main feature of cell-based selection methods is that they do not require a pairwise calculation of the distances for all the compounds. To provide an efficient application of a cell-based approach, the dimensionality of chemical space defined by the molecular descriptors is usually reduced to a 2D map; the map is divided into zones (cells), and from each cell, a compound is extracted.

Although the diverse libraries are usually used in medicinal chemistry, one can use a diverse subset of compounds as a training set in structure-activity modeling. Diversity is not a fundamental, objective property of a compound collection, but it is a rather practical, problem-dependent and therefore vaguely defined concept. Usually, compound diversity is directly related to their dissimilarity, hence to quantitatively measure the diversity of selected compounds is the most straightforward to encode into chemoinformatics software. If molecules are objects in the descriptor space, then dissimilarity is directly related to the distances separating them. Depending on the descriptor space and the therein employed metrics, the distances may vary. At first sight, the above seems like a rigorous mathematical basis for compound selection: the degree of dissimilarity is typically illustrated by the distribution histogram of pairwise distances between the compounds.

Despite the large number of different applications of GTM, it has never been used for diverse compounds library design. Since GTM is producing a 2D map, this map could be used in a cell-based approach for diverse library selection. In this context, GTM could be compared to its non-probabilistic predecessor – Kohonen Self Organizing Map (SOM). It has been reported that SOM is a useful tool for *focused* library design and combinatorial libraries [88, 89]. Nettekoven and Schneider [88] used SOM for the focused-library combinatorial design of selective purinergic receptor (A2A) antagonists. Selzer and Ertl [90]

used SOM to select a representative subset of 5000 compounds from a collection of combinatorial libraries containing nearly 100000 compounds in total. Unfortunately, in that work, the method performance in diverse library selection has been reported in a qualitative way as "low/medium/high" diversity of the selected library, which does not allow quantitative comparison of the results obtained with GTM to the results obtained by SOM in a similar task. In this section, the performance of GTM in diverse library selection has been compared to a classical algorithm of diverse library selection – MaxMin [9, 91].

6.1.2 Data and methods

A substructural search of *primary amines* and *aromatic aldehydes* has been done on SciFinder; the results have been sorted by the number of citations, putting the most cited compounds on the top of the list. The reactants have been selected according to the following criteria:

- Every compound should contain one single amine/aldehyde group, thus leading to only one possible reaction product.
- In the case of amino-acids, the COOH group has been changed into COOMe.
- Thiol containing compounds have been rejected.
- Reactants having a molecular weight > 400 g/mol have not been taken into account.
- Long-chained acetals -C(OR1)(OR2) have been changed to -C(OMe)(OMe).

Two datasets of reactants have been obtained: the *primary amines* dataset containing 300 molecules and 400 compounds dataset of *aromatic aldehydes*. Interaction between all the selected aldehydes and amines could give 120 000 imines. Since MaxMin is a very time-consuming method of $O(n^2N)$ complexity (n and N are the numbers of objects in the initial set and diverse subset, respectively), we decided to perform some methodological tests on a smaller set containing 42658 imines resulted from reactions of 154 amines and 277

aldehydes. Five different ISIDA fragmentations have been used (see section 6.1.2.2 for details):

Table 6-1: Used fragmentation schemes of ISIDA descriptors and their meaning.

Fragmentation scheme	Meaning
IA-FF-FC-2-3	Sequences of atoms colored by force field properties and formal charge with a length from 2 to 3
IAB-FC-1-6	Sequences of atoms and bonds with a formal charge on atoms, with a length from 1 to 6
IIRAB-FF-1-3	Circular fragments of atoms and bonds with restricted length from 1 to 3 atoms, colored by the force field
IAB-FC-2-4	Sequences of atoms and bonds with a formal charge on atoms, with a length from 2 to 4
IIRA-FF-FC-1-2	Circular fragments of atoms with restricted length from 1 to 2 colored by the force field and containing information about the formal charge

6.1.2.1 Dissimilarity-based methods

Since the task is to select the most dissimilar compounds, one should select the compounds that are most different from the ones already selected in the formed set. As the metric of dissimilarity, the Soergel distance [92, 93] has been taken. The following formula defines Soergel distance:

$$S_{AB} = 1 - \frac{\sum_{j=1}^{n} x_{jA} x_{jB}}{\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jA} x_{jB}}$$

Where x_{jA} and x_{jB} represent descriptors vectors of compounds A and B, respectively. It has been shown that Soergel distance can be used as a metric only when the values of the descriptors are non-negative [94]; therefore the ISIDA descriptors are well adapted for this study.

Holliday et al. [9] evaluated the performance of 4 dissimilarity-based methods – MaxMin, MaxMax, MaxSum and MaxMed. These methods follow similar library selection algorithms:

1. Calculation of all N(N-1)/2 pairwise distances (dissimilarities) for N compounds in the dataset.
2. Selection of random compound (Compound 1) from the initial library and addition of it to the subset *S*.
3. Compound 2 is the most remote compound with respect to Compound 1.
4. Identification of the most dissimilar compound to the already selected compounds from the initial dataset using a distance-related score and its addition to the subset *S*.
5. Repetition n-2 times of step 4.

The difference between MaxMin, MaxMax, MaxSum and MaxMed concerns the rules used to evaluate a score used to identify the next object to be included in the subset *S*. Let D_{ij} be the dissimilarity between the i-th molecule in the initial dataset and j-th molecule in the subset *S*. As a function of the algorithm, and the following scores are used: MIN$\{D_{ij}\}$ for MaxMin, MAX$\{D_{ij}\}$ for MaxMax, $\sum\{D_{ij}\}$ for MaxSum and Median$\{D_{ij}\}$ for MaxMed. The object having a maximal score is added to the subset *S*.

It has been shown [9] that MaxMax and MaxMed often led to subsets containing too similar compounds. MaxSum method preferentially selected compounds that were located "at the corners" of the chemical space. These drawbacks are minimized within the MaxMin method, which ensures a relatively uniform selection of compounds from different areas of the chemical space. The most significant drawback of all dissimilarity-based methods is the necessity to calculate the distance matrix for all the compounds. The complexity is proportional to $O(n^2N)$ [91], where *N* is the number of compounds already selected, and *n* is the total number of compounds in the initial library.

Here, the MaxMin algorithm was applied to both reactants and products. For a product-based approach, the MaxMin algorithm has been directly applied, and 225 imines have been selected. Since the algorithm takes the first compound randomly, 100 libraries have been selected, each having a different seed. In the case of the reactant-based approach 15 aldehydes and 15 amines were selected, and the corresponding imines to the pairwise reactions between the selected reactants were directly extracted from the imine library. Since the number of aldehydes and amines is rather low (277 aldehydes and 154 amines) and the algorithm is dependent on the choice of the first compound in the list, for the selection of the most diverse library of amines and aldehydes, an "exhaustive MaxMin" approach has been applied, where each compound has been used as the seed. Thus, the usage of the "exhaustive MaxMin" for the reactant-based approach ensures that the selected libraries of reactants are indeed the most diverse. To compare MaxMin applied on products and on reactants, ten the most diverse libraries of aldehydes and ten most diverse libraries of amines have been selected to generate 100 most diverse libraries of imines.

6.1.2.2 Cell-based diverse-library selection using GTM

In order to use GTM as a method for diverse library design, the compounds were projected on the map, and the map itself was virtually evenly divided into *n* cells, where *n* equals to the number of needed compounds (therefore in this study n=225). From each cell a random compound has been extracted. GTM parameters have to be optimized for this task in order to obtain a map covered as evenly as possible by all the compounds of the initial dataset. To do so, the same protocol of parameter optimization involving the Genetic Algorithm (GA) [70] described previously has been applied. In these calculations, normalized Shannon entropy has been used as the scoring function. The following formula calculates Shannon entropy [68, 95]:

$$E = -\sum_{k} \mathrm{CumR}_k \log(\mathrm{CumR}_k)$$

Where $CumR_k$ is the cumulated responsibilities of compounds in the node k. However, in the GA optimization, the normalized Shannon entropy has been used:

$$E_{norm} = \frac{E}{\log(N)} \times 100$$

Where *N* is the total number of nodes. The normalized entropy ranges within [0; 100], where 0 means that all the compounds are mapped into the same node, and 100 means that the compounds are covering the map uniformly. In other words, the Shannon entropy corresponds to the homogeneity of the distribution of the source library over the map area. The higher the entropy is, the more the objects are dispersed on the map. In such a way, five "best" different descriptors spaces leading to "optimal" maps have been selected (**Table 6-1**).

6.1.2.3 Performance evaluation

The performance of the cell-based approach using GTM has been compared to the performance of the MaxMin algorithm, and the library of randomly selected compounds has been used as the baseline. The quality of selected libraries has been considered taking into account the following criteria:

- Diversity of the selected library. Here two diversity scores were used "All-Soergel" (All_S) and "Min-Soergel" (Min_S); the first calculates the average of pairwise distances for all selected compounds in subset *S*, whereas the latter calculates the average of the distances to the closest neighbor of each compound in the selected subset *S*.

$$All_S = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N-1}\frac{d_{ij}}{N-1}$$

$$Min_S = \sum_{i=1, j=1, i\neq j}^{N}\frac{\mathrm{MIN}(d_{ij})}{N}$$

Notice that the “All-Soergel” score accounts for all pairwise distances in the selected library, whereas “Min-Soergel” shows how dissimilar are the closest neighbors.

- Data coverage. A diverse library of 225 imines will be used to build a model for logK. Which, in turn, will be applied to assess this thermodynamic parameter for the initial set of 42658 imines. However, because of the *fragment control* applicability domain (see section 3.1), some predictions are considered unreliable. In such a way, data coverage is defined as a ration of the molecules for which predictions are considered reliable to the size of the initial data set.

6.1.3 Results

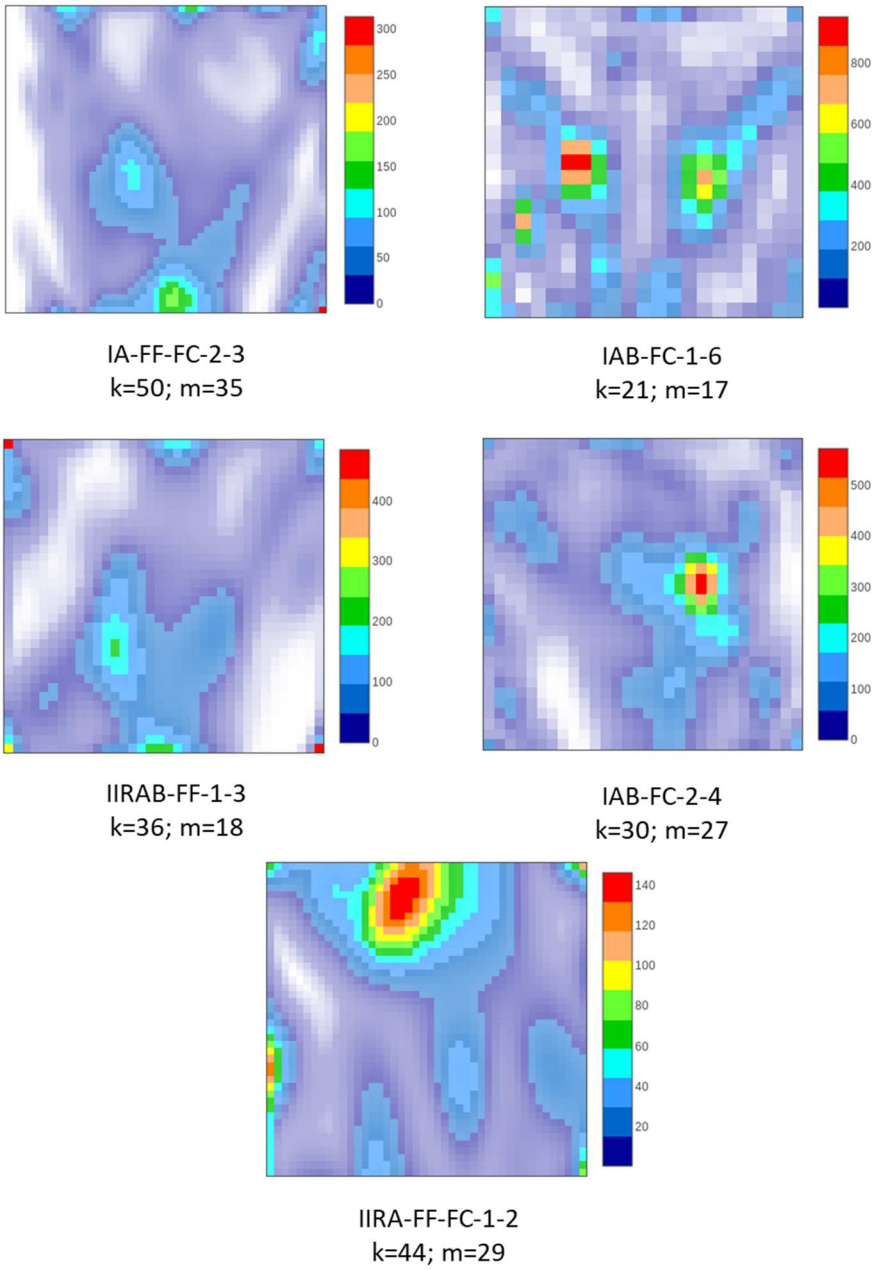


Figure 6.5: Density landscapes of a set of 42658 imines built in 5 different descriptor spaces that have been used for cell-based diverse library selection. Maps parameters are given, where k is the square root of the number of nodes, and m is the square root of the number of RBFs.

According to the Min-Soergel score, MaxMin applied to the dataset of products selects more diverse libraries than any other studied approaches (**Figure 6.6**, **Table 6-2**). In contrast, MaxMin applied to reactants is much less diverse because any imine in the

selected library shares a common aldehyde-substructure with 14 imines and a common amine-substructure with another 14 imines, which reduces the distance to the closest neighbor. According to both All-Soergel and Min-Soergel scores, libraries obtained by cell-based selection are less diverse compared to those selected with the MaxMin algorithm. It can be explained by the fact that MaxMin maximizes the diversity of the selected subset explicitly, while cell-based methods do not consider the distances between the compounds in chemical space. Moreover, MaxMin mainly selects the objects on the "border" of chemical space, while GTM focuses on dense clusters of compounds, and hence, the objects remote from the manifold are not be adequately taken into consideration.

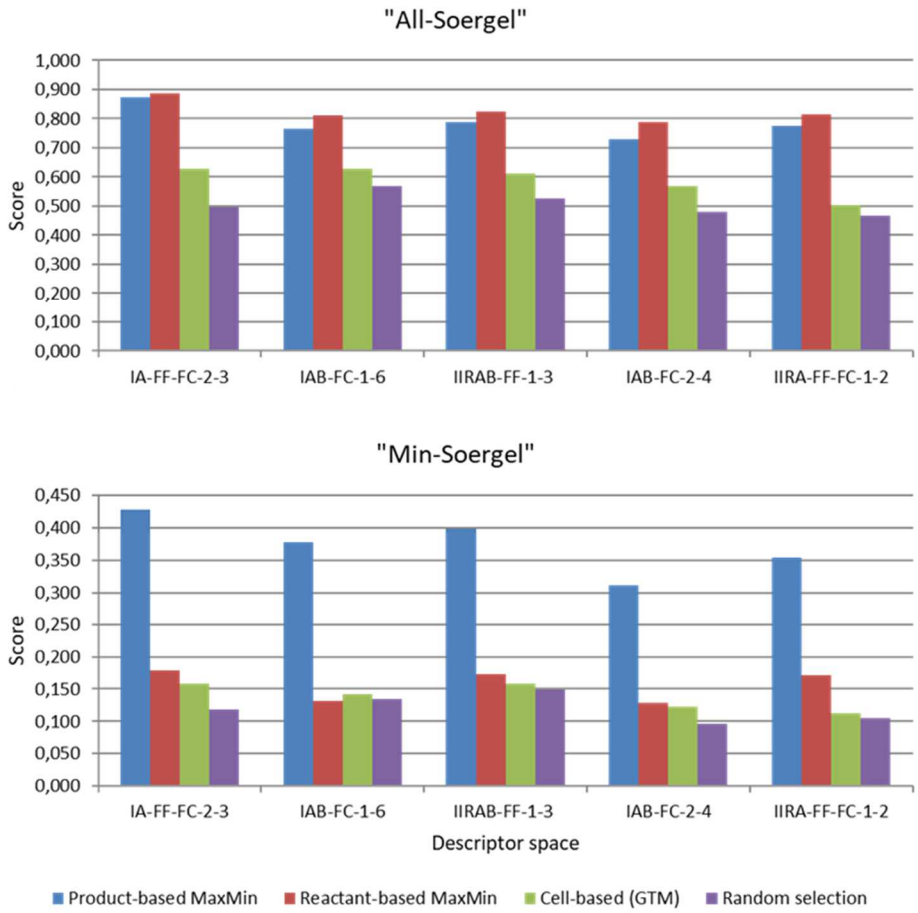


Figure 6.6: Diversity of selected libraries according to the "All-Soergel" score (top) and "Min-Sorgel" score (bottom). Each value is a score average of over 100 diverse libraries selected with different random seeds.

Table 6-2: Diversity of selected libraries in 5 used descriptor spaces. The diversity has been measured using "All-Soergel" and "Min-Soergel" scores. The first calculates the average of the pairwise distances for all the compounds; the latter calculates the average of the distance from each compound to its closest neighbor.

		Product-based		Reactant-based		Cell-based		Random	
		All-Soergel	Min-Soergel	All-Soergel	Min-Soergel	All-Soergel	Min-Soergel	All-Soergel	Min-Soergel
IA-FF-FC-2-3	Mean	0,874	0,428	0,886	0,179	0,628	0,158	0,497	0,118
	Std. Dev	0,002	0,003	0,005	0,003	0,011	0,006	0,024	0,010
	Median	0,874	0,428	0,889	0,179	0,629	0,158	0,495	0,119
	Min	0,869	0,419	0,874	0,172	0,591	0,141	0,463	0,106
	Max	0,881	0,435	0,893	0,187	0,655	0,173	0,547	0,139
IAB-FC-1-6	Mean	0,767	0,378	0,812	0,132	0,628	0,142	0,569	0,135
	Std. Dev	0,003	0,002	0,004	0,005	0,010	0,005	0,028	0,006
	Median	0,767	0,378	0,810	0,130	0,604	0,131	0,583	0,137
	Min	0,761	0,373	0,807	0,125	0,604	0,131	0,506	0,122
	Max	0,774	0,383	0,821	0,139	0,651	0,153	0,604	0,141
IIRAB-FF-1-3	Mean	0,787	0,399	0,825	0,173	0,610	0,158	0,525	0,150
	Std. Dev	0,003	0,002	0,005	0,006	0,013	0,006	0,013	0,006
	Median	0,788	0,399	0,827	0,172	0,609	0,158	0,520	0,152
	Min	0,784	0,394	0,811	0,161	0,587	0,143	0,507	0,138
	Max	0,792	0,402	0,831	0,183	0,644	0,171	0,547	0,156
IAB-FC-2-4	Mean	0,730	0,312	0,790	0,128	0,568	0,122	0,480	0,096
	Std. Dev	0,003	0,002	0,005	0,006	0,011	0,005	0,021	0,007
	Median	0,730	0,312	0,792	0,128	0,569	0,123	0,481	0,097
	Min	0,721	0,307	0,776	0,119	0,538	0,109	0,455	0,087
	Max	0,736	0,316	0,794	0,141	0,591	0,138	0,521	0,107
IIRA-FF-FC-1-2	Mean	0,775	0,354	0,814	0,172	0,504	0,112	0,467	0,106
	Std. Dev	0,003	0,002	0,006	0,004	0,017	0,006	0,021	0,003
	Median	0,775	0,354	0,813	0,173	0,505	0,113	0,470	0,106
	Min	0,766	0,349	0,799	0,162	0,446	0,095	0,424	0,101
	Max	0,780	0,359	0,824	0,177	0,541	0,128	0,497	0,111

On the other hand, the diverse libraries obtained by cell-based selection have several advantages over the classical dissimilarity-based diverse library selection method. First of

all, this concerns high data coverage (**Figure 6.7**, **Table 6-3**). Since every zone of the 2D latent space corresponds to the related zone in the initial N-dimensional chemical space, regular selection objects from the "cells" on GTM corresponds to regular sampling from the initial space.

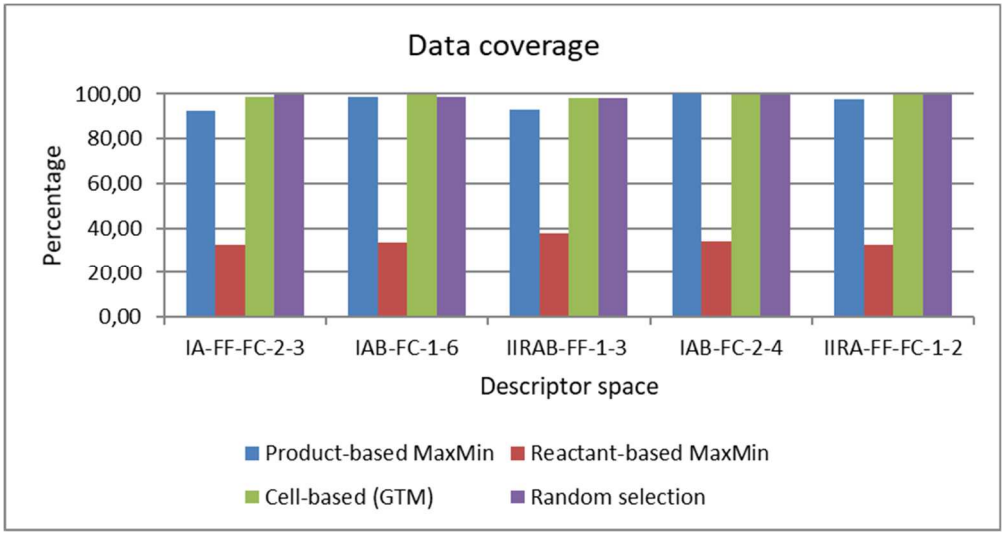


Figure 6.7: Data coverage provided by a diverse library. In the context of this study, the chemical space coverage is defined according to the *fragment control* applicability domain approach.

Table 6-3: Data coverage (%) of the initial dataset of compounds provided by selected diverse libraries.

		Product-based	Reactant-based	Cell-based	Random
IA-FF-FC-2-3	Mean	92,48	32,66	98,78	99,53
	Std. Dev	0,65	6,13	1,46	0,81
	Median	92,21	34,47	99,17	100,00
	Min	92,21	20,81	95,85	97,92
	Max	94,24	39,02	100,00	100,00
IAB-FC-1-6	Mean	98,81	33,63	99,72	98,67
	Std. Dev	0,34	5,68	0,49	1,46
	Median	98,92	34,13	100,00	99,28
	Min	97,83	24,27	98,63	95,93
	Max	98,92	38,92	100,00	100,00
IIRAB-FF-1-3	Mean	93,06	37,68	98,05	98,31
	Std. Dev	2,24	3,78	1,71	1,05
	Median	93,18	37,31	98,99	98,92
	Min	88,18	36,59	95,36	96,35

	Max	96,21	38,67	100,00	99,35
IAB-FC-2-4	Mean	100,00	34,31	99,76	99,90
	Std. Dev	0,00	3,49	0,75	0,32
	Median	100,00	33,32	100,00	100,00
	Min	100,00	31,69	97,63	98,99
	Max	100,00	41,41	100,00	100,00
IIRA-FF-FC-1-2	Mean	97,64	32,63	99,60	99,90
	Std. Dev	2,13	3,63	0,52	0,32
	Median	98,99	33,92	100,00	100,00
	Min	93,80	29,38	98,99	98,99
	Max	100,00	37,19	100,00	100,00

The second advantage of the cell-based approach over MaxMin is the high speed of calculations (**Figure 6.8**). The slowest step of the MaxMin algorithm is pairwise distance calculation. For instance, this step took 1h on average for the calculation of pairwise distances of 42658 compounds. On the other hand, in a cell-based approach, even selection objects on a 2-dimensional map takes several seconds. Moreover, in the latter case, the time of calculations does not depend on the size of the initial dataset

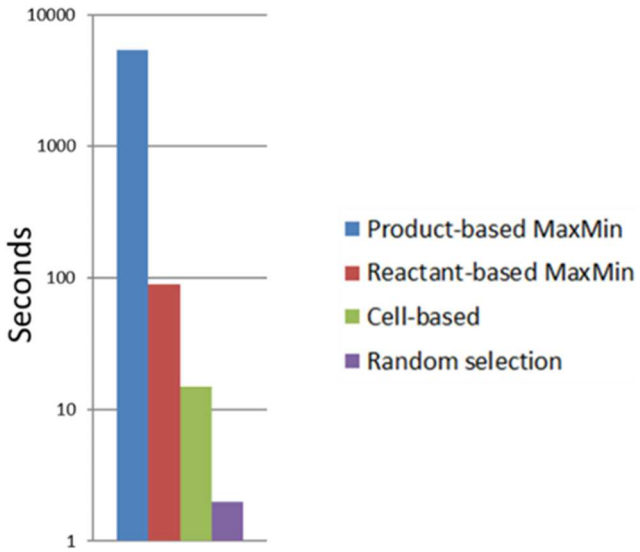


Figure 6.8: The time needed for every approach to select a diverse library of 225 compounds on a computer with Intel(R) Core(TM) i7-6900K CPU and 16 GB of RAM.

One can also envisage a two-step workflow combining MaxMin and cell-based algorithms. In the first step, the cell-based method is applied in order to select an "intermediate" diverse library whose size is larger than of the final set (225 compounds). This "intermediate" set serves as a source library for the MaxMin algorithm, which will select 225 the most dissimilar compounds. The results show (**Figure 6.9**) that the combination of two approaches is a reasonable trade-off between the diversity of the selected library and the time of computations.

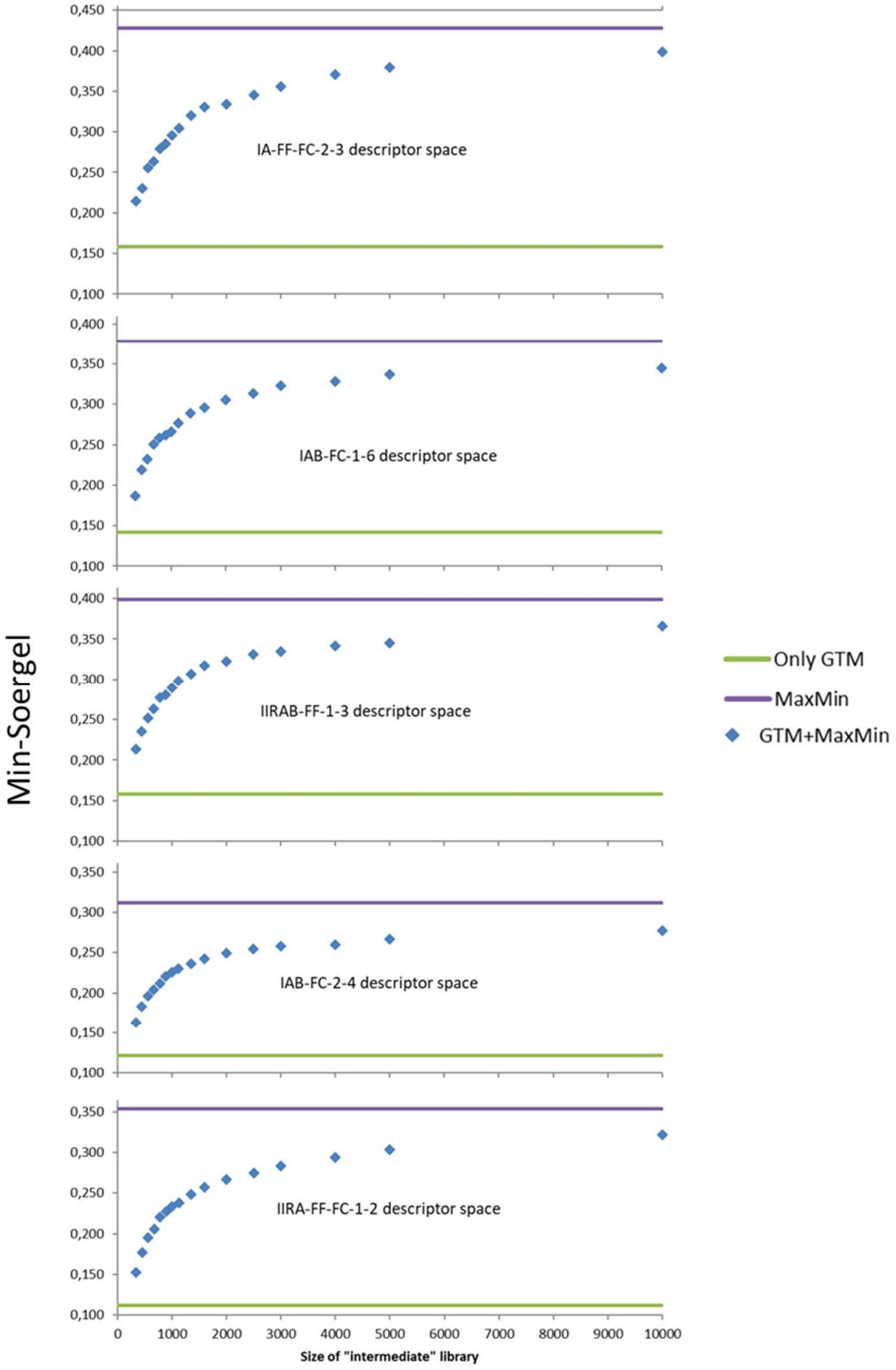


Figure 6.9: Diversity of selected library by consecutive application of cell-based approach and MaxMin algorithm according to Min-Soergel score.

6.1.4 Discussion

The above results demonstrate that GTM is an acceptable method to select diverse library: it performs reasonably well, it is very fast, and its data coverage is almost 100%. On the other hand, all 225 imines selected with GTM may contain unique reactants. This may be a problem for the budget of the experimental laboratory, which needs to purchase 225 amines and 225 aldehydes.

In this regard, the dissimilarity-based MaxMin algorithm applied to reactants might be a reasonable solution despite its relatively small data coverage. Therefore, we decided to use the latter for the selection of a training set for logK of imines formation modeling in a two-steps procedure. In the first step, five different diverse libraries larger than 225 compounds were selected using descriptors spaces mentioned in **Table 6-1**. Their overlap resulted in 15 aldehydes and 15 amines present in all five individual libraries. Then, because of the recommendation of experimentalists, the subset of aldehydes was extended to 24 aldehydes. Their pairwise combinations result in a library of 360 imines, out of which 276 imines were synthesized, and their logK were measured using NMR spectroscopy (see experimental details section).

6.2 Chemoinformatics driven assessment of speciation in dynamic combinatorial libraries

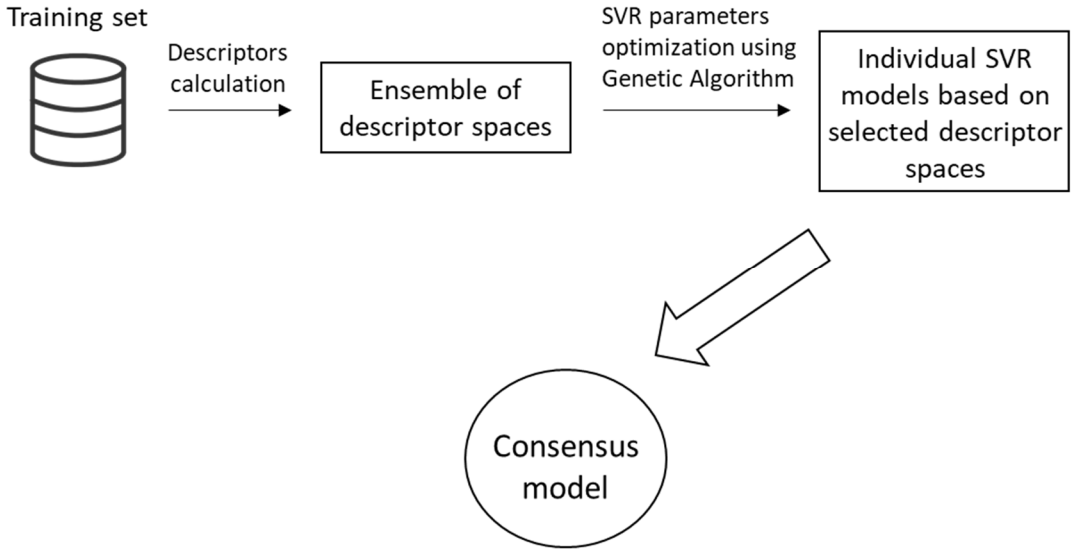


Figure 6.10: Model building workflow applied for both logK and pK_i.

Structures related to logK and pK_i datasets were standardized following the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics at the University of Strasbourg (infochimie.u-strasbg.fr/webserv/VSEngine.html) using the ChemAxon Standardizer. The evolutionary model tuning of the Support Vector Regression [70] approach was applied to grow both stability and CA affinity Support Vector Regression (SVR) models. As the approach can select the best suited molecular descriptors out of a user-provided pool of potentially useful descriptor sets, in both cases, the optimizer was given the freedom to choose its preferred descriptor spaces. However, given the different scopes of the approaches (stability – based on imines representing a combinatorial core of the envisaged DCL/affinity – based on public compounds with reported CA affinity data, most of them not being imines), distinct pools of ISIDA fragment descriptors were fed as possible input. For stability, a series of 18 customized ISIDA fragmentation schemes have been selected as a result of a genetic algorithm; then, in each of the 18 descriptor spaces, an additional evolutionary model tuning has been done. The protein binding predictor was grown, starting with the "default" pool of 100 ISIDA descriptors schemes

acknowledged being potentially useful for biological activity predictions. Also, while the large set of affinity data supported the default "aggressive" 12x repeated 3-fold cross-validation scheme prone by the used GA optimizer tool, logK data is less robust and was subjected to 5-fold cross-validation.

6.2.1 Modeling of equilibrium constants of imines formation

6.2.1.1 Imine formation data

The training set for imines formation was selected using MaxMin algorithms applied to reactants subsets, as explained in section 6.1.4. It contains 276 imines constituted by 24 aldehydes and 15 amines. Stability constants of imines formation (K_{eq}) were measured experimentally using NMR by our collaborators in the Lehn's laboratory (see experimental details below).

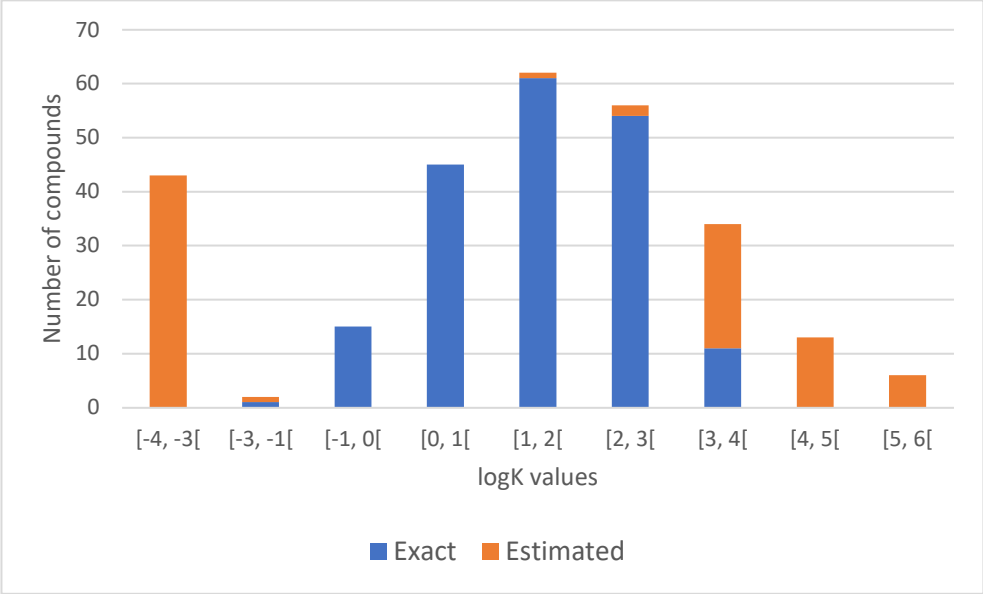


Figure 6.11: logK values distribution. Each measurement was annotated as "exact" and "estimated". Experiments with no detected measuring issues were labeled "Exact". "Estimated" label has been assigned because of (*i*) too weak concentration of reactants/products or (*ii*) peaks superposition, which leads to the difficulties in quantitative identification of compounds.

Figure 6.11 shows the distribution of measured logK values for 276 selected imines. Note that each data point is labeled either "exact" or "estimated". These labels come from NMR limitations. The "estimated" label was given to those logK of imines when the concentration of products/reactants were hard to identify precisely.

6.2.1.2 Experimental details

NMR measurements. Imines were prepared directly in 5 mm NMR tubes by mixing 200 mM stock solutions of components (60 µL each) and diluted with 480 µL of $CDCl_3$ to reach the final concentration of imines of 20 mM. The NMR measurements were performed after 24 h of equilibration at room temperature. ^1H MNR spectra were recorded on 500 MHz Bruker spectrometer with an automated sampler, using standard parameters.

Solvent. In this study, the deuterated chloroform $CDCl_3$ was used to perform all the tests of imine formation for several reasons. Namely, (i) many organic molecules are soluble in chloroform; (ii) the formation of imines in this solvent, in general, is quantitative; (iii) it is a solvent of choice for routine NMR. Before use, chloroform was filtered through the pad of basic alumina to remove residual acid always present. Next, several milliliters of Mili-Q water were added to the bottle to obtain a saturated solution of water in chloroform. This strategy helps to ensure control of the water content to be relatively constant. During the imine formation, one molecule of water is produced, and it participates in the equilibrium. However, it is pretty difficult to measure precisely the amount of water by NMR, so saturation of chloroform with water should solve this problem.

The concentrations of imines, as well as of reactants, were, thus, obtained by integrating the corresponding signals. For most of the cases, the concentrations have been easily identified; hence the associated logK was precisely calculated. However, in some "extreme" cases, the signals' intensity was so weak (either the concentrations of both reactants or imine being low), that it was nearly impossible to integrate the peaks correctly. In this case, the obtained value of logK was labeled as "estimated".

6.2.1.3 logK modeling in chloroform

16 SVR individual models, each built in 16 selected descriptor spaces (**Table 6-4**), contributed to consensus calculations. The resulting consensus model well performs in 5-fold cross-validation: the determination coefficient R^2=0.93 and the RMSE=0.62 log units. A plot of experimental vs. predicted logK values is shown in **Figure 6.12**.

Table 6-4: 5-fold cross-validation performance of the individual logK (in chloroform) regression models that form a consensus model. Fragmentation scheme nomenclature in column 1 denotes the fragment type (I-sequence, II-circular fragments), the nature of captured information (A-atom types are captured, B – bond orders are captured), the coloring scheme (FF – force field type-based labeling supersedes default labeling by atomic symbol), other options (FC – formal charges are considered).

Descriptor space	R^2	RMSE	Data coverage of the initial set of 120k compounds (%)
IAB-1-3	0.92	0.66	66.56
IIAB-1-2	0.92	0.65	52.19
IA-FF-FC-1-2	0.92	0.64	32.62
IAB-1-4	0.92	0.63	23.52
IA-FF-FC-1-3	0.92	0.64	9.91
IAB-1-5	0.92	0.65	7.42
IIA-FF-FC-1-2	0.92	0.63	7.34
IA-FF-FC-1-4	0.92	0.65	4.18
IAB-1-6	0.92	0.65	4.00
IAB-1-7	0.92	0.64	2.81
IA-FF-FC-1-5	0.92	0.64	1.69
IA-FF-FC-1-6	0.92	0.66	1.00
IA-FF-FC-1-7	0.92	0.65	0.93
IIA-FF-FC-1-3	0.92	0.64	0.43
IIAB-1-4	0.92	0.64	0.28
IIA-FF-FC-1-4	0.92	0.65	0.27

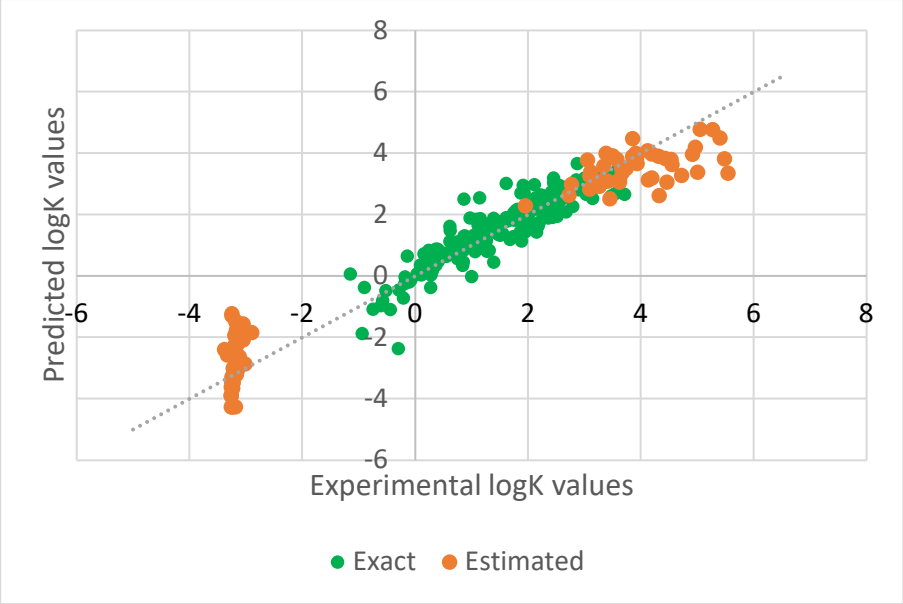


Figure 6.12: Experimental vs. predicted logK values plot of consensus model obtained from 16 SVR models (see Table 1-4). The model's performance is R^2=0.93 and the RMSE=0.62 log units. A gray dotted line corresponds to ideal predictions.

For most molecules, predicted logK values were close to the experiment, whereas the majority of erroneous predictions were detected for the compounds labeled as "estimated".

An important criterion used to estimate the quality of the obtained models is data coverage. Since the 276 imines have been selected from a bigger pull of 120000 imines, it has been decided to identify for how many imines the model can provide reliable predictions. For this purpose, the *fragment control* was used as a model's applicability domain. If a compound has a structural motif not present in the training set structures, then this compound is considered to be out of the applicability domain of the model. Therefore, predictions made for this compound are considered unreliable and should be discarded. It has been found that the SVR consensus model trained on logK of 276 imines is able to provide reliable predictions for 80400 imines, which represents 67% of the entire dataset of 120000 compounds.

The consensus model has been uploaded to the online *Predictor* tool of the Laboratory of Chemoinformatics of the University of Strasbourg (See section 6.3 for details).

6.2.1.4 Chemical space analysis of a set of 80400 imines

THE developed SVR consensus model has been applied to a set of 80400 hypothetical imines identified as being inside of its applicability domain. Distribution of the predicted logK values shows that only 30% of imines have logK> 3 and, therefore, are suited to be used in DCL.

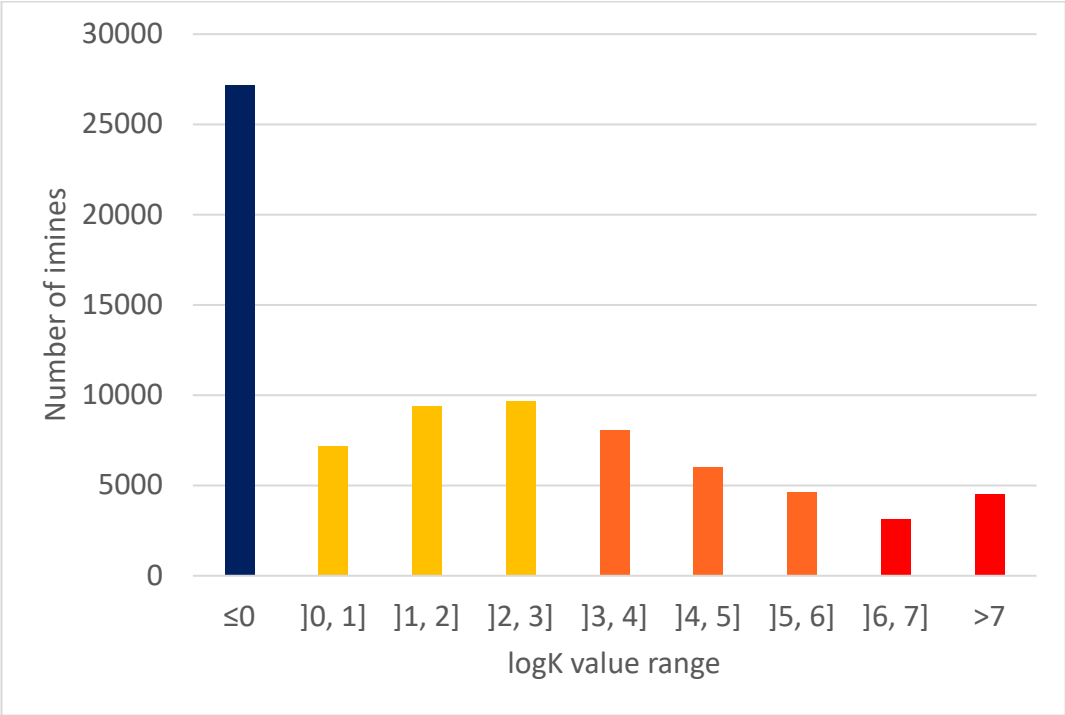


Figure 6.13: Distribution of predicted values of logK of imine formation in chloroform.

GTM has been used to visualize a chemical space of 80400 imines. The manifold has been built on 276 imines with experimentally determined logK values. Then all the imines that are in the applicability domain of the SVR consensus model have been projected on the map, thus creating a property landscape (**Figure 6.14**). One can notice that the majority of the compounds having negative logK values are located on the central and right-central side of the map (green and light/dark blue colors); the compounds having high logK values (orange/red) color are located on the top left and bottom right side of the map. An in-depth analysis of chemotypes has been done in order to see what reactants are usually linked to low and high logK values. Some aldehydes and amines have been identified as “inert” since

they, in 90% they produce imines with a negative logK. As a counterpart to "inert," some "reactive" compounds have been found, in 80% of the cases when a "reactive" aldehyde or amine is involved in imine formation reaction, the outcoming logK values are usually high (> 4). It should be noted that "inert"-amines often bear methyl ester fragment, while 3,5-dibromobenzaldehydes are "inert" (**Figure 6.15**). On the other hand, "reactive" reactants share a common substructure of 3,4,5-trimetoxibenzene and benzodioxole (**Figure 6.16**). Last but not least, some "versatile" reactants were also identified. These reactants can be potentially involved in any imine formation since related logK values range from -8 to 8. The structures of "versatile" reactants are shown in the **Figure 6.17**.

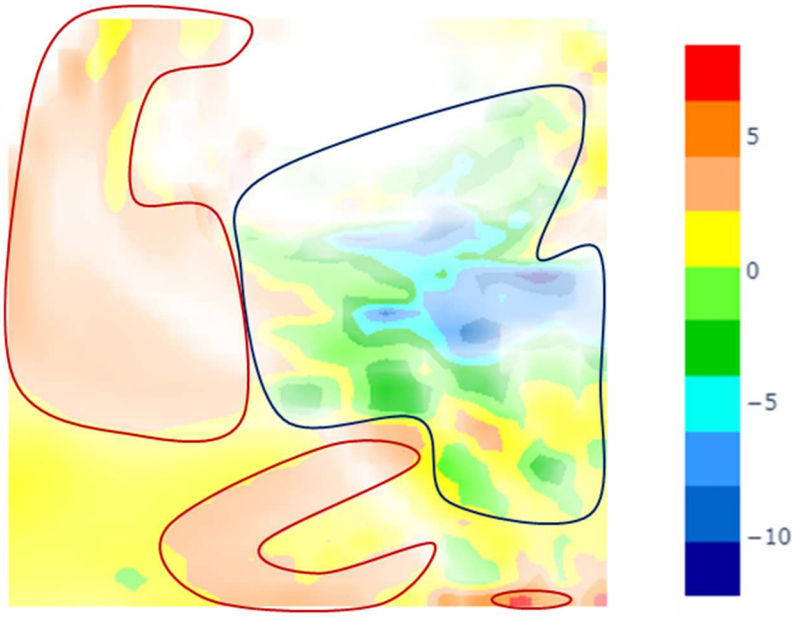


Figure 6.14: Property landscape of the 80k imines that are in applicability domain of the consensus model. The map resolution is 31x31 and the number of RBF is 19x19. Each node is colored according to the mean $logK_{eq}$ value of all the compounds that reside in it. Red lines delineate the zones where "reactive" compounds are located, while the blue line shows the zone populated by "inert".

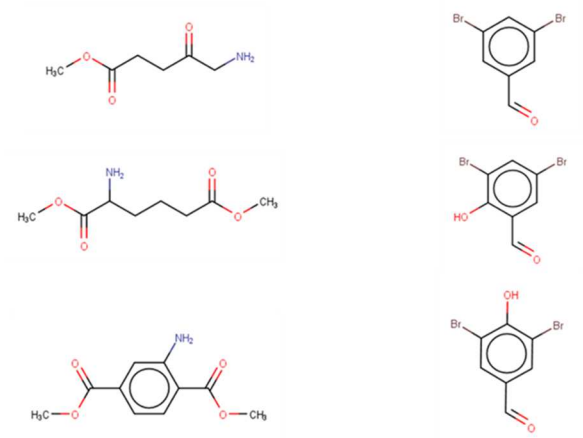


Figure 6.15: Examples of "inert"-amines (left) and "inert"-aldehydes (right). Their interactions with any other aldehydes and amines, respectively, in 90% of cases lead to negatively predicted logK.

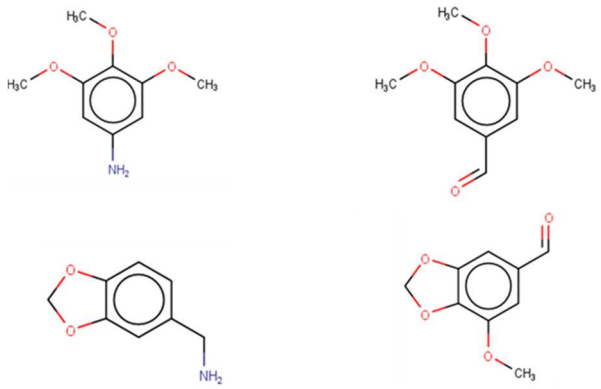


Figure 6.16: Examples of "reactive" amines (left) and "reactive" aldehydes (right). Their interactions with other aldehydes and amines, respectively, in 80% of the cases lead to logK > 4.

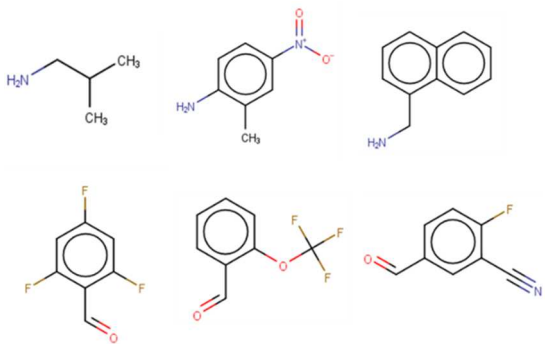


Figure 6.17: Examples of "versatile" reactants. These reactants have been found to yield imines with a very spread range of logK (from -8 to 8).

6.2.1.5 Modeling of logK in water

Since interactions imine – biological target (effector) proceed in aqueous solution, we need to estimate logK in water using the values predicted or measured in chloroform. For this purpose, a thermodynamic cycle shown in **Figure 6.18** was used. The relationship between logK(wat) and logK(chl) is defined by equations below, which require an estimation of solvation energies of reactants and products of imine formation reactions in both solvents.

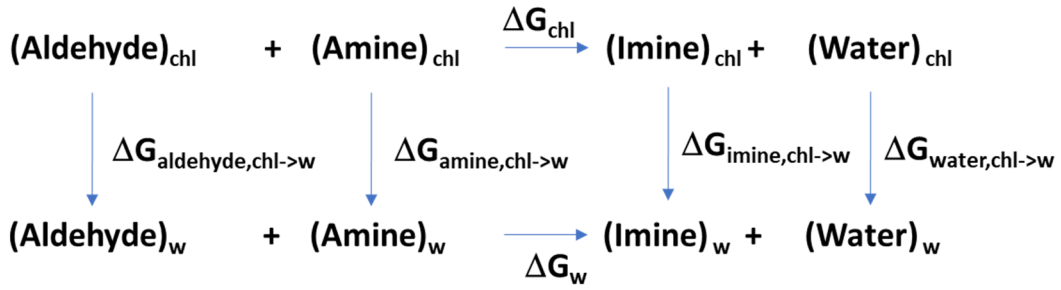


Figure 6.18: Thermodynamic cycle used to establish a relationship between logarithms of stability constant of imine formation in water and chloroform.

$$\Delta G_{chl} = \Delta G_{imine,chl} + \Delta G_{water,chl} - \Delta G_{amine,chl} - \Delta G_{aldehyde,chl}$$

$$\Delta G_{w} = \Delta G_{imine,w} + \Delta G_{water,w} - \Delta G_{amine,w} - \Delta G_{aldehyde,w}$$

$$\Delta G_{chl>w} = \Delta G_{w} - \Delta G_{chl} =$$

$$= \Delta G_{imine,chl>w} + \Delta G_{water,chl>w} - \Delta G_{aldehyde,chl>w} - \Delta G_{amine,chl>w} =$$

$$= \left(\Delta G_{imine,w} - \Delta G_{imine,chl}\right) + \left(\Delta G_{water,w} - \Delta G_{water,chl}\right)$$

$$- \left(\Delta G_{amine,w} - \Delta G_{amine,chl}\right) - \left(\Delta G_{aldehyde,w} - \Delta G_{aldehyde,chl}\right)$$

$$\log K_{wat} = \log K_{chl} + \Delta K_{chl>w} = \log K_{chl} + {}^{\Delta G_{chl>w}}/_{RT}$$

For this purpose, a series of DFT calculations on two selected aldehydes and two amines (**Figure 6.18**, **Table 6-5**), water and four resulting imines have been performed using the *Spartan18* software [95] with ωB97X-D functional in 6-31G* basis and continuous solvation model in water and a non-polar solvent (with a specified dielectric

constant of 7.5). The application of the equations mentioned above resulted in $\Delta K_{chl->w} \approx 0.5$ log units. Note that in water, the solvent is a reaction product. Hence, the massive presence of water would displace equilibrium towards hydrolysis, and might also "capture" a part of the aldehyde reagent under the form of hydrates. However, at this stage, there is no experimental data available to support our calculations, which limits us to the debatable but not absurd working hypothesis that the relative order of imine stability is not too much affected.

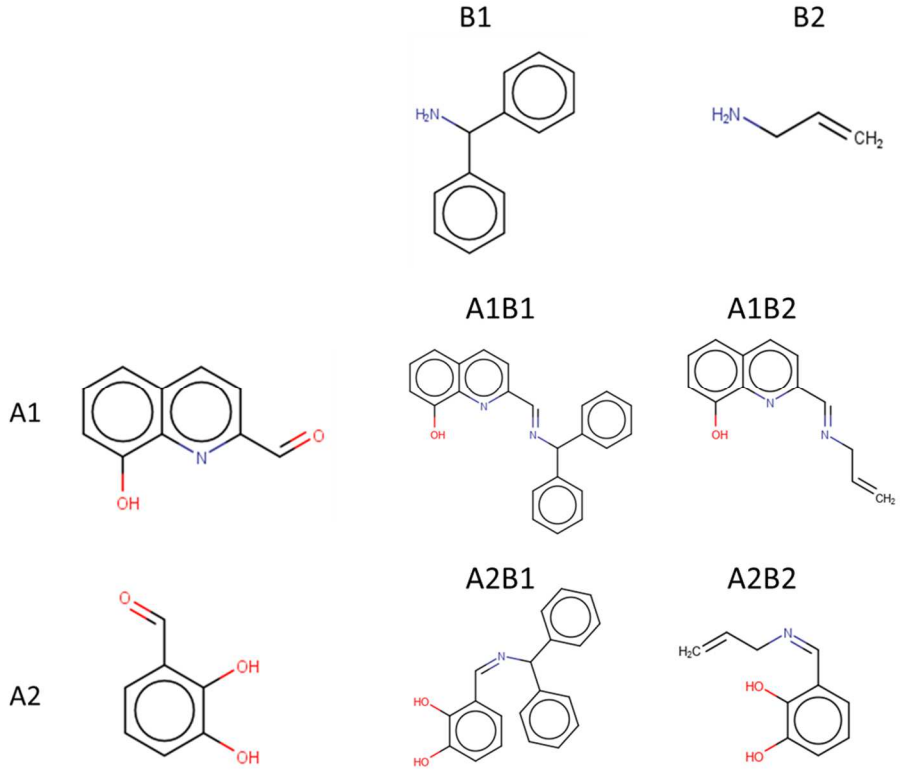


Figure 6.19: Structures of aldehydes, amines and imines forming the DCL.

Table 6-5: The free energies of the compounds used in DCL modeling.

Compound	Water (hartree)	Non-polar solvent (hartree)	ΔG (hartree) solvation	ΔG (kJ/mol) solvation
A1	-590,293	-590,291	-0,00162	-4,30
A2	-495,866	-495,865	-0,00154	-4,08
B1	-557,785	-557,783	-0,00197	-5,23
B2	-173,196	-173,195	-0,00109	-2,88

A1B1	-1071,67	-1071,67	-0,0031	-8,23
A1B2	-687,08	-687,08	-0,00235	-6,22
A2B1	-977,244	-977,24	-0,00323	-8,57
A2B2	-592,654	-592,651	-0,00246	-6,51
Water	-76,398	-76,3965	-0,00148	-3,91

6.2.2 Modeling of pK_i of human CA II

6.2.2.1 Data

ChEMBL database has been used as a source for the inhibition data of human CA II; ChEMBL ID of the target in the 26th version of the database is CHEMBL205. For this target,> 8500 pK_i entries are present. The data have been cleaned, the duplicates, salts and mixtures have been removed, resulting in 4350 unique compounds with experimentally measured pK_i (**Figure 6.20**). 425 compounds out of these 4350 inhibitors of human CA II contain C=N fragment (imines, hydrazones, oximes and Schiff bases); 41 out of 425 compounds having C=N fragments have been identified as imines, most of which have the pK_i values situated between 6 and 9.

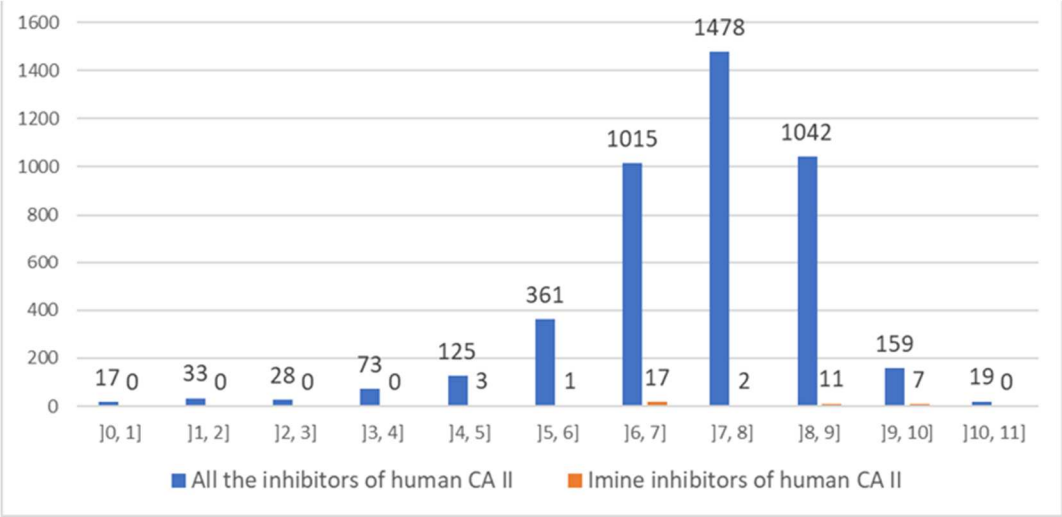


Figure 6.20: The distribution of pKi values of human CA II inhibitors extracted from ChEMBL database.

6.2.2.2 Models description

The affinity model for human CA II is publicly available within the property prediction tool on the web server of the laboratory of Chemoinformatics (see section 6.3). It is composed of a consensus predictor based on the top 5 evolved models using the top 5 best suited ISIDA descriptor spaces. Each of the individual models (**Table 6-6**) is applied to the compounds submitted to the webserver. However, the output of the ones containing the compound to predict within its specific AD (according to the descriptor-specific fragment Control rules) is preferentially used to calculate the returned consensus (mean) value. The mean of all predictions, irrespectively of AD compliance, is also returned – it may be used as a low-trust estimator for compounds that are out of the AD of all the five individual models.

Table 6-6: Cross-validation performance of the five individual pKi prediction models that form the consensus model used in this work to estimate the affinity of imines for the active site of the human carbonic anhydrase II protein.

Descriptor space[a]	RMSE	Q^2
IIAB-FF-1-2	0.37	0.932
IIA--P-FC-1-5	0.27	0.963
IA-FF-P-FC-2-7	0.47	0.889
IIA--1-3	0.28	0.960
IAB-FF-P-2-6	0.40	0.918

[a] Fragmentation scheme nomenclature in column 1 denotes the fragment type (I-sequence, II-circular fragments), the nature of captured information (A-atom types are captured, B – bond orders are captured), the coloring scheme (FF – force field type-based labeling supersedes default labeling by atomic symbol), other options (P – atom pair counts only, FC – formal charges are considered).

The developed pK_i model has been applied to a set of 80400 imines within the applicability domain of the SVR consensus model for imines equilibrium constant. The distribution of *predicted* pK_i values shows that most of the imines should have a pK_i value between 5 and 6 (see Figure 6.21).

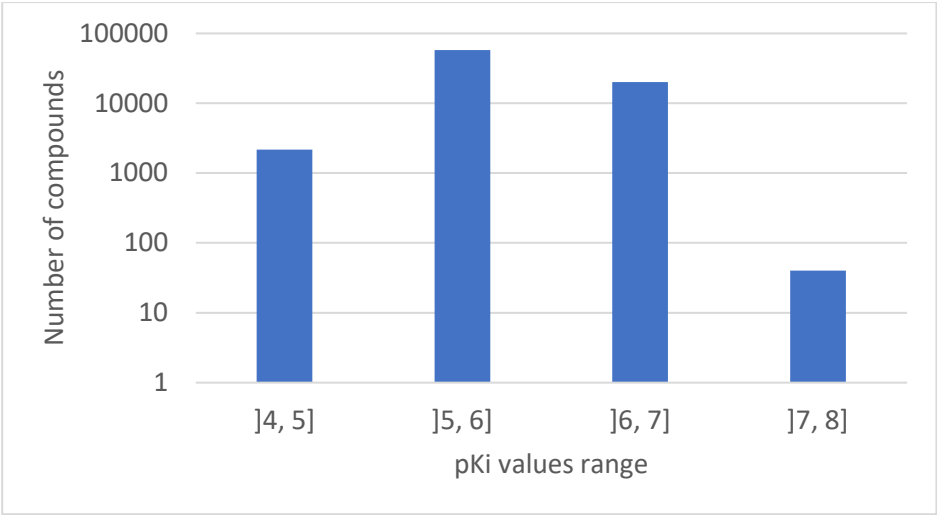


Figure 6.21: Distribution of predicted pK_i values of human CA II for 80400 imines.

6.2.3 Speciation assessment

The ChemEqui program [96] has been used in this project. For the prediction of equilibrium concentration, it uses two groups of equations:

- The law of mass action (Brinkley's representation), where C_i is a reaction product i, βi is the formation constant of component i ν_{ij} is the stoichiometric coefficient in the ith chemical equilibrium involving basic component j with concentration C_j and m is the number of basic components.

$$C_i = exp\left(ln\beta_i + \sum_{j=1}^{m} \nu_{ij} lnC_j\right)$$

- The law of mass conservation, where C^0_j and C^Σ_j are respectively the initial concentration and analytical concentration of component j, r is the number of reactants.

$$\sum_{i=1}^{r} \nu_{ij} C_i = \sum_{i=1}^{r} \nu_{ij} C_i^0 = C_j^\Sigma$$

Then, the following function is minimized:

$$\boldsymbol{y} = \sum_{j=1}^{m}\left[\sum_{i=1}^{r} v_{ij} exp\left(ln\beta_i + \sum_{j=1}^{m} v_{ij} lnC_j\right) - C_j^{\Sigma}\right]^2$$

6.2.3.1 DCL speciation: computational tests.

In order to identify some trends in the variation of equilibrium concentrations of species as a function of equilibrium constants of imines and binding constants protein-imine, three series of computational tests (Batch 1-3) have been performed on a model DCL containing two aldehydes A1 and A2 and two amines B1 and B2 in the absence and the presence of an effector.

Batch 1 involved speciation simulations of DCL without effector where equilibrium concentrations of imines A1B1, A1B2, A2B1 and A2B2 were calculated as a function of logK varied in a narrow range from 4.15 to 4.50. The first simulation was performed for logK = 4.25 for all imines; then, this value was slightly varied. Results given in **Table 6-7** show that the equilibrium concentrations of the products are no longer quasi-equal if :

- logK of only one imine changes on > 0.25 compared to its initial value,
- logK of agonistic imines either both increase or both decrease on > 0.15
- logK of antagonistic imines increases for one and decreases for another one on > 0.10

Table 6-7: Speciation of a DCL formed of two aldehydes and two amines as a function of the logarithm of equilibrium constants of imine formation reaction (logK). The logK values and the proportions correspond to the imines in following order A1B1/A1B2/A2B1/A2B2. The ideal case (equal proportions of products) is given in bold.

logK of 4 imines	*Proportions of obtained imines (%)*
4.25/4.25/4.25/4.25	**47.4 / 47.4 / 47.4 / 47.4**
4.25/4.20/4.25/4.25	48.7 / 45.8 / 46.1 / 48.7
4.25/4.20/4.20/4.25	50 / 44.6 / 44.6 / 50
4.30/4.25/4.25/4.25	48.9 / 46.1 / 46.1 / 48.6
4.30/4.25/4.25/4.20	47.7 / 46.1 / 46.1 / 48.6
4.35/4.25/4.25/4.15	48.5 / 47.4 / 47.4 / 46.8
4.35/4.25/4.15/4.25	53.1 / 42.3 / 41.7 / 52.6

4.40/4.25/4.25/4.40	55.8 / 39.5 / 39.5 / 55.8
4.50/4.25/4.25/4.25	55.1 / 40.8 / 40.8 / 53.8

In batch 2, all equilibrium constants of imines were taken equal (logK = 4.25), whereas the negative logarithm of binding constant pK_i. varied in the range of 2.5 – 4.25. Results given in **Table 6-8** show that the effector exclusively selects only one imine from the solution if the pK_i value for one selected imine is larger than that for the others by > 1.75.

Table 6-8: Speciation of a DCL formed of 2 aldehydes. 2 amines and 1 effector as a function of the logarithm of binding constants (pK_i) of the effector. The pK_i values and the proportions of free (complexated) imines correspond to the imines in following order A1B1/A1B2/A2B1/A2B2. An optimal case (a sole imine is selectively binding to the effector) is given in bold.

pK_i of 4 imines	Proportions of free (complexed) imines (%)
4.25/4.25/4.25/4.25	23.3 (24.9) / 23.3 (24.9) / 23.3 (24.9) / 23.3 (24.9)
4.25/3.25/3.25/4.25	23.5 (48.9) / 23.5 (0.5) / 21.5 (0.5) / 23.5 (48.9)
4.25/4.25/3.25/3.25	5.4 (48.9) / 5.4 (0.5) / 43.2 (0.5) / 43.2 (48.9)
4.25/3.25/3.25/3.25	10.8 (54.6) / 21.3 (10.7) / 21.3 (10.7) / 41.9 (21.1)
4.25/3.0/3.0/3.0	8.6 (60.9) / 20.2 (8.06) / 20.2 (8.06) / 47.6 (18.97)
4.25/2.75/2.75/2.75	6.8 (66.4) / 19.0 (5.9) / 19.0 (5.9) / 53.2 (16.4)
4.25/2.5/2.5/2.5	**5.4 (70.8) / 17.9 (4.1) / 17.9 (4.1) / 58.7 (13.6)**
4.25/2.75/2.5/2.5	5.6 (69.3) / 16.9 (6.7) / 19.1 (4.2) / 58.0 (12.9)
4.25/3.0/2.5/2.5	5.6 (67.0) / 16.9 (10.2) / 19.1 (4.4) / 58.0 (11.9)
4.25/3.25/2.5/2.5	5.9 (63.8) / 13.9 (15.0) / 23.6 (4.5) / 55.3 (10.6)

The third batch of simulations consisted of a DCL of 2 aldehydes. 2 amines, and an effector. The goal of the simulations was to determine how the speciation changes with the variation of logK of imines in the presence of effector. The starting simulations were performed with equal equilibrium constants for all four imines (logK=4.25) and with pKi values favoring selective binding of only one imine A1B1 (pK_i = 4.25/2.5/2.5/2.5). The

only logK varied in the range of 2.5 – 4.25. The results (**Table 6-9**) have shown that the selective binding with the selected imine takes place if:

- logK of selected imine drops by 1.25 with respect to its initial value.
- logK of selected imine and its "agonistic pair" changes by 0.6
- logK of selected imine decreases/increases and its "antagonistic pair" increases/decreases by 0.75

Table 6-9: Speciation of a DCL formed of 2 aldehydes. 2 amines in the presence of an effector as a function of logK of the four imines. The pK_i values and the proportions of free (complexated) imines correspond to the imines in following order A1B1/A1B2/A2B1/A2B2.

logK of 4 imines	Proportions of free (complexed) imines (%)
4.25/4.25/4.25/4.25	5.4 (70.8) / 17.9 (4.1) / 17.9 (4.1) / 58.7 (13.6)
4.0/4.25/4.25/4.25	4.6 (66.9) / 21.0 (5.5) / 21.0 (5.5) / 54.0 (14.0)
3.75/4.25/4.25/4.25	3.8 (62.3) / 24.3 (7.1) / 24.3 (7.1) / 49.0 (14.3)
3.25/4.25/4.25/4.25	2.5 (51.4) / 31.0 (11.4) / 31.0 (11.4) / 38.7 (14.2)
3.0/4.25/4.25/4.25	2.0 (45.3) / 34.3 (14.0) / 34.3 (14.0) / 33.6 (13.8)
3.95/4.25/4.25/3.95	3.6 (61.1) / 25.6 (7.8) / 25.6 (7.8) / 45.7 (13.8)
3.65/4.25/4.25/3.65	2.2 (48.8) / 33.6 (13.2) / 33.6 (13.2) / 32.2 (12.6)
4.0/4.5/4.25/4.25	3.9 (62.6) / 24.8 (7.2) / 24.2 (7.0) / 49.1 (14.2)
3.75/4.75/4.25/4.25	2.6 (52.2) / 32.2 (11.6) / 30.1 (11.1) / 38.6 (13.8)
3.5/5.0/4.25/4.25	1.7 (40.5) / 39.4 (17.1) / 37.6 (16.4) / 28.3 (12.3)

6.2.3.2 Speciation simulation in model DCL based on predicted logK and pK_i values.

It follows from the above computational tests that

- logK of considered imines should be as close as in order to obtain their quasi-equivalent distribution.
- pK_i values should differ by at least 1.5 log units.

Following these recommendations, four imines resulted from the interaction between two aldehydes, and two amines shown in **Figure 6.19** have been selected. Their predicted $logK_{wat}$ and pK_i values are given in **Table 6-10**.

Table 6-10: Predicted logK and pK_i of the species present in DCL.

Compound	logK in water	pK_i
Aldehyde A1	-	4.63
Aldehyde A2	-	5.57
Amine B1	-	6.13
Amine B2	-	3.54
Imine A1B1	4.17	6.11
Imine A1B2	4.26	4.63
Imine A2B1	3.84	6.06
Imine A2B2	3.96	4.99

The modeling of the dynamic behavior of this DCL shows (**Table 6-11**, **Figure 6.22**) that before the addition of the human CA II the concentrations of all the imines in solution are almost equal. According to the predicted pK_i values, the imine A1B1 has the highest binding affinity with the effector (pK_i = 6.11), but binding with its antagonist A2B1 is competitive (pK_i = 6.06). Initially, the concentrations of all the species (reactants + effector) are set to 10 mmol. Following the Le Chatelier principle, it is expected that the effector would mostly bind A1B1 and A2B1, thus leading to a shift of equilibria in DCL toward these imines. In turn, this would decrease the concentration of free reactants A1 and B1 in solution. The calculations well reproduce these effects, see **Table 6-11**.

Table 6-11: Speciation of the DCL before and after the addition of human CA II to the solution. Note that initial concentrations of the two aldehydes, two amines and the effector were set to 10 mmol/L.

Compound	Concentration (mmol)	
	Before	After
Aldehyde A1	0.44	0.34
Aldehyde A2	0.88	0.64
Amine B1	0.73	0.21
Amine B2	0.60	0.66
Imine A1B1	4.77	1.08

Imine A1B2	4.78	4.13
Imine A2B1	4.49	0.94
Imine A2B2	4.62	3.69
A1-CA	-	0.04
A2-CA	-	0.67
B1-CA	-	0.81
B2-CA	-	0.006
A1B1-CA	-	3.91
A1B2-CA	-	0.50
A2B1-CA	-	3.04
A2B2-CA	-	1.02

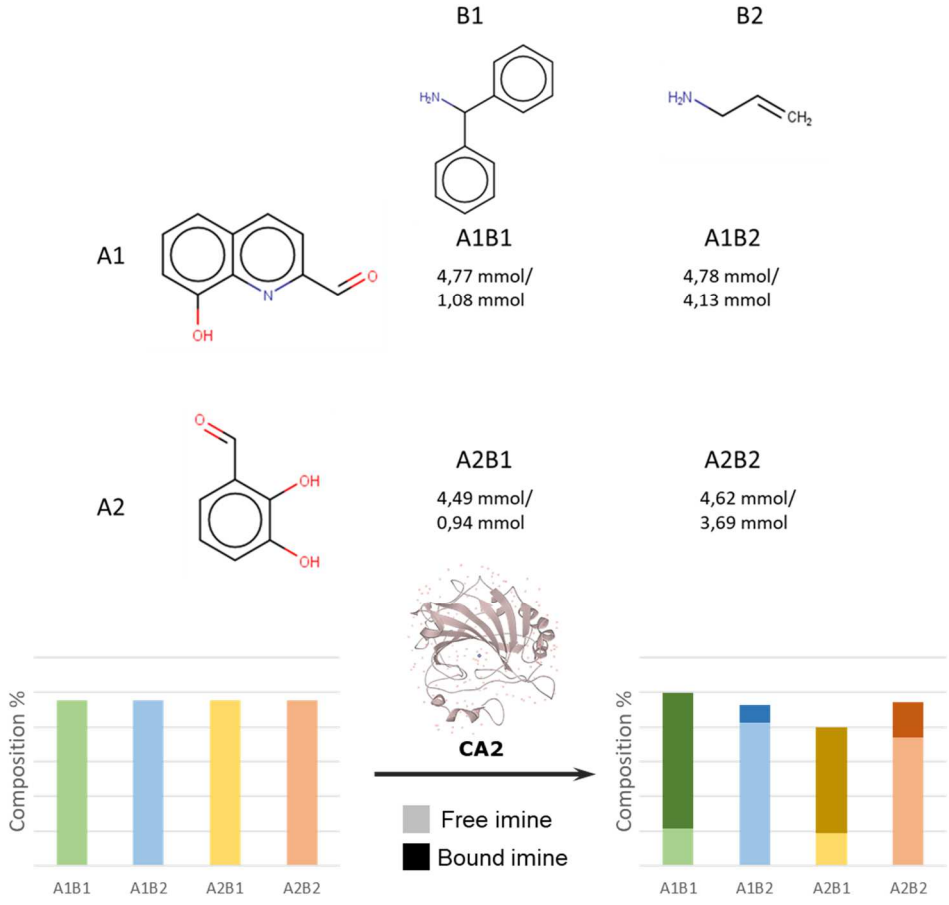


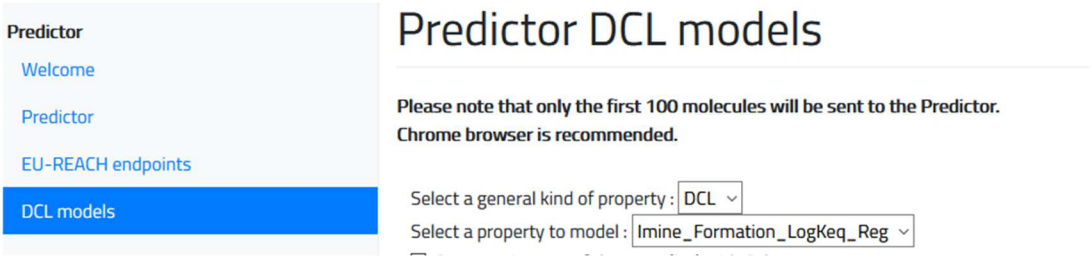
Figure 6.22: Speciation of a hypothetical DCL of 2 aldehydes and 2 amines. The concentrations of resulting imines before/after the addition of human CA II are shown.

6.3 Models implementation

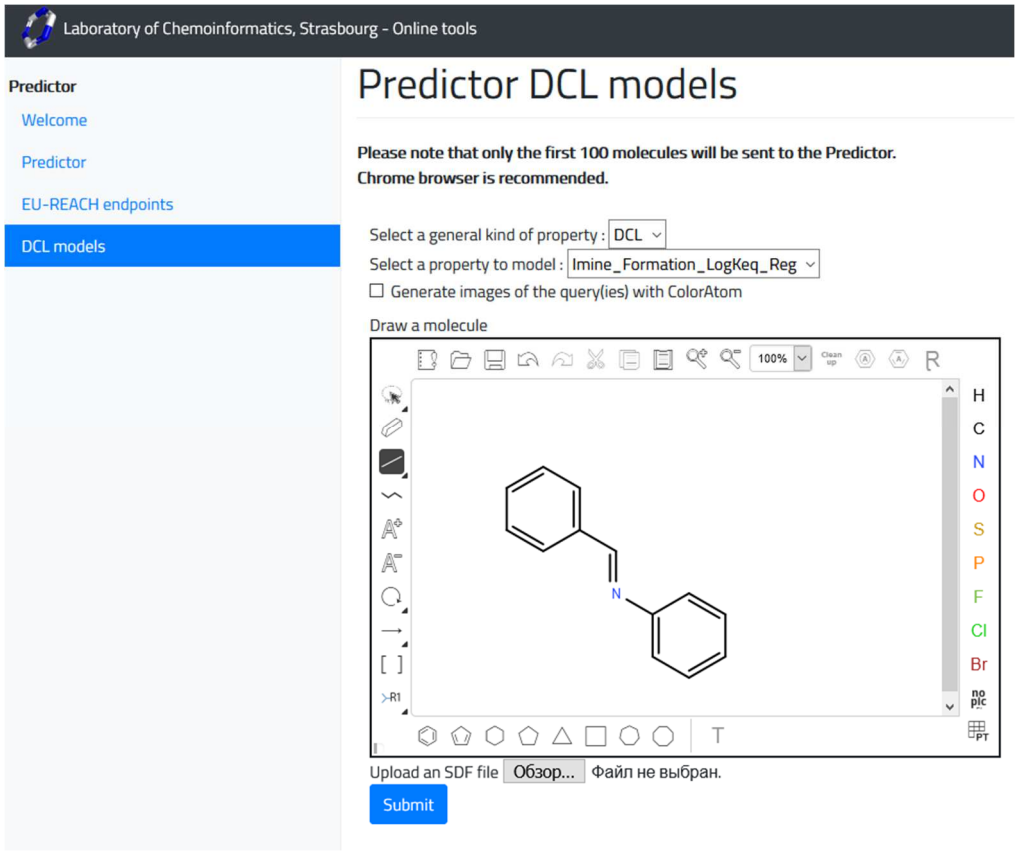
6.3.1 Predictive models of logK of imine formation in chloroform

The obtained consensus SVR models were uploaded on the *Predictor* service of the laboratory of Chemoinformatics (http://infochim.u-strasbg.fr/cgi-bin/predictor_dcl.cgi). A short "user-guide" (Google Chrome v. 84.0 or Mozilla Firefox v. 78.0 browsers) for proper usage of the implemented models is given below:

1. First, select the "DCL model" on the left menu of the predictor and set the "Select a general kind of property" to **DCL** and "Select a property to model" to **Imine_Formation_LogKeq_Reg.**



2. The user can either draw the imine for which he wants to receive a prediction or, in the case of multiple imines, he can give an sdf. Note that the Predictor tool can treat only 100 compounds. Once one of the two input possibilities were chosen, click on the "Submit" button.



3. The information on the input compounds (molecule ID and molecule name) and the results like the *consensus prediction* (Predicted value field), number of applied models and the prediction confidence will be shown on the new webpage, if needed, they can be downloaded.

Predictor DCL models

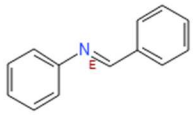
Selected model : Imine_Formation_LogKeq_Reg

Please note that only the first 100 molecules have been sent to the Predictor.

All calculations processed!

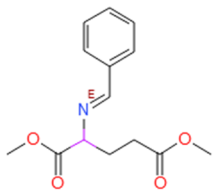
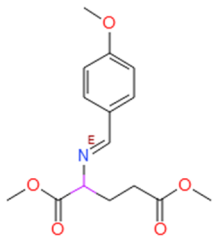
Thank you for your patience.

Explore results with the scrollbar

Molecule Id	Molecule Name	Predicted value	Applied models	Prediction confidence	2D structure
1		0.798	16/16		

Back to Main Menu Download results

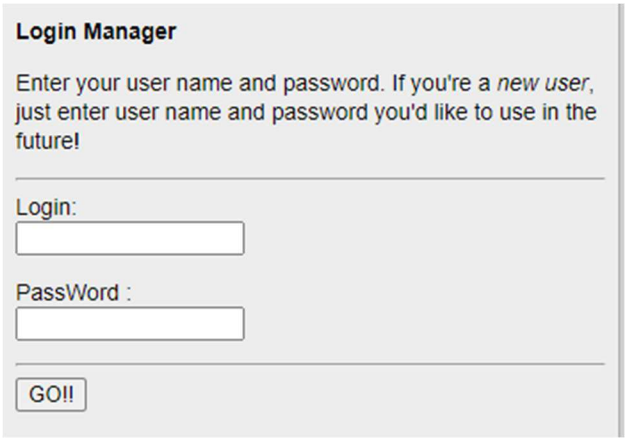
Or in the case of an input file:

Molecule Id	Molecule Name	Predicted value	Applied models	Prediction confidence	2D structure
1		0.536	2/16		 unknown chirality
2		1.115	2/16		 unknown chirality

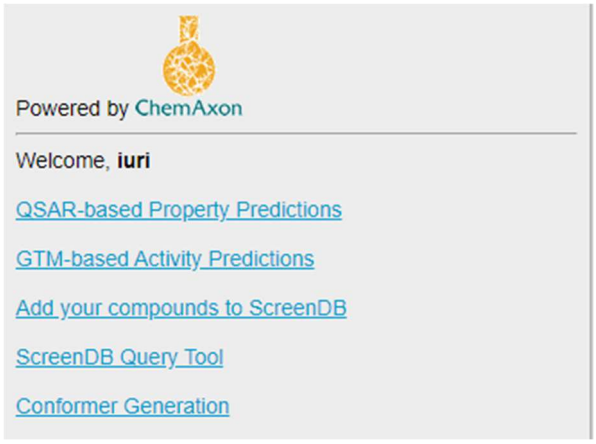
6.3.2 Predictive models of pK_i of human CA II

These models have been uploaded on the web server of the laboratory of Chemoinformatics (http://infochim.u-strasbg.fr/webserv/VSEngine.html) with the help of Dr. Dragos Horvath. A short "user-guide" (Google Chrome v. 84.0 or Mozilla Firefox v. 78.0 browsers) for proper usage of the implemented models is given below:

1. Enter a preferred username and password in the top left corner of the webpage in order to connect to the web server.

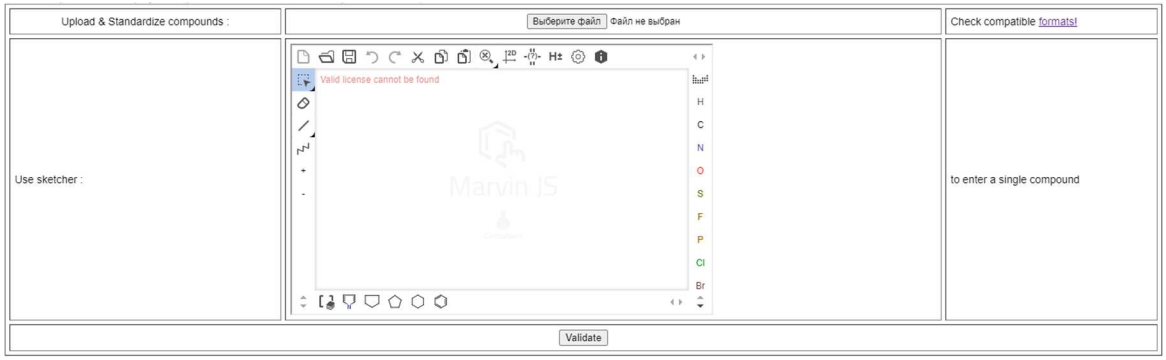


2. Once connected to the web server, click on "QSAR-based Property Predictions" located on the top left corner of the webpage.

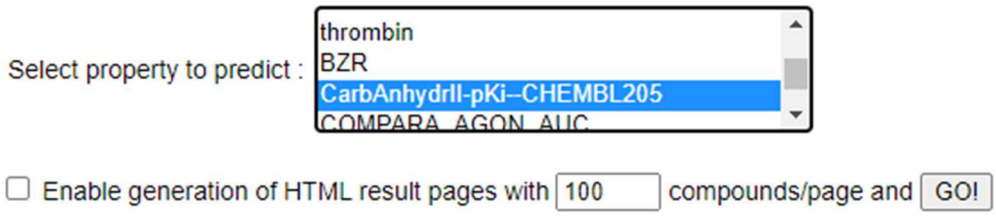


3. The webpage will be refreshed and a new field will appear. asking the user to enter the project name.

4. On a new webpage, the user can either draw a compound or provide an sdf or smiles file containing multiple compounds. The web server will then standardize the provided compound(s). Do note that for large files having thousands of compounds, it may take some time.



5. Once the standardization is done, the user will be redirected to a webpage of *all the implemented models* on it. The model concerning the current project is called **CarbAnhydrII-pKi-CHEMBL205**. Once the model is selected, the user can proceed to the predictions by clicking on the "GO!" button. The webserver tool will start the needed descriptors generation and eventually will proceed to the predictions.

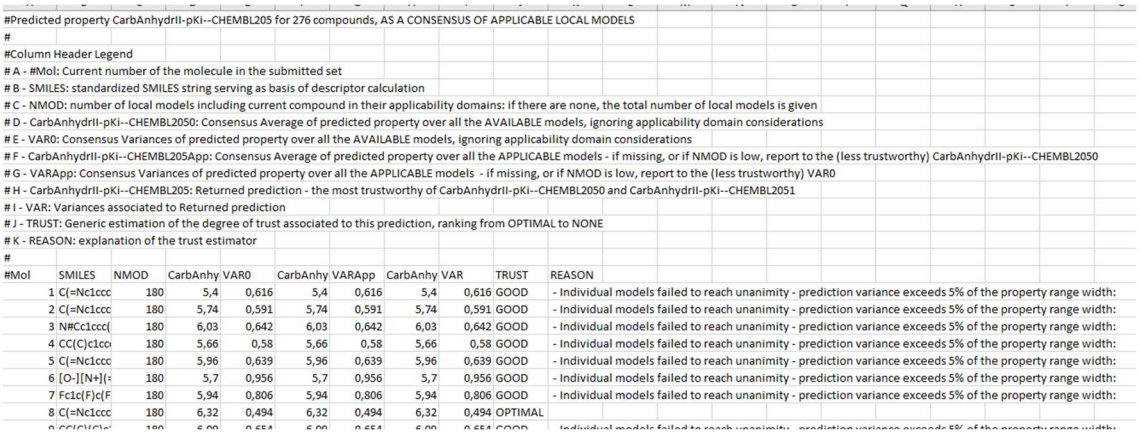


6. Once the predictions are made, the webserver tool will generate a .csv file containing the predicted values.

Available Prediction Results

CarbAnhydrII-pKi--CHEMBL205 .csv format

7. The obtained file will contain a short description of the used model as well as the number of the compounds that have been sent to the prediction tool. An explanation for every column will also be given.



#Predicted property CarbAnhydrII-pKi--CHEMBL205 for 276 compounds, AS A CONSENSUS OF APPLICABLE LOCAL MODELS
#
#Column Header Legend
A - #Mol: Current number of the molecule in the submitted set
B - SMILES: standardized SMILES string serving as basis of descriptor calculation
C - NMOD: number of local models including current compound in their applicability domains: if there are none, the total number of local models is given
D - CarbAnhydrII-pKi--CHEMBL2050: Consensus Average of predicted property over all the AVAILABLE models, ignoring applicability domain considerations
E - VAR0: Consensus Variances of predicted property over all the AVAILABLE models, ignoring applicability domain considerations
F - CarbAnhydrII-pKi--CHEMBL205App: Consensus Average of predicted property over all the APPLICABLE models - if missing, or if NMOD is low, report to the (less trustworthy) CarbAnhydrII-pKi--CHEMBL2050
G - VARApp: Consensus Variances of predicted property over all the APPLICABLE models - if missing, or if NMOD is low, report to the (less trustworthy) VAR0
H - CarbAnhydrII-pKi--CHEMBL205: Returned prediction - the most trustworthy of CarbAnhydrII-pKi--CHEMBL2050 and CarbAnhydrII-pKi--CHEMBL2051
I - VAR: Variances associated to Returned prediction
J - TRUST: Generic estimation of the degree of trust associated to this prediction, ranking from OPTIMAL to NONE
K - REASON: explanation of the trust estimator
#

#Mol	SMILES	NMOD	CarbAnhy	VAR0	CarbAnhy	VARApp	CarbAnhy	VAR	TRUST	REASON
1	C(=Nc1ccc	180	5,4	0,616	5,4	0,616	5,4	0,616	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
2	C(=Nc1ccc	180	5,74	0,591	5,74	0,591	5,74	0,591	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
3	N#Cc1ccc(	180	6,03	0,642	6,03	0,642	6,03	0,642	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
4	CC(C)c1cc	180	5,66	0,58	5,66	0,58	5,66	0,58	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
5	C(=Nc1ccc	180	5,96	0,639	5,96	0,639	5,96	0,639	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
6	[O-][N+](=	180	5,7	0,956	5,7	0,956	5,7	0,956	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
7	Fc1c(F)c(F	180	5,94	0,806	5,94	0,806	5,94	0,806	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 5% of the property range width:
8	C(=Nc1ccc	180	6,32	0,494	6,32	0,494	6,32	0,494	OPTIMAL	

6.4 Conclusions

To understand the behavior of a DCL, the relative propensity of the formation of DCL products and their relative affinities for the target are needed. The first steps in this direction have been done, including the preparation of predictive models for imines stability constants and affinities of organic molecules for the human CA II. While this work underlies the conceptual workflow and reports useful and publicly available models, experimental proof of the reported here *in silico* approach is needed.

The calculation of speciation of any solution does not present any technical difficulty due to numerous software; its utilization remains constrained by the availability of experimental data. The study presented herein aimed to overcome that constrain by the usage of current chemoinformatics methods and tools. First, from the pool of 400 most cited aromatic aldehydes and 300 most cited primary amines (according to SciFinder), diverse subsets of 24 aldehydes and 15 amines (resulting in 276 imines) have been selected.

These imines have been synthesized. Their logK measured and the obtained data served as a training set for building SVM classification and regression models. These models showed high “extrapolation potential” since they were able to provide reliable predictions for more than 80000 imines from the initial pool of 120000.

Although the models themselves are useful and the obtained predictions can be already used for the speciation of a DCL without any effector, they are not sufficient enough for the modeling of a DCL in the presence of any effector. In order to model the presence of the effector, it is needed to quantitatively know the affinity of each constituent in solution to it for this purpose. ChEMBL database has been used as a source of binding affinity (pK_i) on human Carbonic Anhydrase II. More than 4300 compounds have been extracted and used to train SVM regression models.

Developed in this work, predictive models for logK and pKi allowed us to overcome the “experimental data constrain” and thus make possible the usage of speciation software. In considered here hypothetical DCL, the selected quartet of imines has close *predicted* values of logK, thus ensuring a quasi-equal concentration of imines. Moreover, for imines, related pK_i values differ by at least 1.5 log units, thus ensuring a certain level of selectivity. It follows that on the example of “imine-based” DCL modeling, the available speciation software coupled to chemoinformatics tools could be, in principle, used to any type of DCL and any chemical/biological effector if there is a sufficient amount of data for the modeling.

7 Conclusion and Perspectives

This work was devoted in the first place to the application and study of Generative Topographic Mapping in virtual screening. GTM has also been used in a task of selection of diverse libraries of imines, and the quality of the obtained libraries has been compared to the classic dissimilarity-based method of diverse library selection – MaxMin. Last but not least, the first steps towards the modeling of protein-templated dynamic combinatorial libraries have been made.

Two projects focused on the application and exploration of GTM have been done. The first project was rather methodologically-oriented since it provided several universal maps (multi-target GTM-based classification models) that can accommodate more than 1.5M compounds extracted from ChEMBL and to discriminate actives from inactives with high accuracy for 617 targets out of 618. Moreover, this study has shown that the maps are complimentary since each map has been built in different descriptor spaces; therefore, if certain target-specific activities are poorly predicted by one map, there would be another map that will be able to do it. The usage of DUD targets as external-validation sets helped to identify that the correlation between predictions quality in cross-validation is weak. This fact motivated the usage of the universal maps in the consensus model since one could not tell *apriori* what map will show better results in "real-life tasks". It has been shown that universal maps applied in consensus provide undeniable advantages such as i) 100% of data coverage for most of the targets; ii) higher performance in cross-validation *and* external validation according to BA and ROC AUC; iii) higher enrichment factor for top 100 predicted "active" compounds.

The goal of another project was to find new inhibitors of Bromodomain 4 by virtually screening a collection of 2M compounds. In this project, a virtual screening funnel was

composed, including the building of ligand-based pharmacophore models, SVM and GTM-based classification models on publicly available data extracted from REAXIS and ChEMBL. The obtained models have been used in consensus and selected 12k compounds that have been predicted by most of the models to have the highest probability of being active. Then, this subset of 12k compounds has been subjected to a docking procedure, which selected 3k most potent compounds. These compounds have been tested by our collaborators from Enamine. Out of 2992 tested compounds, only 29 compounds have been identified as "active". While this result is still 2.6 times better than the screening of a set of 3k *randomly* selected compounds, 29 confirmed hits objectively is a low success rate. First of all, it has been shown that public data from different sources cannot be fused into a single and rigorously defined dataset adapted for QSAR modeling. It has been found that the active/inactive labels of training data have been assigned according to pK_i and IC_{50} values, while the experiment has been done using Differential Scanning Fluorimetry – a method that measures ΔT_m (thermal denaturation temperature). Moreover, it has been shown that the correlation between IC_{50}/pK_i and ΔT_m. is very low. Still, it has been shown that GTM classification models were able to find 24 out of 29 confirmed hits.

For the first time, GTM has been used for a selection of a diverse library of compounds. The quality of the selected libraries has been compared to the quality of libraries obtained by MaxMin – a classic dissimilarity-based method for diverse libraries selection. Since the term and the metrics of "diversity" are vaguely defined, a score based on Soergel distances of the compounds included in the diverse library has been used. It has been shown that GTM as an individual model cannot provide diverse libraries with the same level of dissimilarity as MaxMin; however, GTM-based diverse library selection selects a diverse library that is more "representative" than the libraries selected by MaxMin. Moreover, it was found that the application of GTM for a pre-selection of a bigger diverse library ("intermediate" diverse library) followed by the application of the MaxMin method

on the “intermediate” diverse library gives much better results than the application of GTM individually.

The first steps were made towards the modeling of the dynamic combinatorial libraries. To understand the behavior of a DCL, the relative propensity of the formation of DCL products and their relative affinities for the target are needed. The calculation of speciation of any solution does not present any technical difficulty due to numerous software, but its utilization remains constrained by the availability of experimental data. First, from the pool of 400 most cited aromatic aldehydes and 300 most cited primary amines (according to SciFinder), diverse subsets of 24 aldehydes and 15 amines (resulting in 276 imines) have been selected. These imines have been synthesized, their logK measured, and the obtained data served as a training set for building SVM classification and regression models. The obtained models have shown high data coverage of the initial pool of 120k imines (67% of the data are in the AD of the consensus model) as well as high predictive performance. In order to model the presence of the effector, it is needed to quantitatively know the affinity of each constituent in solution to it. For this purpose, the ChEMBL database has been used as a source of binding affinity (pK_i) on human Carbonic Anhydrase II. More than 4300 compounds have been extracted and used to train SVM regression models. Both regression models allow the overcoming of the initial “experimental data constrain” and thus make possible the usage of speciation software. The mutual usage of logK and pK_i regression models, as well as ChemEqui speciation software, lead to a possibility of modeling a hypothetical DCL containing 2 aldehydes, 2 amines and human CA II as an effector.

Perspectives

The candidates for DCL in the presence of biological target should be selected in such a way that (*i*) neither aldehydes nor amines efficiently interact with the effector and, (*ii*) only one imine firmly binds to the protein. For this purpose, a series of ligand-to-protein

docking calculations could be envisaged to gain a microscopic insight into imine-protein interactions.

Experimental measurements of DCL equilibria in aqueos solution are very welcome. They would help to build new predictive model for logK in water and to prove our suggested here protocol of logK rescaling from one solvent to another one.

8 References

1. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, and Overington JP (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100--D1107
2. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, and Nowotka M others (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930--D940
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
4. Bishop CM, Svensén M, and Williams CKI (1998) GTM: The generative topographic mapping. Neural Comput 10:215–234
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
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
7. Lehn J-M (1999) Dynamic combinatorial chemistry and virtual combinatorial libraries. Chem Eur J 5:2455–2463
8. Huc I, and Lehn J-M (1997) Virtual combinatorial libraries: dynamic generation of molecular and supramolecular diversity by self-assembly. Proc Natl Acad Sci 94:2106–2110
9. Holliday JD, and Willett P (1996) Definitions of" dissimilarity" for dissimilarity-based compound selection. J Biomol Screen 1:145–151
10. 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
11. Cortes C, and Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
12. Chang C-C, and Lin C-J (2011) {LIBSVM}: A library for support vector machines. ACM Trans Intell Syst Technol 2:27:1--27:27
13. Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci 2:559–572
14. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417
15. Akella LB, and DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330

16. Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, and Chen L (2008) Interface Foundation of America. J Comput Graph Stat 17:444–472

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

18. Oprea TI, and Gottfries J (2001) Chemography: The art of navigating in chemical space. J Comb Chem 3:157–166. https://doi.org/10.1021/cc0000388

19. Balakin K V (2009) Pharmaceutical Data Mining. John Wiley & Sons, Inc.

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

21. Neal RM (2007) Pattern Recognition and Machine Learning. Technometrics 49:366. https://doi.org/10.1198/tech.2007.s518

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

23. Bemis GW, and Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928

24. 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. ChemInform 38:. https://doi.org/10.1002/chin.200715213

25. 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

26. 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

27. Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS, Chang RSL, Lotti VJ, Cerino DJ, Chen TB, Kling PJ, Kunkel KA, Springer JP, and Hirshfield J (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 31:2235–2246. https://doi.org/10.1021/jm00120a002

28. 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

29. Kayastha S, Horvath D, Gilberg E, Gu?tschow M, Bajorath J, and Varnek A (2017) Privileged structural motif detection and analysis using generative topographic maps. J Chem Inf Model 57:1218–1232

30. Horvath D, Orlov A, Osolodkin D, Marcou G, Ishmukhametov AA, and Varnek A (2020) A Chemographic Audit of Anti-Coronavirus Structure-Activity Information from Public Databases (ChEMBL). https://doi.org/10.26434/chemrxiv.12104010.v1

31. 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

32. Erwin E, Obermayer K, and Schulten K (1992) Self-organizing maps: ordering, convergence properties and energy functions. Biol Cybern 67:47–55. https://doi.org/10.1007/bf00201801

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

34. Wold S, Sjöström M, and Eriksson L (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

35. Quinlan JR, and others (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. pp 343–348

36. 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

37. Horvath D, Baskin I, Marcou G, and Varnek A (2017) Generative Topographic Mapping of Conformational Space. Mol Inform 36:1700036. https://doi.org/10.1002/minf.201700036

38. 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. https://doi.org/10.1002/jcc.20035

39. Horvath D, Marcou G, and Varnek A (2017) Monitoring of the Conformational Space of Dipeptides by Generative Topographic Mapping. Mol Inform 37:1700115. https://doi.org/10.1002/minf.201700115

40. Horvath, Marcou, and Varnek (2019) Generative Topographic Mapping of the Docking Conformational Space. Molecules 24:2269. https://doi.org/10.3390/molecules24122269

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

42. 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. https://doi.org/10.1021/acs.jcim.5b00398

43. 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

44. Bort W, Baskin II, Sidorov P, Marcou G, Horvath D, Madzhidov T, Varnek A, Gimadiev T, Nugmanov R, and Mukanov A (2020) Discovery of Novel Chemical Reactions by Deep Generative Recurrent Neural Network. https://doi.org/10.26434/chemrxiv.11635929.v1

45. Lowe D (2017) Chemical reactions from US patents (1976-Sep2016). https://doi.org/10.6084/m9.figshare.5104873.v1

46. Sanchez R, Meslamani J, and Zhou M-M (2014) The bromodomain: from epigenome

reader to druggable target. Biochim Biophys Acta (BBA)-Gene Regul Mech 1839:676–685

47. Liu Z, Wang P, Chen H, Wold EA, Tian B, Brasier AR, and Zhou J (2017) Drug discovery targeting bromodomain-containing protein 4. J Med Chem 60:4533–4558
48. Enamine Ltd. https://enamine.net/
49. Wolber G, and Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160–169
50. Gerhard Wolber and Inte:Ligand GmbH (2017) LigandScout 4.1
51. 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
52. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, and others (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57:4977–5010
53. Todeschini R, and Consonni V (2008) Handbook of molecular descriptors. John Wiley & Sons
54. Varnek A, Fourches D, Hoonakker F, and Solov'ev VP (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J Comput Aided Mol Des 19:693–703
55. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko I V, and Marcou G (2008) ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 4:191
56. Ruggiu F, Marcou G, Solov'ev V, Horvath D, and Varnek A ISIDA Fragmentor2015-User Manual
57. Mitchell JBO (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481
58. Marill KA (2004) Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med 11:94–102
59. Hassoun MH, and others (1995) Fundamentals of artificial neural networks. MIT press
60. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32:241–254
61. Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf theory 28:129–137
62. Leach AR, and Gillet VJ (2007) An introduction to chemoinformatics. Springer
63. Golbraikh A, and Tropsha A (2002) Beware of q2! J Mol Graph Model 20:269–276
64. Browne MW (2000) Cross-validation methods. J Math Psychol 44:108–132
65. Sayad S (2020) An Introduction to Data Science: Suppor Vector Machine. https://www.saedsayad.com/support_vector_machine.htm. Accessed 3 Jun 2020

66. Drucker H, Burges CJC, Kaufman L, Smola AJ, and Vapnik V (1997) Support vector regression machines. In: Advances in neural information processing systems. pp 155–161

67. Gaspar HA, Sidorov P, Horvath D, Baskin II, Marcou G, and Varnek A (2016) Generative Topographic Mapping Approach to Chemical Space Analysis. In: Frontiers in Molecular Design and Chemical Information Science - Herman Skolnik Award Symposium 2015: Jürgen Bajorath. American Chemical Society, pp 211–241

68. Lin A, Baskin II, Marcou G, Horvath D, Beck B, and Varnek A (2020) Parallel Generative Topographic Mapping: an Efficient Approach for Big Data Handling. Mol Inform

69. 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

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

71. Reddy AS, Pati SP, Kumar PP, Pradeep HN, and Sastry GN (2007) Virtual screening in drug discovery-a computational perspective. Curr Protein Pept Sci 8:329–351

72. 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

73. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705

74. Jain AK, and Barton MC (2017) Bromodomain histone readers and cancer. J Mol Biol 429:2003–2010

75. Stathis A, and Bertoni F (2018) BET proteins as targets for anticancer treatment. Cancer Discov 8:24–36

76. Marcotte R, Sayad A, Brown KR, Sanchez-Garcia F, Reimand J, Haider M, Virtanen C, Bradner JE, Bader GD, Mills GB, and others (2016) Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell 164:293–309

77. French CA, Rahman S, Walsh EM, Kühnle S, Grayson AR, Lemieux ME, Grunfeld N, Rubin BP, Antonescu CR, Zhang S, and others (2014) NSD3--NUT fusion oncoprotein in NUT midline carcinoma: implications for a novel oncogenic mechanism. Cancer Discov 4:928–941

78. French CA, Ramirez CL, Kolmakova J, Hickman TT, Cameron MJ, Thyne ME, Kutok JL, Toretsky JA, Tadavarthy AK, Kees UR, and others (2008) BRD--NUT oncoproteins: a family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells. Oncogene 27:2237–2242

79. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, Morse EM, Keates T, Hickman TT, Felletar I, and others (2010) Selective inhibition of BET bromodomains. Nature 468:1067–1073

80. Sanchez R, and Zhou M-M (2009) The role of human bromodomains in chromatin biology and gene transcription. Curr Opin Drug Discov Devel 12:659

81. Ganellin C, Lindberg P, and Mitscher L (1998) Glossary of terms used in medicinal chemistry. Pure Appl Chem 70:1129–1143

82. Gasteiger J, and Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36:3219–3228

83. Borysko P, Moroz YS, Vasylchenko O V, Hurmach V V, Starodubtseva A, Stefanishena N, Nesteruk K, Zozulya S, Kondratov IS, and Grygorenko OO (2018) Straightforward hit identification approach in fragment-based discovery of bromodomain-containing protein 4 (BRD4) inhibitors. Bioorg Med Chem

84. Giuseppone N, and Lehn J-M (2006) Protonic and temperature modulation of constituent expression by component selection in a dynamic combinatorial library of imines. Chem Eur J 12:1715–1722

85. Men G, and Lehn J-M (2017) Higher order constitutional dynamic networks:[2x3] and [3x3] networks displaying multiple, synergistic and competitive hierarchical adaptation. J Am Chem Soc 139:2474–2483

86. Osypenko A, Dhers S, and Lehn J-M (2019) Pattern generation and information transfer through a liquid/liquid interface in 3D constitutional dynamic networks of imine ligands in response to metal cation effectors. J Am Chem Soc 141:12724–12737

87. Clark RD (1997) OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. J Chem Inf Comput Sci 37:1181–1188

88. 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. https://doi.org/10.1021/cc020092j

89. Schneider P, Tanrikulu Y, and Schneider G (2009) Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. Curr Med Chem 16:258–266

90. Selzer P, and Ertl P (2006) Applications of self-organizing neural networks in virtual screening and diversity selection. J Chem Inf Model 46:2319–2323

91. Holliday JD, Ranade SS, and Willett P (1995) A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. Quant Struct Relationships 14:501–506

92. Gower JC (1985) Measures of similarity, dissimilarity and distance. Encycl Stat Sci Johnson CB Read 5:397–405

93. Willett P, Barnard JM, and Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996

94. Lipkus AH (1999) A proof of the triangle inequality for the Tanimoto distance. J Math Chem 26:263–265

95. Wavefunction Inc, and Https://www.wavefun.com/ Spartan 18

96. Solov'ev VP, and Tsivadze AY (2015) Supramolecular complexes: Determination of stability constants on the basis of various experimental methods. Prot Met Phys Chem Surfaces 51:1–35

Iuri CASCIUC

Cartographie chimique et modélisation de systèmes complexes

Résumé

Cette thèse concerne l'utilisation de la Cartographie Topographique Générative (Generative Topographic Mapping – GTM) pour l'analyse et la visualisation de jeux de données, le criblage virtuel et la conception de chimiothèques. La performance en criblage virtuel de modèles GTM de classification multi-cibles (uGTM) a été étudiée et l'utilisation de plusieurs uGTM en consensus a été proposée. Un criblage virtuel utilisant une combinaison de la GTM avec d'autres techniques de chémoinformatique a permis de découvrir 29 nouveaux inhibiteurs de BRD4 dont l'activité a été prouvée expérimentalement. La GTM a été comparée à la méthode MaxMin comme outil de conception de chimiothèques. Il a été trouvé que malgré le fait que les chimiothèques obtenues avec MaxMin sont plus diverses que celles obtenues avec la GTM, cette dernière est plus rapide et peut être appliquée à des jeux de données plus larges. Un protocole de modélisation pour l'analyse de spéciation de bibliothèques combinatoires dynamiques basées sur la réaction de formation d'imines en absence et en présence d'une protéine a été proposé. Les modèles développés sont disponibles au public sur le site de Laboratoire de Chémoinformatique.

Mots-clés : GTM, QSAR ; criblage virtuel, visualisation de données, conception de chimiothèques, Bibliothèques Combinatoires Dynamiques.

Résumé en anglais

This work concerns application of Generative Topographic Mapping method to different tasks including data analysis and visualization, virtual screening and library design. Performance of multi-target GTM-based classification models (uGTM) in virtual screening was investigated and consensus usage of several uGTMs has been suggested. Virtual screening involving a combination of GTM with some other chemoinformatics techniques allowed to discover 29 new BRD4 inhibitors, activities of which were experimentally confirmed. As a library design tool, GTM was compared to the MaxMin method. Although diversity of MaxMin libraries is systematically larger than those obtained with GTM, the latter is much faster and, therefore, can be recommended for large datasets. A modeling workflow for speciation analysis in imine-based Dynamic Combinatorial Libraries in absence and presence of a protein has been suggested. Developed models are publicly available at the site of the Laboratory of Chemoinformatics.

Key words: GTM, QSAR, virtual screening, data visualization, library design, Dynamic Combinatorial Libraries