Université des Sciences et Technologies - Lille 1

Les triplets pharmacophoriques flous Développement et applications

Thèse présentée pour obtenir le grade de

Docteur de l'Université de Lille 1

par

Fanny BONACHERA

Discipline :

Biomolécules, pharmacologie, thérapeutique

Spécialité :

Chimie théorique, physique et analytique

soutenue le 12 décembre 2011 devant le jury composé de

Pr. Luc MORIN-ALLORY	Orléans	Rapporteur
Dr. Michel PETITJEAN	Paris	Rapporteur
Dr. Jean-Claude MICHALSKI	Villeneuve d'Ascq	Examinateur
Dr. Dragos HORVATH	Strasbourg	Co-Directeur de thèse
Dr. Guy LIPPENS	Villeneuve d'Ascq	Directeur de thèse

Année 2011 Ecole Doctorale Biologie Santé

Remerciements

Je tiens à remercier le Dr. **Guy Lippens** pour m'avoir donné l'opportunité de m'inscrire en doctorat, ainsi que pour sa compréhension, son soutien et la grande liberté qu'il m'a laissée pour accomplir mes travaux.

Je voudrais remercier ensuite le Dr. **Dragos Horvath** pour avoir toujours été présent, disponible et toujours prêt à m'aider quel que soit le sujet sur lequel nous avons travaillé. De même, je souhaiterais remercier le Pr. **Alexandre Varnek** pour m'avoir accueillie en stage dans son laboratoire pour ma dernière année de thèse. Leurs conseils et leur soutien m'ont été précieux.

Ensuite, je voudrais adresser ma sincère reconnaissance au Pr. **Luc Morin-Allory**, au Dr. **Michel Petitjean** ainsi qu'au Dr. **Jean-Claude Michalski** pour avoir accepté de juger mon travail et de faire partie du jury de thèse.

Puis, j'aimerais remercier le Dr. **Gilles Marcou** pour m'avoir conseillée, guidée et aiguillée de nombreuses fois, ainsi que pour nos discussion très intéressantes.

Je tiens également à remercier tous les collaborateurs et les collègues avec qui nous avons travaillé. Plus particulièrement, mes collègues chémoinformaticiens et post-doctorants **Frank Honnakker**, **Igor Baskin**, **Vitaly Solov'yev**, **Natalia Kireeva**, **Vladminir Chupakhin** et mes collègues doctorants **Aurélie De Luca**, **Fiorella Ruggiu**, **Ioana Oprisiu**, **Christophe Muller**, **Laurent Hoffer**, **Tetiana Khristova**, **Evgeny Kondratovich**, pour leur sympathie, leur bonne humeur et leur intérêt scientifique.

Merci à tous les collègues qui m'ont proposé leur aide et leurs conseils durant cette thèse, que ce soit au niveau administratif (je pense particulièrement à **Gaëlle Vanstaevel**, **Martine Ratajczak**, **Sandrine Garcin** et **Danièle Ludwig**) ou au niveau scientifique et humain (**Natalie Sibille**, **Chrystelle Le Danvic**, **Emmanuel Maes**, **Christophe Biot**, **Xavier Trivelli**,**Yann Guérardel**, **Patricia Nagnan-Le Meillour**).

Enfin, mes pensées vont tout naturellement à mon compagnon, ma famille et mes amis.

Les triplets pharmacophoriques flous Développement et applications Résumé

L'un des challenges de la chémoinformatique est d'être capable de décrire de manière simple des composés afin de pouvoir les utiliser dans des études de similarité (pour trouver de nouveaux composés potentiellement intéressants) ou de pouvoir prédire leur activité en se basant sur les informations contenues dans les composés déjà connus.

Une des façons de décrire les molécules consiste à retranscrire de façon chiffrée leurs caractéristiques. Ceci permet notamment de pouvoir comparer *in silico* deux molécules entre elles. Il existe de nombreux descripteurs (bidimensionnels, tridimensionnels, se basant uniquement sur certains types d'informations contenues dans les molécules, ...). Cependant, il est important de prendre en compte non seulement les informations structurales des composés mais aussi leurs propriétés chimiques. Ainsi, nous avons développé des descripteurs combinant ces deux types d'informations en apportant des améliorations ayant du sens chimiquement parlant. Ce manuscrit vise à expliquer la mise en place de ces descripteurs puis à montrer leur efficacité dans différents types d'utilisations.

Les triplets pharmacophoriques flous se basent sur l'énumération en trois points de caractéristiques pharmacophoriques, combinée à la prise en compte des distances topologiques entre chacune de ces caractéristiques.

Ainsi, ces descripteurs représentent deux types particuliers d'informations contenues dans les molécules de manière simple :

- La distance topologique entre les atomes (nombre de liaisons interposées entre deux atomes),
- Le type pharmacophorique de chaque atome (6 types pharmacophoriques différents sont pris en compte : cations, anions, hydrophobes, aromatiques, donneurs et accepteurs de liaisons hydrogène).

Ces deux types d'informations sont réunis par 3 (dans un triplet pharmacophorique de type $T1.D2,3-T2.D1,3-T3.D1,2$ où Tn correspond au type pharmacophorique de l'atome n et $D(n,\ m)$ correspond à la distance topologique entre les atomes n et m) et sont énumérés dans une molécule.

Outre l'énumération de caractéristiques, les 2D-FPTs (triplets pharmacophoriques flous bidimensionnels) introduisent 2 améliorations : La projection floue des triplets d'atomes sur les triplets pharmacophoriques de base (permettant de mimer la tolérance naturelle des récepteurs par rapport à leurs ligands), et le marquage par pharmacophores dépendants du pK_a (permettant de prendre en compte l'équilibre protéolytique).

De plus, une nouvelle formule de calcul de similarité est introduite, qui prend en compte l'absence simultanée d'un triplet comme moins contraignante et probante qu'une présence simultanée.

Le développement des triplets est détaillé dans une première publication. Puis, plusieurs applications des triplets sont étudiées.

Les triplets sont tout d'abord utilisés dans des études de relation structure-activité ($QSAR$). Vue leur haute dimensionalité, la sélection des éléments pertinents pour des corrélations structure-activité nous a incité à mettre en place une nouvelle méthode de sélection : le SQS (Stochastic $QSAR$ Sampler). Cet échantillonneur $QSAR$ basé sur un algorithme génétique original, piloté par un méta-algorithme génétique est comparé au SR (outil de régression pas à pas) afin de sélectionner les meilleurs ensembles de descripteurs sur 3 jeux de données de petite taille. Les 2D-FPTs montrent de bons résultats, bien que la propension à conduire à des équations validantes est liée au jeu de données. En effet, les 2D-FPTs sont plus efficaces pour construire certains types de modèles, ce qui est expliqué par leur faculté à encoder certains types d'informations chimiques.

Une grande étude de performance a été réalisée à la suite de ces observations. L'utilisation des FPTs dans des études de $QSAR$ a donc été étudiée plus en profondeur, en comparaison avec des descripteurs déjà existants sur un ensemble de 11 jeux de données provenant de la littérature. De plus, nous avons comparé différentes versions de triplets entre elles. Les études $QSAR$ basés sur les 2D-FPTs se sont extrêmement bien déroulées tout au long de cet exercice de performance. Les modèles basés sur les FPTs ont égalisé, voire dépassé significativement les modèles publiés basés sur l'index 2D et 3D, sauf dans les cas où la nature même des composés n'est pas compatible avec ce type de descripteurs. Le flou optimal à appliquer lors de la projection des triplets trouvés sur les triplets de base ainsi que la représentativité des descripteurs sont étudiés. De plus, nous avons donné quelques pistes pour faire face aux artefacts inévitables avec les jeux de données étudiés.

Enfin, nous avons utilisé les FPTs pour construire des cartes auto-organisatrices (SOMs) afin de les utiliser pour accélérer les recherches par similarité dans une base de données. Un ensemble de cartes a été construit sur deux jeux de données d'entrainement différents en variant les paramètres d'entrainement ainsi que les paramètres de taille, de topologie et de fonctions de voisinage. Chacun de ces ensembles a été décrit par les 2D-FPTs. Ces cartes servent à projeter des molécules organiques dans cet espace chimique en assignant chaque composé sur un neurone dit "de résidence".

Ces cartes serviront à focaliser la recherche par similarité : une requête pharmacophorique est d'abord positionnée sur la carte et elle sera comparée explicitement uniquement avec les molécules résidentes dans son nœud ou dans les nœuds voisins, limitant ainsi l'effort de calcul. Les composés qui se trouvent en dehors du voisinage sont considérés comme implicitement "dissimilaires", ce qui a un impact sur le taux de récupération des ensembles initiaux de *Hits* virtuels. Un compromis est donc à trouver entre accélération et perte de *Hits* virtuels.

Un critère conciliant ces tendances opposées a donc été défini, afin de caractériser l'accélération par la SOM contre l'efficacité du taux de récupération. Ce critère simple permet la comparaison des performances relatives de toutes les cartes créées. Nous avons grâce à ce critère étudié l'impact des choix de construction des cartes (taille de l'ensemble d'entrainement, taille et géométrie des cartes, critère de convergence imposé, choix des fonctions de voisinage) sur leur efficacité.

Il est démontré dans ces travaux qu'augmenter la taille du set d'entrainement au delà d'une certaine limite se fait au détriment de la qualité : trop de composés d'entrainement entrainent des problèmes de convergence. De plus, l'impact de l'entrainement est analysé en profondeur et il est montré qu'il est aussi important de bien entrainer une carte que de bien choisir l'ensemble sur lequel cet entrainement est fait.

La meilleure carte ressortant des comparaisons est décryptée et présentée dans ces travaux. Enfin, les tests grandeur nature sur 4 cartes sélectionnées pour leurs bons résultats démontrent que notre critère permet de prédire le comportement des cartes sur de nouveaux jeux de données.

Au delà des résultats publiés, les 2D-FPTs ont été adoptés par des partenaires industriels dans le cadre de collaborations et ont effectivement servi à la découverte de nouveaux composés bioactifs. (Résultats non publiables).

Fuzzy pharmacophores triplets developement and applications. Summary

One of the challenges faced by Chemoinformatics is the following : to be able to describe compounds in a simple way, in order to use them in similarity studies (to discover new druglike compounds) or to predict their activity, based on informations contained in already known compounds.

One of the possible ways to describe a compound is to encode its features as numeric data. This allows *in silico* comparisons (ie calculating distance between two vectors). There are a lot of existing descriptors (bidimensional, tridimensional, based only on certain type of features,...). However, it is important to take into account not only structural information but also chemical properties. Thus, we developed our descriptors, combining these two types of informations and adding chemically-relevant improvements. This work explains the creation of these descriptors as well as their use in different type of applications.

The fuzzy tricentric pharmacophores are based on the enumeration of 3 pharmacophoric features points, combined with the topological distances between each of these features. These descriptors represent two particular types of informations contained in any compound in a simple way :

- the topological distance between atoms (number of interposed bonds)
- the pharmacophoric type of each atom (6 different pharmacophoric types are taken into account : cations, anions, hydrophobic, aromatic, H-bonds donors and acceptors).

These informations are reunited by 3 (in a pharmacophoric triplet $T1.D2,3-T2.D1,3-T3.D1,2$ where Tn corresponds to the atom n pharmacophoric type and $D(n,m)$ is the topological distance between atoms n and m) and are enumerated in a molecule.

Besides features enumeration, the 2D-FPTs (bidimensional fuzzy pharmacophoric triplets) introduce two improvements : the fuzzy mapping of molecular triplets on basis pharmacophoric triplets (this mimics the natural tolerance of receptors towards their ligands), and pK_a-dependant pharmacophore flagging (which takes into account the proteolytic equilibrium).

Moreover, a new similarity calculation formula is introduced, which accounts for the simultaneus absence of a triplet as a less-constraining indicator of similarity than its simultaneous presence.

The fuzzy triplets development is detailed in a first publication. Then, several applications are studied.

The FPTs are first used in quantitative structure-activity relationship studies (QSAR). Considering their high dimensionality, the selection of relevant elements for structure-activity correlation prompted us to develop a new selection method : the SQS (Stochastic QSAR Sampler), a QSAR sampler based on an original genetic algorithm driven by a genetic meta-algorithm, is compared to the SR (Stepwise Regression) tool, in order to select the best matches of descriptors on 3 small data sets. The FPts shows good results even if their propensity to lead to validating equations is related to the data set. Indeed, the FPTs are more effective when it comes to build certain types of models. This is explained by their faculty to encode some types of chemical informations. A big performance study was conducted after these observations.

The use of FPTs in QSAR studies was deeply examined and compared with existing descriptors on an ensemble of 11 data sets from litterature. We also compared different triplets versions. The QSAR studies based on FPTs went extremely well in this benchmark. The models based on FPTs were as good as, and sometimes even significantly better than published models based on 2D and 3D index, except in the case where the intrisec compounds nature was not compatible with this kind of descriptors. The optimal fuzziness that should be used when mapping molecular triplets on basis triplets was studied, as well as the descriptors representativity. We also gave some advice in order to deal with inevitable artefacts found within the studied data sets.

In our last study, we used FPTs to build self-organizing maps (SOM), in order to use them as an attempt to accelerate similarity searches in a database. Many maps were built on two different data sets, varying the training parameters as well as size, topology and neighbourhood functions. Each of these data set was exclusively described with FPTs. These maps are used to map organic compounds in this chemical space and each compound is assigned to its "resident" neuron.

These maps were used to focus the similarity search : a pharmacophoric query is first positioned on the map and will explicitely be compared only with compounds located on its neuron or on neighbouring neurons, which limitates the computational effort. The compounds that are outside the neighbourhood are implicitely considered as "dissimilars", which impacts the retrieval rate of initial virtual Hits ensembles. A compromise must then be found between acceleration and virtual Hits loss.

A criterion, conciliating these two opposed tendencies has been defined in order to caracterize the SOM speed-up against retrieval rate. This simple criterion allows the comparison of relative performances of all created maps. We studied the impact of the building choices (training set size, size and geometry of maps, imposed convergence criterion, neighbourhood functions impact) on their efficacy.

It is shown that increasing the size of the training set beyond a certain limit becomes detrimental to map quality. Moreover, the training impact is analyzed and it is shown that it is as important to properly train the map than to carefully choose the training set. The best map is presented and analyzed. Then, real-life tests conducted on 4 maps selected because of their good results show that our criterion allows to correctly predict maps behaviour on new data sets.

Next to published results, FPTs have been adopted by industrial partners in collaborations and have been used to discover new bioactive compounds (Results non publishable).

Mots-Clés

Descripteurs moléculaire, pharmacophores, chémoinformatique, similarité, QSAR, criblage virtuel, logique floue, drug design, scaffold hopping.

Keywords

Molecular descriptors, pharmacophores, chemoinformatics, similarity, QSAR, virtual screening, fuzzy logic, drug design, scaffold hopping.

Laboratoires de rattachement

Equipe RMN et Modélisation
Unité de Glycobiologie Structurale et Fonctionnelle - UMR 8576
Université des Sciences et Technologies de Lille (Université Lille I)
Bâtiment C9
59655 Villeneuve d'Ascq Cedex
France

Dernière année de thèse en stage dans le

Laboratoire d'Infochimie
Institut de Chimie de Strasbourg - UMR 7177
Université de Strasbourg
1, rue Blaise Pascal
67000 Strasbourg
France

Table des matières

Première partie

Introduction

1 Généralités sur la Chémoinformatique.

Introduit à la fin des années 1990, le terme de "Chémoinformatique" est apparu afin de décrire l'utilisation en plein essor de l'informatique pour résoudre des problèmes chimiques. L'utilisation de l'outil informatique est en effet devenue évidente pour manipuler les informations structurales des molécules qui ont été au cours des dernières années stockées de manière numérique. De plus, la multiplication des données exploitables par les chimistes a donné lieu à une obligation de numérisation, afin d'être capable de stocker, visualiser et traiter ces mêmes données.

La Chémoinformatique traite de nombreux problèmes à la fois dans le domaine chimique mais aussi en biologie. Ses utilisations sont très variées et vont de la création et l'utilisation de base de données de petites molécules à la manipulation de fichiers en passant par les études statistiques. Cependant, son application la plus communément admise est dans le domaine de la recherche de nouveaux médicaments (*drug discovery*), domaine dans lequel elle joue un rôle central dans l'analyse et l'interprétation des données de structures et de propriétés collectées au cours des criblages à haut débit (technique visant à identifier des molécules nouvelles et potentiellement actives dans des bases de données de composés).

Les définitions les plus populaires et les plus anciennes de la Chémoinformatique (ou Chéminformatique) sont les suivantes :

The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.[11]

Chem(o)informatics is a generic term that encompasses the design, creation, organisation, management, retrieval, analysis, dissemination, visualization and use of chemical information.[93]

L'émergence de la Chémoinformatique peut être mise en parallèle avec la multiplication des données chimiques stockées numériquement. En effet, les quantités de données générées par les nouvelles approches de *drug design* n'ont eu de cesse d'augmenter et il s'est avéré nécessaire, pour traiter les résultats de criblage à haut débit ou encore de la chimie combinatoire, de développer et de d'utiliser des techniques informatiques. [76]

Les avancées technologiques des dix dernières années ont rendu possibles de nombreuses découvertes et applications inaccessibles auparavant. Par exemple, le nombre de composés disponibles dans les études de criblage a augmenté de manière exponentielle. En parallèle, les développements techniques dans le domaine de l'informatique et des technologies de communication ont permis la création de bases de données de composés comportant des millions d'entrées.

Un exemple parfait pour illustrer ces avancées est la base de données PubChem développée par le NIH. [10] Avec un contenu de plus de 31 millions de composés reliés à leur activité biologique, ce genre de bases de données nécessite le développement et l'utilisation d'outils mathématiques et statistiques afin de pouvoir accéder à de nouvelles découvertes en termes de développement de nouveaux médicaments et à la compréhension des relations entre structure et activité.

Bien que certaines des techniques utilisées en Chémoinformatique sont basées sur des concepts établis depuis de nombreuses années, [40], il est évident qu'elle reste un domaine en plein essor et surtout très vaste.

Nous nous focaliserons dans ce manuscrit sur l'utilisation de descripteurs spécifiques à 2 dimensions (les *Fuzzy Pharmacophore Triplets : 2D-FPT*) pour aider au criblage virtuel basé sur la structure du ligand (*Ligand-based virtual screening*), ainsi que pour la mise en place d'études de relations Structure-Activité (*QSAR – Quantitative Structure-Activity Relationships*). Ces descripteurs seront aussi utilisés afin de construire des cartes auto-organisatrices (*SOM - Self Organizing Maps*) dans le but d'accélérer les recherches par similarité dans des bases de données.

Nous allons d'abord définir ce qu'est le Criblage Virtuel, puis nous nous intéresserons plus en détail aux méthodes de criblage virtuel par similarité basées sur les descripteurs.

Ensuite, nous définirons les études *QSAR*. Nous introduiront également de manière succincte le principe des cartes auto-organisatrices, utilisées dans les les travaux du dernier chapitre de ce manuscrit. Pour terminer cette introduction, nous accorderons une attention particulière aux descripteurs topologiques pharmacophoriques, afin de dresser un état de l'art dans ce domaine et de placer nos descripteurs spécifiques par rapport à l'ensemble des descripteurs déjà existants.

2 Le Criblage virtuel

Le fait d'avoir accès à des bases de données de composés de plus en plus fournies a entrainé une nécessité d'utiliser des outils informatiques afin d'identifier les candidats potentiels, ce qui permet de réduire les tests in vitro en n'envoyant en laboratoire que les composés susceptibles de se lier à la cible d'intérêt.

Devenu ainsi depuis une dizaine d'année une partie intégrante du processus de recherche de nouvelles molécules bioactives, le Criblage Virtuel, tel qu'il a été défini en 1998 par Walters, et al. [92],

automatically evaluating very large libraries of compounds.

désigne l'action de rechercher à l'aide de programmes informatiques (*in silico*), grâce à de larges bases de données, des *Hits* virtuels pouvant selon la prédiction se lier à des cibles macromoléculaires ayant un intérêt pharmaceutique, ou posséder les propriétés souhaitées. L'intérêt du Criblage Virtuel est de permettre de découvrir des composés "nouveaux", dans le sens où certains composés ayant des structures inhabituelles par rapport aux ligands communément utilisés peuvent être mis à jour au cours du criblage. Ces composés, suffisamment différents des composés requêtes, peuvent potentiellement être considérés comme une nouvelle classe d'agents thérapeutiques.

Les *Hits* choisis sont envoyés pour être testés biologiquement et découvrir si de nouveaux composés actifs sont identifiés parmi eux.

Le Criblage Virtuel est habituellement divisé en 2 sous-catégories : Le Criblage virtuel basé sur la structure de la cible et le Criblage Virtuel *ligand-based* (basé sur la structure du ligand).

2.1 Le Criblage Virtuel basé sur la structure de la cible

Le Criblage Virtuel basé sur la structure de la cible utilise la structure tri-dimensionnelle de la macromolécule cible pour découvrir des petites molécules capables de s'y accrocher et prédire leur affinité de liaison (*scoring*). Ce processus, appelé le *docking*, peut être efficace pour trouver des candidats mais reste compliqué à mettre en place pour des études impliquant des millions de ligands potentiels. [54, 62]

Le processus de *docking* cherche à prédire l'orientation et la conformation d'un ligand dans le site de liaison d'une macromolécule cible – puis, à partir de ceci, calcule l'affinité estimée du ligand pour la cible, justifiant ou pas de considérer ce ligand comme un *hit* potentiel. [56]

2.2 Le Criblage Virtuel basé sur la structure du ligand

Lorsque la structure de la cible n'est pas connue, une autre méthode de Criblage Virtuel est appliquée afin de déterminer des candidats. Cette méthode, basée sur la structure et sur diverses caractéristiques des ligands connus de la cible, implique de cribler des bases de données de composés en utilisant ces informations comme requêtes. [27, 84]

Il existe plusieurs manières de mettre en place un Criblage Virtuel basé sur la structure du ligand. Nous pouvons ainsi distinguer ces méthodes :

2.2.1 Méthodes dites "locales"

basées sur des caractéristiques pré-définies comme étant déterminantes pour l'activité biologique. Auparavant, il est donc nécessaire de savoir quels sont les sous-ensembles de la structure globale qui impactent sur l'activité – information à extraire à partir d'exemples d'actifs et d'inactifs déjà connus.

- Recherche via des modèles *QSAR* (Relation Structure-Activité) : L'analyse *QSAR* est capable d'extraire l'information sur l'activité biologique d'une molécule à partir d'un ensemble de molécules de référence. (voir Chapitre 3.2.1)
- Recherche par sous-structure : La recherche par sous-structure renvoie tous les composés contenant la sous-structure requête. Cette méthode est basée uniquement sur la comparaison des composés en prenant en compte leur structure. C'est un cas particulier de *(Q)SAR*, basé sur un modèle postulant que l'activité est conditionnée par la présence d'un fragment spécifique – vu dans toutes les structures d'actifs connus, et absent dans les inactifs. Note : une recherche sous-structurale est souvent utilisée dans un contexte de planning synthétique : pour faire des amides, il faut chercher acides "-COOH" et amines "N substitué par carbones saturés" dans la base – ici, le modèle structure-activité n'est pas conçu spécifiquement pour ce problème, mais il est appris en cours de chimie organique.
- Recherche via des pharmacophores 3D : On utilise lors d'une recherche via des pharmacophores 3D un ensemble de conformations 3D obtenues à partir de molécules actives et inactive de référence. Il est possible de déterminer à partir de cet ensemble quelles sont les parties des molécules qui interagissent avec la cible – ce qui, encore une fois, relève du domaine du *(Q)SAR*. L'arrangement dans l'espace de ces points d'interaction est utilisé comme requête dans la base de données et les molécules trouvées dans la base de données qui présentent des arrangements similaires au pharmacophore sont sélectionnées comme candidates.

2.2.2 Methodes dites “globales”

qui prennent en compte la structure moléculaire complète (car on ne sait pas où là-dedans se cachent les vrais points-clés définissant l'activité). Elles se basent sur le principe de la similitude : “des molécules similaires ont une plus forte chance de présenter des activités similaires ” (par rapport à n'importe que paire de composés choisis aléatoirement). Les *Hits*, dans ce contexte-là, seront alors les molécules les plus similaires à la structure d'un actif connu, le composé-requête. Il reste à définir sur quelle base cette similitude sera évaluée :

- Recherche basée sur les graphes : Dans cette méthode, les molécules sont représentées comme des graphes, c'est à dire des ensembles de nœuds (les atomes) reliés entre eux par des arcs (les liaisons). La recherche peut ainsi se faire par sous-graphes, en utilisant des parties de la molécule comme requête ou par graphe complet. On va donc rechercher la meilleure correspondance entre les atomes et les liaisons de la requête par rapport aux atomes et aux liaisons des composés de la base de données (c'est à dire le sous-graphe connexe commun maximal, contenant le plus de nœuds). Cette méthode peut prendre ou non en compte les propriété physico-chimiques des atomes et des liaisons.
- Recherche basée sur la superposition : Cette technique essaye de superposer une molécule sur une autre. Pour les graphes moléculaires, cette superposition, improprement qualifiée de "2D" se fait en cherchant la correspondance entre les atomes de la molécule A et les atomes de la molécule B. En 3 dimensions, elle implique de trouver la meilleure superposition entre les deux objets tridimensionnels, en se basant soit sur les distances entre atomes, soit sur une distance par exemple entre les champs entourant les atomes. Il existe de nombreuses méthodes de superposition. [80, 2, 17]
- Recherche basée sur des descripteurs : Cette méthode considère les molécules comme un ensemble de descripteurs (habituellement chiffrés), représentant des propriétés structurales ou physico-chimiques. Ainsi, une molécule est considérée comme étant un point dans un espace multidimensionnel de descripteurs – indices topologiques, propriétés physico-chimiques calculées, histogrammes de distribution des différentes propriétés locales (comptage de fragments, de multiplets d'atomes, etc). On détermine la similarité entre 2 molécules grâce à des fonctions de *scoring*. Ce type de méthode sera décrit plus en détail dans la section suivante.

3 Recherche par similarité basée sur les descripteurs

Comme indiqué précédemment, contrairement aux études de $QSAR$ ou aux études basées sur les pharmacophores, la recherche par similarité requiert une représentation globale des molécules qui prenne en compte leur structure complète et non pas uniquement certaines caractéristiques.

Lors d'une recherche par similarité, on évalue la ressemblance entre le composé requête et les éléments d'une base de données. Une fois la similarité évaluée grâce à différentes distances ou indices de dissimilarité, les composés sont classés en fonction des résultats obtenus. Partant toujours du même postulat qu'une relation existe entre structure et activité, les composés les plus similaires à la requête seront considérés comme ayant le plus de chance de partager son activité biologique.

Avant même de lancer une recherche de similarité basée sur les descripteurs, il est bien entendu nécessaire de mettre en place ces descripteurs moléculaires pour tous les composés de la base de données sur laquelle nous souhaitons lancer la recherche.

3.1 Définition de l'espace de référence et choix des descripteurs

3.1.1 L'espace de référence

Afin de pouvoir définir et analyser la similarité entre des molécules, il est nécessaire d'établir un cadre de référence qui relie les structures moléculaires entre elles et facilite les calculs de comparaisons.

Ainsi, un aspect très important de la recherche par similarité basée sur des descripteurs est la définition de l'espace chimique théorique dans lequel les composés à analyser seront projetés. Un espace chimique théorique est décrit par un ensemble de descripteurs moléculaires, chaque descripteur rajoutant une dimension à l'espace. Les molécules sont placées dans l'espace de référence par rapport aux valeurs de leurs descripteurs (les "coordonnées" correspondant aux valeurs prises par les descripteurs). Une bonne représentation de l'espace chimique devrait projeter des composés similaires dans des régions proches les unes des autres. De cette manière il est possible de calculer la similarité entre 2 molécules : c'est la distance inter moléculaire dans l'espace de référence.

Le plus grand challenge dans la mise en place de l'espace chimique est de choisir les représentations moléculaires de manière à ce que des activités biologiques proches ou des propriétés similaires soient reflétées par des distances inter moléculaires faibles.

3.1.2 Les descripteurs

Plusieurs type de descripteurs moléculaires sont généralement utilisés.

De manière simple, un descripteur moléculaire est une représentation mathématique d'une molécule, qui contient à la fois des informations sur la structure, et donc, implicitement ou explicitement, sur ses propriétés physico-chimiques. Cette information peut être encodée par des valeurs scalaires, des vecteurs ou des chaines de bits.

Les descripteurs sont fréquemment classés par rapport à la dimensionalité de la représentation moléculaire sur laquelle ils sont calculés : On parlera alors de descripteurs 1D, 2D, ou 3D. [25]

Les descripteurs 1D Ces descripteurs sont appelés "descripteurs constitutionnels" et sont faciles et rapides à calculer. Ils sont basés sur la formule brute de la molécule et contiennent des informations simples comme le poids moléculaire ou le nombre d'atomes.

Les descripteurs 2D Les descripteurs qui utilisent la représentation des molécules en tant que graphes sont dits 2D et contiennent, par exemple, des informations à propos de la connectivité ou à propos de certains fragments moléculaires, mais aussi des estimations des propriétés physicochimiques. C'est à partir de ce niveau que l'on peut espérer la capture d'informations chimiques pertinentes pour la prédiction de la majorité des propriétés moléculaires.

On trouvera dans cette catégorie les descripteurs suivants :

- Les descripteurs topologiques, qui considèrent la structure du composé comme un graphe, les atomes étant les nœuds et les liaisons les arcs. De nombreux indices quantifiant la connectivité moléculaire ont été développés en se basant sur cette approche, comme par exemple l'indice de Wiener [95], qui compte le nombre total de liaisons dans les chemins les plus courts entre toutes les paires d'atomes (en excluant les hydrogènes). D'autres indices basés sur les chemins ont été développés [74, 3, 78]. Les informations sur les électrons de valence peuvent être inclues dans les descripteurs topologiques [53, 33]. Enfin, des descripteurs combinant les informations de connectivité avec d'autres propriétés sont aussi à disposition, comme par exemple les descripteurs BCUT, qui se présentent sous la forme de matrices de connectivités des atomes, avec sur la diagonale la charge atomique, la polarisabilité ou les valeurs du potentiel de liaisons hydrogènes, et des termes additionnels hors diagonale. [12, 85]

– Les descripteurs basés sur les fragments se basent sur des motifs sous-structuraux. Par exemple, les empreintes BCI [4] sont des ensembles de bits indiquant la présence ou l'absence de certains fragments dans une molécule. Les fragments prennent en compte les atomes et leurs plus proches voisins, les paires d'atomes et les séquences ou encore les fragments basés sur des cycles. L'approche des clés MDL est une approche similaire comprenant la recherche des 166 fragments MDL [26]

Les descripteurs 3D Les descripteurs 3D comme le volume moléculaire, la surface ou les motifs pharmacophoriques requièrent une conformation 3D de la molécule expérimentale ou prédite. De plus, l'information sur la structure de la cible (protéine) est parfois requise. On pourra ainsi distinguer les descripteurs 3D qui nécessitent un alignement de la molécule guidé par l'étude des complexes ligand-cible (ou, au moins, par des contraintes visant d'optimiser le recouvrement spatial des champs électriques et stériques des ligands, faute d'information sur le vrai mode de fixation dans la cible) avant d'être calculés, comme par exemple les descripteurs CoMFA [21].

– Les descripteurs géométriques sont basés sur l'arrangement spatial des atomes constituant la molécule. Ces descripteurs incluent des informations sur la surface moléculaire obtenue par les aires de Van Der Waals et leur superposition [58]. Les volumes moléculaires peuvent être obtenus par les volumes de Van Der Waals [42]. L'information sur l'arrangement spatial des atomes dans la molécule peut aussi être obtenue par les moments principaux de l'inertie et les indices gravitationnels [52]. Un autre descripteur géométrique est la surface totale accessible au solvant. [99, 94].
– Les empreintes moléculaires (*fingerprints*) sont des descripteurs beaucoup utilisés dans les recherches par similarité. Ces empreintes sont des représentations des structures moléculaires sous formes de chaines de cases, chaque case contenant02 un descripteur scalaire. Le plus souvent, les empreintes sont des chaînes de bits dans lesquelles l'information moléculaire est contenue sous format binaire. En parallèle des simples empreintes 2D, il existe certains types d'empreintes utilisant l'information contenue dans la géométrie 3D. C'est le cas de certains types d'empreintes pharmacophoriques, qui proposent tous les motifs pharmacophoriques possibles dans les conformères d'une molécule. Nous reviendrons sur les empreintes pharmacophoriques plus tard (voir Chapitre 9.1).
– Les descripteurs électrostatiques et quantiques prennent en compte l'information contenue dans la nature électronique de la molécule. On trouvera ainsi des descripteurs fondés sur les charges partielles [67], les charges positives et négatives ainsi que sur la polarisabilité moléculaire [13]. Les surfaces accessibles au solvant partiellement chargées positivement ou négativement sont aussi utilisées comme descripteurs électrostatiques dans le cas de la modélisation des liaisons hydrogènes intermoléculaires [86]. Enfin, on trouvera aussi des descripteurs basés sur les énergies des orbitales moléculaires. [55, 100]

D'autres descripteurs ne prennent pas en compte la position de la molécule dans l'espace. Ainsi, il n'est pas nécessaire de faire un alignement pour comparer deux molécules. C'est le cas des descripteurs CoMMA [81] ou encore de VolSurf [24, 23].

Une fois les descripteurs sélectionnés et tous les composés de la base de données décrits, il est possible de lancer des requêtes sous la forme d'une ou plusieurs molécules.

3.2 Calcul de la similarité

3.2.1 Similarité basée sur les descripteurs

Quelle que soit la représentation moléculaire utilisée, il est important de noter que les relations de voisinage entre les molécules sont liées au choix de l'espace chimique de référence. En effet, la similarité entre molécules ne peut être évaluée que par rapport à un type de représentation moléculaire donné.

Comme indiqué ci-dessus, la similarité ou dissimilarité moléculaire est mesurée comme étant la distance inter-moléculaire dans l'espace de référence choisi.

Les mesures de distances conventionnelles telles que la distance Euclidienne ou la distance de Hamming mesurent la distance entre deux molécules dans l'espace chimique, alors que les coefficients de similarité (comme Tanimoto, Dice ou Cosine) évaluent directement la similarité intermoléculaire [96]. La plupart des coefficients de similarité produisent des valeurs comprises entre 0 (représentant une dissimilarité maximale) et 1 (représentant une similarité maximale), ou peuvent être normalisés. On les appelle des coefficients d'association.

Lorsque des empreintes binaires sont utilisées, le recouvrement (*overlap*) des chaines de bits sert de mesure de la similarité moléculaire.

Le coefficient d'association le plus utilisé dans les applications chimiques est le coefficient de Tanimoto. (Tc), qui compte le nombre de bits en commun entre 2 empreintes binaires par rapport au nombre total de bits dans chaque empreinte. Le Tc pour 2 représentations d'empreintes binaires A et B est calculé comme suit :

$$Tc(A,B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (1)$$

Où N_{AB} est le nombre de bits fixés dans chaque empreinte et N_A et N_B le nombre de bits fixés dans A et B respectivement.

La similarité basée sur les descripteurs telle que nous l'avons présentée et telle que nous l'utiliserons comme base au développement d'un nouveau score de similarité dans la suite de ce manuscrit n'est qu'un volet des différents types de mesures de similarité possibles entre deux composés. Il existe d'autres approches indépendantes de l'utilisation des descripteurs telles que la similarité basée sur la superposition, la similarité basée sur les représentations moléculaires en graphes [35], les comparaisons d'histogrammes [80, 6], ou encore le traitement Brownien des molécules [37].

4 Les Relations Structure-Activité Quantitatives (*QSAR*)

Lors d'une étude de *QSAR*, on étudie les relations entre la structure et l'activité d'un composé, par exemple les effets d'une variation chimique locale sur une molécule à l'activité connue. En effet, certains changements chimiques sur certaines parties d'une molécule peuvent entrainer des variations de son activité biologique en agissant sur l'interaction avec la cible. [25]

Ainsi, la méthodologie *QSAR* permet de trouver un modèle mathématique qui mette en corrélation l'activité et la structure au sein d'une famille de composés. De nombreuses méthodes conceptuellement différentes peuvent être utilisées pour mettre en place les modèles mathématiques permettant de détecter des relations de type *QSAR*. Ainsi, les études *QSAR* sont basées sur des méthodes statistiques dans lesquelles on peut trouver les méthodes de régression linéaire, multilinéaire ou encore des méthodes d'apprentissage machine (*Machine Learning*), dont les réseaux de neurones sont une sous-branche.

Les grandes phases de la mise en place d'un modèle *QSAR* peuvent être décrites comme suit : Extraire les descripteurs à partir de la structure moléculaire, choisir les descripteurs adaptés à l'étude par rapport à l'activité analysée, et utiliser les descripteurs comme variables explicatives pour définir une relation qui les corrèle à l'activité en question. Chaque modèle doit être validé sur des jeux de données de test.

4.1 Préparation des données d'entrée

Une bonne préparation des données d'entrée est nécessaire pour une étude *QSAR* efficace (ou pour n'importe quelle étude chémoinformatique). Cette préparation implique plusieurs étapes de travail sur les composés :

- S'assurer que les conditions expérimentales dans lesquelles ont été obtenues les mesures d'activité des molécules sont similaires,
- Éliminer les doublons,
- Appliquer les mêmes règles de standardisation pour les structures des composés,
- Éliminer les mélanges (sauf si, bien sûr, il faut prédire les propriétés de mélanges – un cas particulier qui ne sera pas approfondi ici).

4.2 Génération des descripteurs moléculaires à partir des structures.

Une fois les données d'entrée préparées, la première étape de l'étude consiste à obtenir un ensemble de descripteurs pour chacune des molécules.

En effet, les composés, encodés comme des ensembles de liaisons covalentes et d'atomes, ne peuvent être utilisés directement *in silico*. Les structures chimiques ne contiennent pas habituellement d'information explicite les reliant à l'activité. Cette information doit être extraite grâce au descripteurs variés qu'il est possible de mettre en place. De cette manière, des propriétés implicites contenues dans la structure de la molécule vont être mises en avant, sachant que seules certaines d'entre elles peuvent éventuellement corréler avec l'activité. Comme indiqué précédemment (chapitre I.3.), ces descripteurs sont basés non seulement sur la structure des composés mais aussi sur un ensemble de propriétés physico-chimiques.

Pour des raisons techniques, les descripteurs sont représentés sous la forme de vecteurs de nombres pour chaque molécule. Ces vecteurs doivent tous être de la même longueur. En effet, la plupart des méthodes utilisées pour la prédiction requièrent comme données d'entrée des objets comparables de taille constante, ce qui n'est pas naturellement le cas des structures des composés, qui sont diverses en taille et en nature.

4.3 Sélection des descripteurs les plus adaptés

La plupart des applications sont capable de générer un grand nombre (des centaines voire des milliers) de descripteurs moléculaires différents. Il est important de pouvoir choisir parmi la grande quantité de descripteurs qui peuvent être générés ceux qui sont effectivement corrélés avec l'activité. De plus, de nombreux descripteurs sont inter-corrélés, ce qui peut avoir des effets négatifs sur l'analyse *QSAR*. De manière générale, un trop grand nombre de descripteurs fait baisser la robustesse de l'étude. De plus, certaines méthodes statistiques nécessitent d'avoir un nombre de descripteurs significativement plus restreint que le nombre de composés donnés en entrée.

Pour cette raison, des méthodes spécifiques doivent être utilisées afin de réduire le nombre de descripteurs aux descripteurs les plus informatifs. Le set de descripteurs doit donc être le plus petit possible mais le plus riche en informations possible.

Les méthodes utilisées pour sélectionner les meilleurs descripteurs peuvent être regroupées en deux catégories [39]. Les méthodes dites "d'emballage" (*wrapper*) construisent et évaluent une série de modèles *QSAR* afin d'évaluer la qualité des sous-ensembles de descripteurs. Dans les méthodes de filtrage, aucun modèle n'est construit, mais les descripteurs sont évalués via d'autres critères (Critères statistiques tels que le ratio de Fisher [60] ou encore critère de corrélation de Pearson [65, 38]).

4.4 Relation des descripteurs à l'activité.

Une fois les descripteurs moléculaires utiles calculés et sélectionnés, la dernière étape est de créer une fonction reliant leurs valeurs à l'activité analysée. La valeur qui quantifie l'activité sera donc exprimée comme étant une fonction des valeurs des descripteurs.

Les meilleures fonctions sont en général mises en place en se basant sur l'information contenue dans l'ensemble d'entrainement (les composés pour lesquels l'activité est connue).

Il existe une gamme de familles de fonctions très vaste, incluant des fonctions linéaires, qui ont été utilisées depuis le début du *QSAR*. Ces fonctions prédisent l'activité comme étant une fonction linéaire des descripteurs moléculaires. En général, ces fonctions linéaires sont facilement interprétables et suffisamment précises pour de petits ensembles de composés similaires, spécialement lorsque les descripteurs sont sélectionnés avec soin pour une activité donné. Ces méthodes sont par exemple la régression linéaire multiple (*MLR – Multiple Linear Regression*) [79, 91, 49], La méthode des moindres carrés (*PLS – Partial Least Squares*) [98, 97] ou encore l'analyse discriminante linéaire (*LDA – Linear Discriminant Analysis*) [31].

D'autres méthodes, non-linéaires, étendent l'approche à des relations plus complexes. Ces modèles se révèlent être plus précis, spécialement pour des ensembles de données plus larges et plus divers. Cependant, ces modèles nous heurtent parfois à des difficultés de compréhension et sont parfois en proie à l'*overfitting* (ils se borneront dans ce cas à décrire du bruit au lieu de la relation sous-jacente entre descripteurs et activité). On utilisera comme méthodes non-linéaires la classification de Bayes [89], la méthode des k plus proches voisins (*k-NN*) [20], des réseaux de neurones [50, 75], mais aussi des arbres de décision [73, 34], ou des méthodes des machines à vecteurs de support (*SVM – Support Vector Machines*) [19].

5 Les Cartes auto-organisatrices

Une carte auto-organisatrice (*SOM - Self-Organizing Map*) est une classe de réseau de neurones artificiel basée sur des méthodes d'apprentissage non supervisé. L'algorithme de classification à la base des SOMs a été développé par Teuvo Kohonen en 1982. [7]

Ce type de carte est beaucoup utilisé en Chémoinformatique, principalement pour décrire l'espace chimique mais aussi pour détecter des relations entre composés en terme de similarité (La mesure de similarité obtenue entre deux objets projetés sur une SOM sera de type distance Euclidienne entre 2 vecteurs à n coordonnées, n représentant le nombre de descripteurs calculés pour chaque composé.)

Les caractéristiques des SOMs en font des outils faciles à utiliser. En effet, il est possible de construire des cartes à partir de données d'entrée sous forme de vecteurs de descripteurs à n coordonnées. L'application de l'algorithme de Kohonen sur ces données permet de réduire leur dimensionalité (en général à une dimension de 2) tout en préservant sa topologie. Ainsi, une SOM sera visuellement représentée sous la forme d'une grille à 2 dimensions de neurones (ou nœuds), chacun étant représenté par un vecteur de poids (ou vecteur code) contenant n coordonnées.

Cette représentation permet une visualisation rapide des relations entre composés. La préservation de la topologie de l'espace d'entrée entraine une représentation sous forme de *clusters* : si deux objets sont assignés à un même neurone, alors on pourra considérer qu'il existe une relation de similarité entre eux. De plus, lorsqu'un objet nouveau est projeté sur un neurone, il sera considéré comme ayant des caractéristiques communes avec les autres membres de ce neurone.

Nous verrons dans la suite de ce manuscrit que notre utilisation des cartes de Kohonen, construites avec nos descripteurs, se portera sur leur utilité pour accélérer les recherches par similarité dans de larges bases de données.

Maintenant que les principales méthodes utilisées dans ce manuscrit ont été présentées, nous allons nous intéresser plus en détails au sujet principal du développement de nos travaux : les descripteurs pharmacophoriques topologiques.

Deuxième partie

Les descripteurs topologiques pharmacophoriques

6 Introduction

Les pharmacophores sont définis comme étant "l'ensemble des caractéristiques stériques et électroniques nécessaires pour assurer des interactions supramoléculaires optimales avec la structure d'une cible biologique spécifique et pour déclencher (ou bloquer) sa réponse biologique". Ce concept provient de la volonté des chimistes de tenter de rationaliser les relations entre structure et propriétés.

Grâce à la compréhension de la nature tri-dimensionnelle des molécules et des contraintes stériques qui déterminent leurs conformations privilégiées, la liaison entre un ligand et une macromolécule a pu être expliquée par le paradigme très simplifié de la serrure et de la clé : la liaison se fait par complémentarité de forme. La nature des forces de liaison non covalentes (électrostatique, par liaison hydrogène et différentes contributions dispersives comme la solvatation ou les effets hydrophobes) est cependant extrêmement complexe.

Les prédictions d'affinité basées sur des études en profondeur des interactions physico-chimiques entre le ligand, sa cible et le solvant (par exemple des études de *docking* flexible ou des simulations de perturbation de l'énergie libre) sont habituellement trop consommatrices de temps pour pouvoir être utilisées à large échelle, bien qu'elles soient déjà des approximations de la réalité physique de par l'utilisation de calculs d'énergie basés sur des champs de force empiriques. A la place de ces calculs complexes, le principe de complémentarité des groupes fonctionnels (les cations interagissent favorablement avec les anions, les donneurs de liaisons hydrogènes avec les accepteurs et les hydrophobes entre eux) a été adopté pour devenir le second pilier du concept de pharmacophore aux cotés du concept de complémentarité de formes. L'essentiel de l'approche pharmacophorique est donc le marquage des atomes non plus selon leur élément chimique, mais par "type pharmacophorique" définissant, en grandes lignes, leur comportement physico-chimique. Les types pharmacophoriqes classiquement pris en compte sont : hydrophobe, aromatique (parfois considérés simplement comme hydrophobes, et non comme un type indépendant), accepteur de liaisons d'hydrogène, anions, donneurs de liaison H, cation.

En chimie médicinale, le pharmacophore est souvent vu comme complémentaire au châssis moléculaire (*scaffold*) de la molécule, c'est à dire à sa topologie. Le *scaffold hopping* est devenu un paradigme central dans la recherche de nouveaux médicaments. Cette technique recherche des structures bioisostériques, topologiquement différentes, qui orientent néanmoins dans l'espace leurs groupes d'interaction de la même manière que le composé de départ, leur permettant d'avoir des interactions similaires à celle du composé de départ avec la cible biologique. L'importance que revêt cette technique provient de sa capacité à découvrir de nouveaux chemins de synthèse une fois que tous les analogues autour d'un *scaffold* ont été explorés. Elle permet aussi de découvrir des molécules ayant des capacités pharmacocinétiques différentes tout en ayant une affinité de liaison à la cible similaire à celle des composés déjà connus. Ainsi, l'optimisation du prototype (*lead*) peut être orientée de 2 manières conceptuellement différentes : l'échantillonnage de *scaffolds* variés compatibles avec un motif pharmacophorique, ou l'échantillonnage d'un certain nombre de motifs pharmacophoriques pouvant être portés par un *scaffold* donné.

7 Les pharmacophores topologiques

Le pharmacophore, basé sur le concept de la clé et de la serrure ainsi que sur l'idée de *scaffold*, est intrinsèquement tridimensionnel. De même, la découverte d'un pharmacophore implique un minimum de connaissances – idéalement, ça sera la structure du complexe ligand-cible, ou l'on "voit" les points d'ancrage définissant le pharmacophore.

On peut tenter d'extraire une liste de points conservés dans des actifs et absents dans des inactifs (validés expérimentalement) à partir de leurs motifs pharmacophoriques (c'est-à-dire, la nature des groupements représentatifs des types pharmacophoriques et leur disposition relative, sans toutefois savoir lesquels de ces groupement interagiront de manière effective avec une cible) afin de constituer une hypothèse de pharmacophore. Il est cependant nécessaire, pour que l'hypothèse soit valide, d'avoir à disposition un ensemble de composés suffisamment divers.

Il est cependant possible de décrire les motifs pharmacophoriques à partir des structures bidimensionnelles et donnant néanmoins de bons résultats dans les études de similarité. Nous allons tout d'abord détailler plus avant le développement de modèles et de pharmacophores 3D avant d'entrer plus en détail dans la description des différents types de descripteurs pharmacophoriques topologiques bidimensionnels. Cette section n'a pas vocation à être une revue complète des différents types de pharmacophores existants, mais plutôt une introduction au concept de motif pharmacophorique topologique, catégorie à laquelle les triplets pharmacophoriques flous que nous avons développés appartiennent.

8 Les modèles et pharmacophores 3D

Grâce à l'avancée des techniques d'aide à la création de nouveaux médicaments par ordinateur, le concept intuitif de pharmacophore a rapidement été adopté par les chémoinformaticiens. Les outils modernes de recherche par sous-structure peuvent être facilement adaptés pour reconnaître des groupes fonctionnels spécifiques et les classer en caractéristiques pharmacophoriques : Les donneurs de liaisons hydrogènes (HD) interagissent avec les accepteurs (HD), les cations (PC – *positive charge*) forment des ponts salins avec les anions (NC), alors que les hydrophobes (Hp) sont complémentaires entre eux. Le plus souvent, les aromatiques (Ar) sont considérés comme une catégorie à part, complémentaires à la fois entre eux mais aussi avec les hydrophobes.

Définissons NT comme étant le nombre de types considérés. ($NT = 6$ dans l'énumération précédente). Formellement, l'information du pharmacophore peut être représentée sous la forme d'une matrice de drapeaux pharmacophoriques binaire $F(a, T)$, avec $F(a, T) = 1$ si un atome a est du type T et $F(a, T) = 0$ dans le cas contraire. Un pharmacophore de liaison présumé peut alors être découvert en générant l'alignement 3D des actifs et en sélectionnant les régions de l'espace dans lesquelles toutes les molécules de l'ensemble d'entrainement placent des groupes pharmacophoriquement équivalents. Les hypothèses pharmacophoriques [63, 51] sont alors une liste de caractéristiques conservées et qui se recouvrent, trouvées dans le modèle d'alignement. N'importe quel candidat possédant ce type de groupements, placés aux mêmes endroits que les caractéristiques conservées de l'ensemble d'entrainement (le recouvrement de ces caractéristiques pouvant être délimité par une zone) peut alors être considéré comme correspondant au "pharmacophore de liaison", et par là même, considéré comme étant actif (voir figure 1).

Une grande variété d'outils utilisant la même idée générale a été développée, prouvant ainsi la grande popularité du paradigme du pharmacophore. Les sphères pharmacophoriques peuvent être remplacées par des "champs pharmacophoriques" flous ($ComPharm$) [43] dont on vérifie la correspondance avec les molécules candidates alignées. Il est même possible de complètement laisser de coté le marquage pharmacophorique en faveur d'un contrôle des champs stériques et électrostatiques ($CoMFA$) [22]. Le principe reste le même : étant donné un alignement commun, les zones possédant des valeurs de champs ou d'occupation qui corrèlent avec les activités expérimentales trouvées sur l'ensemble des exemples du set d'entrainement entrent dans le "pharmacophore". [5, 32, 61] Au stade suivant, les molécules de test et les candidats externes au criblage virtuel doivent prouver leur capacité à s'aligner sur le modèle ainsi qu'à occuper correctement (ou générer les intensités de champ appropriées dans) ces zones importantes.

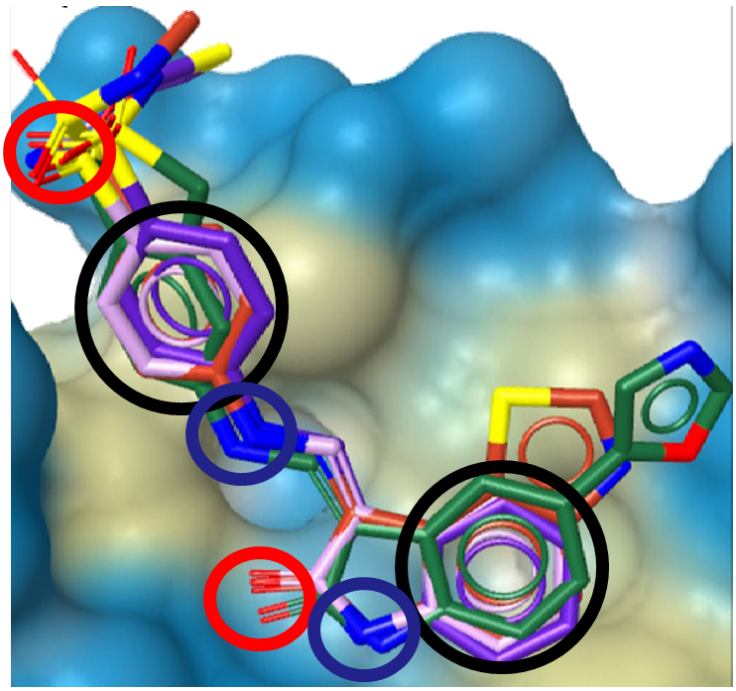

FIGURE 1 – Modèle pharmacophorique issu de la superposition de plusieurs ligands. La superposition est censée représenter la position relative dans le site actif (à l'arrière plan). Les sphères dénotent le positionnement des groupements consensuels de ces actifs.

Il n'y a qu'un pas (conceptuel) à franchir pour prétendre que les zones importantes correspondent réellement à des régions de l'espace dans lesquelles les interactions entre le ligand et le site se réalisent vraiment. Cependant, ce pas doit être pris avec une extrême précaution, car les relations observées entre champ et activités ne sont pas des preuves d'une relation de cause à effet.

Alternativement à ces méthodes, des pharmacophores de liaison peuvent être extraits comme des "images en négatif" du site de liaison d'une protéine [9], par des programmes recherchant les localisations les plus appropriées de sondes hydrophobes, polaires et chargées dans le site, puis en les combinant en une ou plusieurs hypothèses de pharmacophores.

Les modèles par superposition sont cependant très consommateurs de temps et limités à des ensembles de données qui partagent une sous-structure significative commune, en l'absence de laquelle aucun alignement ayant du sens ne peut être fait. Pour contourner ces inconvénients, des empreintes pharmacophoriques non dépendantes de l'alignement, représentant le motif pharmacophorique de la molécule, ont été introduits. (Comme indiqué précédemment, le motif pharmacophorique peut être défini comme l'arrangement spatial relatif de toutes les caractéristiques pharmacophoriques, qu'elles soient impliquées dans les interactions site-ligand ou non).

Une manière simple de caractériser le motif pharmacophorique est de générer les histogrammes de distribution de densité des paires d'atomes correspondant à chaque combinaisons de caractéristiques pharmacophoriques, par rapport à la distance qui les sépare [29, 30]. Le modèle pharmacophorique peut être caractérisé par un vecteur dans lequel chaque élément $T_i - T_j\Delta_k$ représente le nombre de paires d'atomes dans lesquelles le premier est du type pharmacophorique T_i, le second représente la caractéristique T_j et la distance qui les sépare est comprise dans la gamme de distances Δ_k.

Des triplets [70] ou des quadruplets [64] pharmacophoriques peuvent aussi être recherchés à la place des paires. Dans les empreintes pharmacophoriques binaires à trois points, les triangles de base i sont spécifiés entièrement par une liste de 3 types pharmacophoriques $T_j(i)$ - chaque type de Tj étant associé à un coin $j = 1, 2, 3$ du triangle - ainsi que par un ensemble de gammes de tolérance $[d_{kj}^{min}(i), d_{kj}^{max}(i)]$ qui spécifient les contraintes pour les longueurs des cotés des triangles. Les triangles de base doivent être considérés comme les mailles d'une grille sur lesquelles une molécule est projetée. Si l'on considère un triplet d'atomes $(a1, a2, a3)$ dans une molécule, ce triplet est considéré comme correspondant à un triplet de base i si :

1. Chaque atome a_j est du type pharmacophorique $T_j(i)$, c'est à dire $F[a_j, T_j(i)] > 0$ pour chaque sommet j
2. Les distances inter-atomiques calculées (géométriquement ou d'une autre manière) $dist(a_j, a_k)$ se retrouvent chacune dans leurs gammes de tolérance respectives : $d_{kj}^{min}(i) \leq dist(a_j, a_k) < d_{kj}^{max}(i)$

Si dans une molécule M un triplet d'atomes remplit les conditions mentionnées ci-dessus simultanément, alors l'empreinte de M sélectionnera le bit i correspondant au triangle de base.

Les empreintes de ce type sont des descripteurs statiques du motif pharmacophorique global dans la molécule. Elles décrivent comment les représentants existants de chaque type pharmacophorique sont orientés les uns par rapport aux autres, mais ne donnent aucune affirmation à propos du véritable ensemble de groupes fonctionnels qui participent à l'attachement du ligand (ou le bloquent). C'est pourquoi ces empreintes ont été principalement, et avec succès, utilisées dans les calculs de similarité moléculaire, en rapport avec le principe de similarité [47, 45] qui statue de manière simple que "des molécules similaires ont des propriétés similaires", ou plus précisément que "des molécules similaires ont plus de chances de partager des propriétés similaires que n'importe quelle paire de composés pris au hasard." Les empreintes sont les représentants d'une version plus solide du principe de similarité : "Les molécules présentant des modèles pharmacophoriques similaires ont des chances de partager le même type de comportement de liaison réversible non-covalente à des cibles biologiques."

Au delà des applications basées sur la similarité, les techniques d'apprentissage machine [69] devraient choisir les éléments de description spécifiques qui semblent corréler avec l'activité dans un set d'entrainement. A la différence des modèles par superposition dans lesquels il existe un lien évident entre des sphères ou des "champs" pharmacophoriques dans l'espace et leurs atomes sources, les paires (triplets, etc) d'atomes dans une molécule qui représentent les éléments important de description doivent être d'abord établies pour pouvoir se donner une idée des mécanismes de liaison.

9 Les pharmacophores topologiques

Les pharmacophores sont intrinsèquement tridimensionnels. Que signifie donc "pharmacophore topologique" ? Ce chapitre met en avant les aspects clés de ce sujet en s'appuyant sur des études publiées. Son but est de faire une introduction générale aux concepts et problèmes des pharmacophores 2D.

9.1 Les pharmacophores topologiques basés sur des alignements 2D

La géométrie moléculaire est une fonction de la connectivité. Cependant, pour construire des hypothèses de pharmacophores comme décrits dans la section précédente, des conformères stables des composés concernés doivent d'abord être générés. Cependant, les efforts sont payants lorsqu'il est question d'éviter l'étape d'échantillonnage conformationnel, qui est particulièrement consommatrice de temps et très bruitée, en particulier pour les molécules flexibles où les procédures stochastiques retournent des ensembles de conformères différents et incomplets à chaque étape. De plus, le succès des indices topologiques dans les études de $QSAR$ indiquent qu'il est possible de se passer d'une étape de modélisation 3D explicite et que la détection de caractéristiques pharmacophoriques est purement basée sur des considérations de connectivité.

Cependant, les performances relatives des descripteurs 2D doivent être comparées à celles des descripteurs 3D, opposant la simplicité et la robustesse contre le plus grand contenu informationnel. Les pharmacophores basés sur la superposition 3D sont très efficaces si les conformations bioactives des actifs dans le set d'entrainement utilisé pour la calibration du modèle sont connues. Dans le cas contraire, quelle géométrie doit être utilisée ? Les plus stables, si tant est que les structures sont suffisamment rigides pour que l'échantillonnage soit reproductible ? Ou peut être faudrait-il choisir des conformères pris au hasard parmi les plus stables retournés par une recherche stochastique ? Que se passe-t'il si la géométrie des deux analogues très proches se trouve être radicalement différente ?

Bien que la superposition de représentations 2D de molécules ne soit pas représentative d'une quelconque similarité tridimensionnelle, des méthodes visant à projeter les groupes d'une molécule sur des groupes équivalents dans une autre molécule existent et peuvent être utilisées pour mettre en évidence des motifs conservés dans les actifs. Typiquement, les composés du set d'entrainement sont fusionnés pour former une "hypermolécule", avec leurs atomes équivalents chimiquement et topologiquement fondus en sommets uniques de l'hypermolécule. De cette manière, les sommets spécifiquement peuplés dans les actifs et dans les inactifs peuvent être appris.

Les sous-ensembles qui apparaissent (presque) exclusivement au sein des actifs sont nommés "pharmacophores" et sont considérés comme étant responsables de l'activité, alors que ceux qui sont spécifiquement vus dans les inactifs sont nommés "antipharmacophores" et sont désignés comme empêchant le ligand de se lier à un site actif. Ces sous-ensembles ne représentent pas forcément des graphes contigus, c'est pourquoi ces méthodes devraient en principe être capable de faire du *scaffold hopping*.

Malheureusement, les méthodes basées sur des hypermolécules exploitent aussi souvent les informations 3D. Il n'y a apparemment pas d'étude explicite qui décrive les avantages et les inconvénients des modèles par superposition basés sur la topologie ou sur la géométrie dans les études $QSAR$ et dans les élucidations de pharmacophores.

Générer un alignement 2D, le premier pas de l'élucidation de pharmacophores 2D consiste à établir tout d'abord, pour chaque atome de la molécule alignée une liste de correspondances (atomes équivalents, au sens des automorphismes du graphe moléculaire) [59] dans le composé cible (auquel nous nous référerons par la suite par "modèle"). A part dans le cas où le composé aligné est un analogue proche du modèle, ce problème n'est pas trivial, car certains atomes peuvent ne pas avoir d'équivalents appropriés dans le composé partenaire. Les paires d'atomes que l'on suppose correspondants sont des nœuds de nature similaire, localisés dans des voisinages similaires (avec des sphères de coordination successives similaires). Cette information (atome, nature, voisinage) peut être traduite par des Indices Topologiques (TI – *Topological Indices*) atomiques spécifiques.

Les atomes correspondants les plus probables dans le modèle pour un nœud donné du graphe aligné peuvent être déterminés facilement comme étant ceux ayant les valeurs de TI les plus proches. Ensuite, une carte unique d'équivalence doit être établie, qui relie chaque atome de la molécule alignée à au moins un atome du modèle. Les alternatives possibles de projection doivent être évaluées pour permettre de chercher la plus optimale : des "bonus" sont considérés, à chaque fois que des atomes ont été projetés avec succès sur un équivalent ayant une valeur de TI très similaire, et des "pénalités d'écart" sont soustraites pour des atomes non-correspondants.

A la différence du domaine bioinformatique, où des outils d'alignement de séquence [72] ont émergé et sont utilisés comme standards dans l'industrie, il n'existe pas d'outil d'alignement 2D universellement accepté pour les composés organiques. Il n'y a tout d'abord pas de consensus concernant la "coloration" des graphes moléculaires (celle utilisée par défaut étant les symboles atomiques, bien qu'une coloration par caractéristiques pharmacophoriques pourraient être significative). De plus, la nature 2D des ligands organiques complique la construction de la carte d'équivalence, ce qui contraste avec les procédures d'alignement de séquences 1D qui sont plutôt simples. Les approches basées sur des alignements 2D ont émergé dans les années 80 [82], en introduisant le concept "d'hypermolécule", un graphe obtenu par fusion de graphes moléculaires individuels en faisant correspondre les atomes équivalents (voir ci-dessus). Bien que le critère d'alignement soit de nature topologique (*MTD – Minimal Topological Difference*, c'est à dire mettre en correspondance les nœuds équivalents topologiquement), l'approche était typiquement utilisée [57, 18] pour piloter la superposition de modèles moléculaires 3D pour construire des pharmacophores 3D ou des modèles de pseudo-récepteurs. Les nœuds des hypermolécules peuvent être classifiés en points de "cavité" (présents spécifiquement dans les composés actifs du set d'entrainement, amenant ainsi des interactions favorables site-ligand), des points de "murs de site" (présents dans les inactifs, et représentant ainsi des points de l'espace qui ne doivent pas être occupés par le ligand), et des points "indifférents".

Une version purement topologique des *MTDs* [66] se base sur l'alignement de chaque molécule candidate sur l'hypermolécule, pour calculer des indices topologiques spécifiques par rapport aux sous-ensembles d'atomes qui sont projetés sur les points de "cavités" et de "murs" respectivement.

Une stratégie d'alignement purement topologique basée sur la correspondance de sous-arbres, l'approche *MTree* [41] a permis de mettre en évidence des pharmacophores topologiques significatifs. Les molécules sont d'abord réduites à des "arbres de caractéristiques" dans lesquels chaque nœud représente des groupes fonctionnels inter-connectés et est coloré en fonction du type pharmacophorique du groupe représentatif. Un nouvel algorithme d'alignement par paires (la recherche dynamique de correspondances [101]) permet d'arriver à un alignement moléculaire topologique efficace basé sur la correspondance (chimiquement raisonnable) de groupes fonctionnels. En se basant sur cet alignement, un nouvel arbre (le modèle *MTree*) qui combine l'information contenue dans plusieurs arbres de caractéristiques peut être créé.

Les nœuds représentent les correspondances entre les caractéristiques des sous-arbres projetés, et les arcs sont formés en suivant les topologies des arbres de caractéristiques d'entrée. Chaque nœud du *MTree* peut être coloré par son degré de conservation dans les molécules actives. Les modèles *MTree* ne sont pas conceptuellement équivalents aux pharmacophores topologiques mais plutôt aux hypermolécules sus-mentionnées. Ils représentent une fusion des nœuds rencontrés dans les composés du set d'entrainement, certains étant conservés dans de nombreuses molécules, d'autres se rencontrant moins souvent. Pour générer un pharmacophore topologique à partir d'un modèle *MTree*, des poids devraient être assignés aux nœuds pour représenter les occurrences relatives de chaque nœud au sein des actifs, par contraste par rapport aux inactives.

9.2 Les pharmacophores topologiques basés sur des empreintes

L'analogie entre les motifs pharmacophoriques 3D et les descripteurs pharmacophoriques topologiques [14] est assez évidente : il suffit de remplacer les distances euclidiennes dans les empreintes 3D par des distances inter-atomiques topologiques du plus court chemin (*shortest-path*). Ces distances sont des entiers et simplifient le schéma de regroupement dans des cases basé sur la distance (les paires à la même distance topologique entrent dans la même case). Les empreintes pharmacophoriques 2D peuvent être utilisées exactement comme leurs homologues 3D, que ce soit pour du criblage basé sur la similarité ou encore pour de l'apprentissage machine de modèles pharmacophoriques basés sur les descripteurs sélectionnés. Si les distances inter-atomiques 3D réelles corrèlent bien avec leurs séparations topologiques, alors les descripteurs pharmacophoriques 2D et 3D devraient se comporter de la même façon.

Bien entendu, il arrive que certaines molécules se replient de manière à rendre proches des atomes séparés par de nombreuses liaisons, une situation dont les descripteurs 2D ne parviennent pas à rendre compte. Néanmoins, en pratique, le conformère bioactif peut ne pas être connu, c'est pourquoi les empreintes 3D se basent souvent sur un ensemble de conformères plutôt que sur une seule géométrie.

Il a été montré que les descripteurs 3D basés sur une seule géométrie se comportent de manière erratique, car les programmes de construction de géométries peuvent retourner des structures différentes pour des composés topologiquement similaires (et des programmes différents peuvent retourner différents conformères pour le même composé). Des analogues proches peuvent alors être représentés par des empreintes très différentes. Alternativement, les moyennes des distances inter-atomiques calculées sur un ensemble de géométries sont corrélées bien plus fortement avec les distances topologiques, si la diversité conformationnelle est suffisante (c'est à dire, si l'ensemble énumère à la fois les structures repliées et non repliées). En ce sens, les empreintes 3D moyennes par rapport à un ensemble de conformères ne sont pas rigoureusement tridimensionnelles [71, 44] : ces descripteurs devraient plutôt être nommés empreintes "2.5D". Les empreintes pharmacophoriques 2D devraient présenter de meilleures performances à la fois dans les calculs par similarité mais aussi en apprentissage machine. Même lorsque leurs performances sont moins bonnes que celles de leurs homologues 3D, elles devraient tout de même être utilisées eu regard à leur cout bien moins élevé en terme de puissance de calcul.

De nombreuses empreintes pharmacophoriques topologiques, indicateurs binaires de présence ou comptes flous de niveaux de population de paires [30, 47, 14, 77, 16] ou de triplets [71, 1] d'atomes pharmacophoriquement marqués ont été développés.

Typiquement, la mise en place de pharmacophores topologiques consiste en plusieurs étapes : importation de la molécule, standardisation (délétion des contre ions, réparation des ordres de liaisons ambigus, ajout/délétion des atomes d'hydrogène, etc...) et analyse topologique (calcul de la matrice de distances des plus courts chemins topologiques). Après cela, le marquage pharmacophorique des atomes / groupes fonctionnels est fait et les paires ou triplets de caractéristiques sont détectés et classifiés en fonction des séparations topologiques et des types pharmacophoriques.

Bien que les empreintes pharmacophoriques topologiques soient considérées comme conceptuellement différentes de l'énumération de fragments, la limite entre ces 2 catégories n'est pas si nette. En fait, des empreintes pharmacophoriques 2D peuvent être obtenues [68] via une procédure de recherche par sous-structure générique guidée par les *SMARTS* [90] des composés, et les motifs pharmacophoriques comptés pour chaque élément de l'empreinte pharmacophoriques sont des fragments génériques correspondant aux caractères de remplacement ("*wildcard matching*"). Par exemple, l'expression Hp * * * HA, dans laquelle *Hp* et *HA* sont les définitions génériques dans *SMARTS* pour "Hydrophobes" et "Accepteurs" respectivement, correspond à n'importe quelle paire d'hydrophobe-accepteur séparée par 4 liaisons – un terme typique de *CATS* [77]. Techniquement, le choix d'une détection de pharmacophores guidée par *SMARTS* est assez puissant, car il peut implicitement autoriser n'importe quel degré arbitraire d'affinement des définitions des types pharmacophoriques (par exemple, les catégories des donneurs et accepteurs pourraient être subdivisées en sous types) sans avoir à modifier le logiciel.

9.2.1 Les empreintes basées sur les paires pharmacophoriques topologiques

Un des descripteurs par paires le plus utilisé, *CATS* [77] (*Chemically Advanced Template Search*) représente le compte de 150 types différents de paires d'atomes définies comme la combinaison de 5 caractéristiques pharmacophoriques (*HA*, *HD*, *PC*, *NC* et lipophiles = *Hp* + *Ar*), multipliées par 10 distances de plus courts chemins (1 à 10 liaisons). Le même principe est utilisé pour toutes les empreintes pharmacophoriques basées sur des paires. Les différences viennent des types pharmacophoriques explicitement considérés (une distinction entre *Hp* et *Ar* peut être mise en place [44]), des règles pharmacophoriques appliquées et en termes de distances considérées.

L'énumération floue des motifs pharmacophoriques a été initialement introduite comme un moyen de réduire le bruit du aux distances 3D dépendantes de l'échantillonnage dans les empreintes 3D [87]. Il a néanmoins été montré qu'il est bénéfique de flouter les frontières entre les catégories strictes de distances entières topologiques. Ceci pourrait expliquer la tolérance implicite des récepteurs, qui tolèrent une insertion / délétion d'un groupe -CH2- dans les lieurs (*linkers*) sans que ceci implique de changements spectaculaires de l'affinité. L'utilisation de graphes réduits [88] représente un pas de plus vers le flou, car les détails structuraux sont fusionnés en groupes fonctionnels génériques.

9.2.2 Les triplets pharmacophoriques

Afin d'énumérer et compter les triplets pharmacophoriques présents dans une molécule, un ensemble de triplets pharmacophoriques de référence est d'abord choisi – limitant ainsi la recherche des triplets possibles à cet sous-ensemble bien précis. Cet ensemble énumère toutes les combinaisons à prendre en compte, sur toutes les combinaisons possibles de caractéristiques pharmacophoriques pour chaque coin, multipliées par toutes les longueurs de cotés obéissant à l'inégalité triangulaire, dans une gamme finie $[E_{min}, E_{max}]$ (E = longueur des côtés). Les *TGT* (*Typed Graph Triangles*) [1] sont un exemple simple de des empreintes binaires, qui suivent la présence / absence de chacun des triangles pharmacophoriques considérés. Un certain nombre d'améliorations [28] ont été apportées par la suite, comme par exemple autoriser les recouvrements entre les types pharmacophoriques (c'est à dire, autoriser les groupes fonctionnels à représenter plusieurs types, les carboxylates comptant par exemple à la fois comme accepteurs et anions), ou encore suivre les comptes de triplets au lieu de leur présence / absence. Ceci a conduit à des améliorations des performances en criblage virtuel.

Les triplets pharmacophoriques flous (2D-FPT) ont été conçus pour améliorer des aspects spécifiques détaillés plus loin. Le reste de ce manuscrit présentera nos travaux de développement des Triplets pharmacophoriques Flous 2D, ainsi que leur utilisation dans des études de *QSAR*. Nous les utiliserons aussi comme descripteurs de base dans le développement de cartes auto-organisatrices, afin d'améliorer les performances des recherches par similarité dans notre base de données, en terme de vitesse.

Troisième partie

Les empreintes pharmacophoriques tricentriques floues – 1ere partie : Les triplets pharmacophoriques flous et les fonctions de calcul de similarité adaptées à ces nouveaux descripteurs

Dans ce chapitre, nous allons nous intéresser en détail aux triplets pharmacophoriques topologiques flous bidimensionnels (2D-FPTs) introduits brièvement plus haut.

Ces descripteurs représentent deux types particuliers d'informations contenues dans les molécules de manière simple :

- La distance topologique entre les atomes (nombre de liaisons interposées entre deux atomes),
- Le type pharmacophorique de chaque atome (6 types pharmacophoriques différents sont pris en compte : cations, anions, hydrophobes, aromatiques, donneurs et accepteurs de liaisons hydrogène).

Ces deux types d'informations sont réunis par 3 (dans un triplet pharmacophorique de type $T_1D_{2,3} - T_2.D_{1,3} - T_3.D_{1,2}$ où T_n correspond au type pharmacophorique de l'atome n et $D_(n, m)$ correspond à la distance topologique entre les atomes n et m) et sont énumérés dans une molécule.

De plus, les 2D-FPTs apportent 3 améliorations par rapport aux autres types d'empreintes pharmacophoriques :

Le comptage des triplets L'empreinte d'une molécule est représentée par le niveau de population d'un jeu de triplets de référence. Ce jeu de référence doit être choisi afin de couvrir les triplets susceptibles d'être responsables d'une activité biologique, donc toutes les combinaisons envisageables décrivant des triangles porteurs des 6 propriétés pharmacophoriques et des dimensions comprises dans un intervalle donné $[E_{min}, E_{max}]$ (E = longueur des côtés). Avec 6 propriétés pharmacophoriques et des distances topologiques comprises entre 2 et 12, il existe 31846 triplets possibles. Un des avantages de la logique floue utilisée ici est la possibilité de réduire le jeu de triplets de référence a un petit sous-ensemble du nombre total des triplets possibles car le traitement flou permet de projeter des triplets moléculaires non explicitement présents dans le jeu de référence sur des triplets de référence similaires. Une telle réduction du jeu de référence n'est pas possible avec un schéma de comptage binaire dans lequel la présence ou absence de chaque triplet possible doit être surveillée de manière explicite. La projection d'un triplet moléculaire sur les triplets de base avoisinants se fait proportionnellement au score de similitude des triplets. Ces scores sont calculés selon la méthodologie ComPharm en considérant les triangles comme pseudo-molécules à superposer. Au delà de l'intérêt d'un jeu de référence réduit, la logique floue sert à rendre compte de la tolérance naturelle des récepteurs à accommoder des ligands de tailles différentes. Le niveau de flou, qui se contrôle via les paramètres de superposition ComPharm, a un impact sur la capacité de la méthode à reconnaitre la similarité d'un même motif pharmacophorique porté par des châssis moléculaires différents (*Scaffold Hopping*).

L'utilisation de l'équilibre protéolytique La seconde amélioration majeure apportée par les 2D-FPTs réside dans la mise en place de descripteurs dépendants du pK_a. Cette dépendance consiste en l'énumération de tous les états de protonation significativement contributeurs, à un pH donné, selon les valeurs prédites du pK_a des groupes ionisables [15]. Le 2D-FPT moléculaire renvoyé est donc une moyenne des empreintes des états de protonation considérés, pondérée par la population. Le marquage basé sur des règles considère les états de protonation des groupes fonctionnels en dehors de leur contexte moléculaire (par exemple, si l'on applique mécaniquement la règle "les amines secondaires sont protonées", un fragment ethylenediamino $R1-NH-CH2-CH2-NH-R2$ sera marqué comme doublement protoné $R1-NH2^{+}-CH2-CH2-NH2^{+}-R2$). Par contre, le marquage dépendant du pK_a retournera l'empreinte moyenne de deux espèces dominantes, $R1-NH2^{+}-CH2-CH2-NH-R2$ et $R1-NH-CH2-CH2-NH2^{+}-R2$, avec deux différences notables par rapport au marquage basé sur des règles.

Tout d'abord, le triplet sensible au pK_a ne renverra aucun triplet peuplé qui présente un coté cation-cation, car les deux cations putatifs ne sont jamais présents simultanément dans la même espèce. Il sera alors assez similaire au triplet présent dans $R1 - NH - CH2 - CH2 - O - R2$, ce qui est plutôt proche de la réalité. En effet, une piperazine se comporte de manière assez similaire à la morpholine ou même à la cyclohexylamine, alors que le marquage basé sur des règles aurait suggéré des différences significatives à cause de la charge additionnelle.

De plus, des changements affectant $R1$ et $R2$, incluant des substitutions qui n'entraineraient aucune différence dans un marquage basé sur des règles (par exemple, le remplacement d'un groupe $-CH3$ par un $-Cl$, les deux étant des hydrophobes) affecteraient par contre significativement l'empreinte obtenue si ils affectent le pK_a des groupes amino. Cela entrainerait un changement dans les niveaux relatifs de population des deux états principaux de protonation et par là même altérerait les participations des empreintes respectives lors du calcul de la moyenne moléculaire 2D-FPT.

La mise en place d'une nouvelle formule de calcul de similarité Dans le calcul de similarité entre deux molécules, l'absence simultanée d'un triplet doit être prise en compte comme étant moins contraignante et probante qu'une présence simultanée. Pour rendre compte de cette observation, un nouveau score de similarité a été mis en place. L'utilisation de cette formule semblait de prime abord donner de bons résultats de comportements au voisinage (*neighborhood behaviour*), dépassant même en performance les descripteurs pharmacophoriques 2D ou 3D binaires. Cependant, une étude très récente [48] nuance ces propos. Nous reviendrons sur ce sujet en fin de manuscrit (voir Chapitre X).

Le logiciel de création et de calcul de 2D-FPTs a été développé en utilisant des outils du set de développement chémoinformatique de Chemaxon (www.chemaxon.com).

Quatrième partie

Fuzzy Tricentric pharmacophore Fingerprints. 1. Topological Fuzzy pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes

Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes

Fanny Bonachéra,[†] Benjamin Parent,[†] Frédérique Barbosa,[‡] Nicolas Froloff,[‡] and Dragos Horvath*[,†]

Unite Mixte de Recherche 8576 Centre Nationale de la Recherche Scientifique − Unité de Glycobiologie Structurale & Fonctionnelle, Université des Sciences et Technologies de Lille, Bât. C9-59655 Villeneuve d'Ascq Cedex, France, and Cerep, Department of Molecular Modeling, 19 Avenue du Québec, 91951 Courtaboeuf Cedex, France

Received June 15, 2006

This paper introduces a novel molecular description—topological (2D) fuzzy pharmacophore triplets, 2D-FPT—using the number of interposed bonds as the measure of separation between the atoms representing pharmacophore types (hydrophobic, aromatic, hydrogen-bond donor and acceptor, cation, and anion). 2D-FPT features three key improvements with respect to the state-of-the-art pharmacophore fingerprints: (1) The first key novelty is fuzzy mapping of molecular triplets onto the basis set of pharmacophore triplets: unlike in the binary scheme where an atom triplet is set to highlight the bit of a single, best-matching basis triplet, the herein-defined fuzzy approach allows for gradual mapping of each atom triplet onto several related basis triplets, thus minimizing binary classification artifacts. (2) The second innovation is proteolytic equilibrium dependence, by explicitly considering all of the conjugated acids and bases (microspecies). 2D-FPTs are concentration-weighted (as predicted at pH = 7.4) averages of microspecies fingerprints. Therefore, small structural modifications, not affecting the overall pharmacophore pattern (in the sense of classical rule-based assignment), but nevertheless triggering a pK_a shift, will have a major impact on 2D-FPT. Pairs of almost identical compounds with significantly differing activities ("activity cliffs" in classical descriptor spaces) were in many cases predictable by 2D-FPT. (3) The third innovation is a new similarity scoring formula, acknowledging that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence. It displays excellent neighborhood behavior, outperforming 2D or 3D two-point pharmacophore descriptors or chemical fingerprints. The 2D-FPT calculator was developed using the chemoinformatics toolkit of ChemAxon (www.chemaxon.com).

1. INTRODUCTION

Rational drug design[1,2] largely relies on the paradigm of site−ligand shape and functional group complementarity in order to explain the affinity of a ligand for its macromolecular receptor. While molecular modeling may offer a deeper insight into ligand recognition mechanisms—molecular dynamics simulations[3] or free energy perturbation calculations[4] might, in principle, also account for the entropic effects at binding—it did not succeed to displace the more straightforward concept of binding pharmacophores[5−7] from the minds of medicinal chemists.

The idea that ligand-site affinity can be broken down into pairwise contributions from interacting functional groups is, after all, not all that far-fetched. Ligand binding is entropically penalizing—a ligand would not restrict its freedom of translation, rotation, and conformational flexibility by binding to a receptor unless this cost is compensated by enthalpic gains. The existence of at least one ligand pose making favorable contacts with the active site is a necessary, albeit not sufficient condition—but even so, a virtual filtering procedure, discarding all molecules failing to show enough complementarity to the site, might well score significant enrichment in actives. Complementarity, in the pharmacophoric sense, must be understood as the ability to form stabilizing interactions—hydrophobic contacts, hydrogen bonds, and salt bridges—between a ligand and a site. The exact chemical nature of the interacting functional groups can be dropped in favor of their pharmacophore type[8] *T*—hydrophobic (Hp) or aromatic (Ar), hydrogen-bond acceptor (HA) or donor (HD), and positively charged (PC) or negatively charged (NC) ions. Pharmacophorically equivalent functional groups are considered replaceable, ignoring the specific ways in which their chemical environment may modulate their properties (the hydrogen-bonding strengths, for example). Formally, pharmacophore-type information can be represented under the form of a binary pharmacophore flag matrix $F(a,T)$, with $F(a,T) = 1$ if atom a is of type T and $F(a,T) = 0$ otherwise.

While the pharmacophore paradigm had been introduced as a purely qualitative framework to explain ligand affinity and specificity for a given site, it has been recently taken over and used as a fundament for various chemoinformatics approaches—empirical algorithmic approaches for rational in silico compound selection, on the basis of some numeric descriptors[9,10] of the distribution pattern of pharmacophoric groups in the molecule. This overall pattern, mathematically represented by a fingerprint (vector) in which every component refers to a specific combination of types at given separations, accounts for the nature and relative position (in terms of topology or geometry) of all of the groups that are

* Corresponding author tel.: +333-20-43-49-97; fax: +333-20-43-65-55; e-mail: dragos.horvath@univ-lille1.fr, d.horvath@wanadoo.fr.
[†] Université des Sciences et Technologies de Lille.
[‡] Cerep.

10.1021/ci6002416 CCC: $33.50
Published on Web 10/21/2006

potentially involved in site−ligand interactions (the actually involved ones are not necessarily known at this stage). Pharmacophore fingerprints may be exploited in both similarity searches[11] and predictive quantitative structure−activity relationships (QSARs).[12] Similarity searches assume that molecules described by covariant fingerprints have similar overall pharmacophore patterns and, hence, a higher chance to share a common binding pharmacophore (and to bind to a same target) than any pair of randomly chosen compounds. In QSAR, model fitting may select[13] several key fingerprint components as arguments to enter an empirical (linear on nonlinear) function estimating the expected activities.

Despite their simplicity and potential pitfalls,[14] pharmacophore-based empirical models have been shown to be successful chemoinformatics tools. A key factor to success is the proper definition of underlying pharmacophore descriptors, with a minimal loss of chemically relevant information. One widely used approach is to monitor the numbers of pharmacophore group pairs[9,15] as a function of the pharmacophore-type combination they represent and the distance separating them. Distribution density plots of such pairs with respect to geometric or topological distance have been shown to display excellent neighborhood behavior (NB),[16] in the sense of selectively attributing high pharmacophore similarity scores to compound pairs with similar experimental properties. The use of fuzzy logics[17] at the descriptor buildup and similarity scoring stages appeared to be paramount in order to smooth out conformational sampling or categorization artifacts. Higher-order descriptors[18−20] monitor the triplets or quadruplets of pharmacophore types and, therefore, furnish a much more detailed description of the overall pharmacophore pattern but become more costly to evaluate and, more important, much more prone to categorization artifacts. This is the case of the binary three-dimensional three- and four-point fingerprints, which were found to show deceivingly low NB compared to their fuzzy two-point counterparts.[16] The main reason for this is the uncertainty of the assignment of a pharmacophore-type triplet or quadruplet to one of the predefined basis triangles or tetrahedra corresponding each to one of the fingerprint elements. In the context of a binary three-point fingerprint (see Figure 1), a basis triangle i is fully specified by a list of three pharmacophore types $T_j(i)$—each type T_j being associated with a corner $j = 1-3$ of the triangle—plus a set of three tolerance ranges $[d_{kj}^{\min}(i), d_{kj}^{\max}(i)]$ specifying constraints for triangle edge lengths. Basis triangles should thus be understood as the meshes of a grid onto which a molecule is being mapped. Considering an atom triplet $\{a_1, a_2, a_3\}$ in a molecule, this triplet is said to match a basis triangle i if (1) each atom a_j is of pharmacophore type $T_j(i)$, in other terms, $F[a_j, T_j(i)] > 0$ for each corner j and (2) the calculated—geometric or other—interatomic distances $\mathrm{dist}(a_j, a_k)$ each fall within the respective tolerance ranges: $d_{kj}^{\min}(i) \leq \mathrm{dist}(a_j, a_k) < d_{kj}^{\max}(i)$.

If in a molecule M an atom triplet simultaneously fulfilling the above-mentioned conditions can be found, then the fingerprint of M will highlight the bit i corresponding to this basis triangle. The risk taken here is that in a very similar compound M'—or, if $\mathrm{dist}(a_j, a_k)$ are taken as geometric interatomic distances, in a slightly different conformation of the same molecule M—the equivalent atom triplet $\{a'_1, a'_2, a'_3\}$ may fail to match the basis triangle i. It is sufficient to have one of the three distances $\mathrm{dist}(a'_j, a'_k)$ exceeding by little one of the boundaries in order to highlight a completely different basis triangle i' in the fingerprint of M'. Basis triangles i' and i are similar, but this is ignored by a binary similarity scoring scheme failing to find either bit i or bit i' set in both compounds. In two-point descriptors, where elements standing for successive distance ranges are assigned successive indices $i' = i \pm 1$, the fingerprint scoring function could be trained to account for the covariance of neighboring bins. Such a straightforward fuzzy logics correction is no longer applicable here. There are, for example, three "successive" triangles of i {with the same $[d_{kj}^{\min}(i), d_{kj}^{\max}(i)]$ ranges for two of the edges and using the successive tolerance range for the third} but only one slot at position $i + 1$ of the fingerprint. The direct consequence is that relatively small differences in interatomic distances may trigger apparently random jumps (symbolized by the arrow of Figure 1, upper part) of the highlighted bits from one location in the fingerprint to another.

This paper shows that fuzzy tricentric pharmacophore descriptors can be successfully constructed and used. The current work reports the buildup of the topological fuzzy pharmacophore triplets (2D-FPT) using shortest-path topological distances as an indicator of pharmacophore group separation. The descriptor reports basis triangle population levels in a molecule instead of a binary presence/absence indicator. An atom triplet in the molecule will contribute to the population levels of all of the related basis triangles by an increment which is directly related to their fuzzy matching degree (Figure 1, below). In the fuzzy approach, it is sufficient to characterize basis triangles i by a set of three nominal edge lengths $d_{jk}(i)$ instead of the above-mentioned tolerance ranges. The fuzzy degree by which an atom triplet is said to match a basis triangle will be 100% if interatomic distances perfectly equal nominal edge lengths, $\mathrm{dist}(a_j, a_k) = d_{jk}(i)$, and smoothly decrease—according to a law to be detailed further on—as discrepancies between real and nominal distances become important.

While 2D-FPTs are obviously not subject to conformational sampling artifacts, fuzzy-logics-based descriptors nevertheless present essential advantages:

- Their tolerance with respect to the limited variability of topological distances between pharmacophore groups mimics the natural fuzziness of ligand recognition by active sites, which may tolerate the insertion or deletion of linker bonds in a series of analogues.
- Their size may be significantly reduced by an appropriate choice of the basis triangle set. In the fuzzy approach, it is, for example, possible to keep only basis triangles with edge sizes being multiples of 2, 3, or 4. Within the strict buildup procedure, any atom triplet featuring two atoms separated by an odd number of bonds would fail to highlight any of the basis triangles of even edge lengths—it would, in other words, slip between the meshes of the grid. A fine grid enumerating all basis triplets with all possible combinations of nominal distances must then be used—but many more of these will be required in order to cover the same global span in terms of possible distances.

A second element of originality introduced here is the pharmacophore-type assignment scheme for ionizable compounds. Classical rule-based pharmacophore typing ignores the mutual long-range influence of multiple ionizing groups,

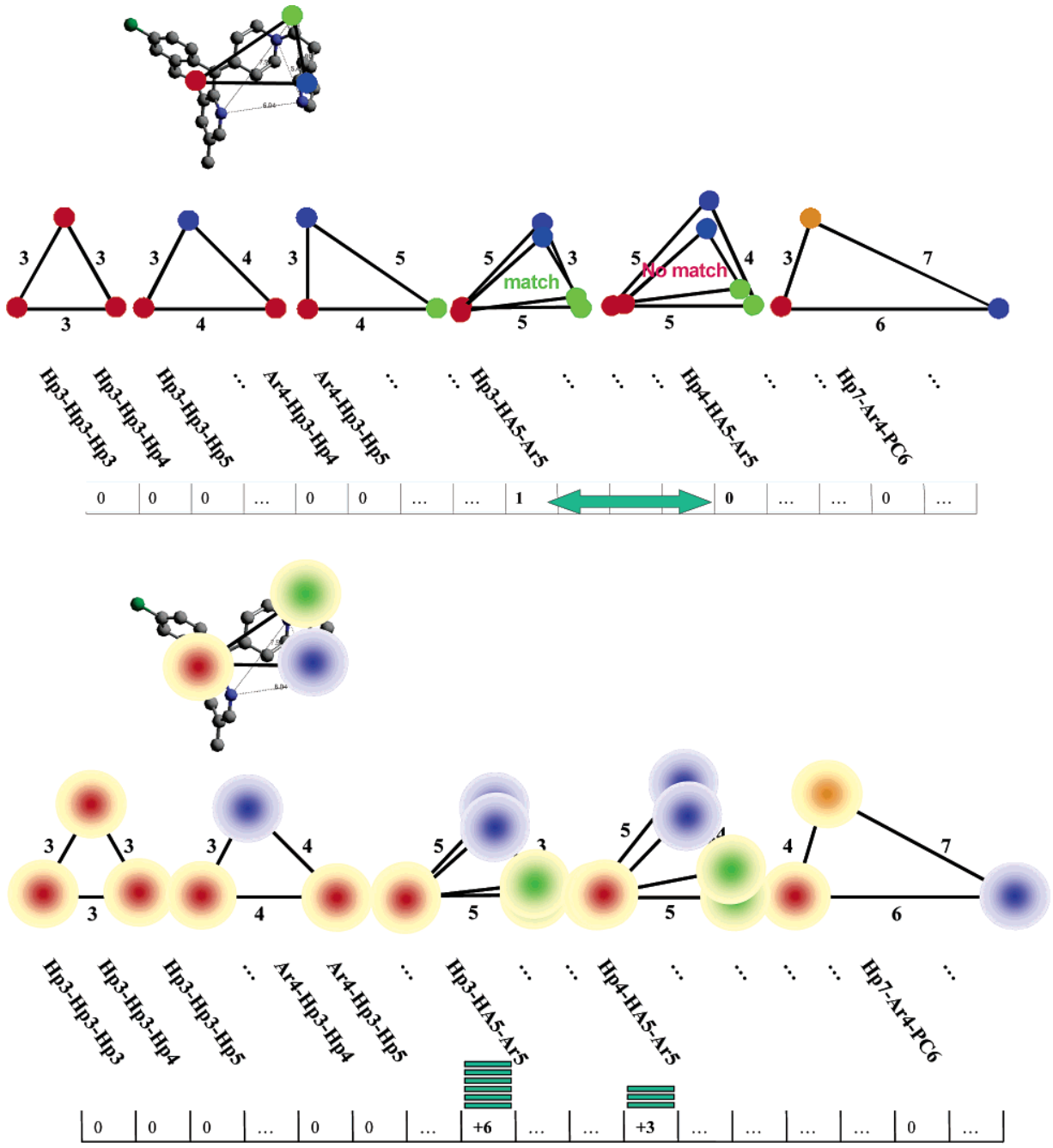

Figure 1. Buildup of a binary (above) and a fuzzy (below) pharmacophore triplet fingerprint, a vector in which every element stands for the presence (binary) or occurrence count (fuzzy) of given basis triplets. A triplet in a molecule (a) highlights a binary fingerprint component of the one best matching basis triangle or (b) increments the integer components of all of the matching basis triangles by amounts dependent on the match quality.

where each one of these is typed according to its protonation state of an isolated functional group at the considered pH. This leads to a typical overestimation of the occurrence of cation−cation or anion−anion pairs in polyamines and polyacids, respectively, and skews the molecular similarity measure upon the deletion of an ionizable group. Also, classical pharmacophore descriptors are not sensitive to electronic effects, being, for example, largely invariant upon the replacement of a methyl group (hydrophobe) by chlorine (another hydrophobe). This is acceptable unless, for example, the mentioned substitution prevents a neighboring amino group from accepting a proton in order to form a salt bridge at its binding site. To address these issues, 2D-FPT relies on the analysis of calculated[21] populations of all of the ionic or neutral forms involved in proton exchange equilibria—the "microspecies" μ, as they will be called throughout the paper—at a given pH. Each of these microspecies is mapped onto the basis triangle set, taking the actual anions and cations and donors and acceptors into account. The molecular fingerprint is rendered as the weighted average of microspecies fingerprints with respect to the predicted concentrations $c\%(\mu)$ of each microspecies μ at the considered pH of 7.4. In many cases, 2D-FPT-based analysis successfully proved that apparently near-identical compounds with puzzlingly different activities are not really as similar as they seem: the apparently minor (in the sense of classical rule-based

Table 1. Parameters Controlling 2D-FPT Buildup—Two Considered Setups

parameter	description	FPT-1	FPT-2
E_{min}	minimal edge length of basis triangles (number of bonds between two pharmacophore types)	2	4
E_{max}	maximal triangle edge length of basis triangles	12	15
E_{step}	edge length increment for enumeration of basis triangles	2	2
e	edge length excess parameter: in a molecule, triplets with edge length $> E_{max} + e$ are ignored	0	2
D	maximal edge length discrepancy tolerated when attempting to overlay a molecular triplet atop of a basis triangle	2	2
$\rho_{Hp} = \rho_{Ar}$	Gaussian fuzziness parameter for apolar (hydrophobic and aromatic) types	0.6	0.9
$\rho_{PC} = \rho_{NC}$	Gaussian fuzziness parameter for charged (positive and negative charge) types	0.6	0.8
$\rho_{HA} = \rho_{HD}$	Gaussian fuzziness parameter for polar (hydrogen bond donor and acceptor) types	0.6	0.7
l	aromatic−hydrophobic interchangeability level	0.6	0.5
	number of basis triplets at given setup	4494	7155

assignment) functional group substitutions actually had major impacts on ionization at the given pH. Many "activity cliffs" seen in classical descriptor spaces can be "leveled out" with pK_a-shift-sensitive 2D-FPT.

At last, the problem of appropriate similarity metrics to be used with 2D-FPT will be discussed, and an original scoring function, better adapted to such a high-dimensional descriptor, will be introduced. A plethora of various recipes have already been suggested[11] for comparing the descriptor sets (vectors) of two compounds m and M in order to determine a molecular dissimilarity score $\Sigma(m,M) = f[\vec{D}(M),\vec{D}(m)]$ (the distance in the structure space where each molecule is seen as a point localized by its vector of descriptors). 2D-FPT is, however, a large and potentially sparse fingerprint: out of the several thousands of basis triplets, only a few will be populated in simple molecules. Euclidean or Hamming distances may thus overemphasize the relative similarity of two simple molecules, while correlation coefficient-based metrics may be biased in favor of pairs of complex compounds. The original working hypothesis used here is to explicitly acknowledge that the simultaneous absence of a triplet in both molecules is a less-constraining indicator of similarity than its simultaneous presence, whereas its exclusive presence in only one of the compounds is a clear proof of dissimilarity. Specific partial distances are calculated with respect to the shared, exclusive, and null triplets in a fingerprint. A linear combination of these contributions leading to optimal neighborhood behavior was selected and used as the specific 2D-FPT similarity score.

For validation purposes, the NB of 2D-FPT was checked with respect to an activity profile featuring activity data (pIC_{50} values) of each molecule with respect to more that 150 targets, according to a previously outlined methodology.[22] Activity dissimilarity scores for $\sim 2.5 \times 10^6$ compound pairs were generated by Cerep, on the basis of the data in the BioPrint database[23,24] and according to a novel profile similarity scoring scheme. A second NB study has been carried out on publicly available data, by merging various QSAR data sets,[25−27] for different targets into an activity profile, assuming that each one of the molecules does not bind to any target except the one(s) for which pIC_{50} values above the micromolar threshold have been reported. Eventually, a validation study featuring virtual screening simulations will be presented. Virtual similarity screenings using 2D-FPT descriptors and metrics were performed by "seeding" a large commercially available compound collection (MayBridge) of 50 000 molecules with two sets of compounds (not used for 2D-FPT calibration) of known activities (featuring both actives and inactives) with respect to the dopamine receptor D2 and the tyrosine kinase c-Met, respectively. The ability of the 2D-FPT approach to retrieve the known actives and to avoid the selection of known inactives was benchmarked with respect to ChemAxon fuzzy pharmacophore fingerprints.[15]

2. METHODS

2.1. 2D-FPT Buildup: Fuzzy Mapping of Molecular Triplets onto Basis Triplets. Two prerequisite tasks must be completed prior to the actual construction of 2D-FPT.

Pharmacophore Flagging. This aspect will be detailed later on, because it is a central issue in ensuring the pK_a sensitivity of the fingerprints. At this time, the pharmacophore flag matrix $F_m(a,T)$, equaling 1 if atom a in the structure m is of type $T \in$ {"Hp", "Ar", "HA", "HD", "PC", "NC"} and zero otherwise, should be taken as granted. To account for the fact that aromatics and hydrophobes may, to some extent, interchangeably bind to the same binding pocket, in this work, aromatics are also flagged as lower-weight hydrophobes and vice versa. This requires the introduction of a fuzzy pharmacophore-type matrix $\Phi_m(a,T)$, identical to the binary flag matrix F for all of the polar types. For hydrophobes and aromatics, however, $\Phi_m(a,T) = \max[F_m(a,T), lF_m(a,T')]$ where T' stands for "aromatic" when T stands for "hydrophobic" and vice versa. $0 < l < 1$ is a tunable aromatic−hydrophobic compatibility parameter (Table 1). For example, an aromatic atom a has $F_m(a,\mathrm{Ar}) = \Phi_m(a,\mathrm{Ar}) = 1.0$, but $F_m(a,\mathrm{Hp}) = 0$ while $\Phi_m(a,\mathrm{Hp}) = l$.

Choice and Nonredundant Enumeration of the Basis Triplets Defining a Particular Version of 2D-FPT. The selection of a series of basis triplets to be monitored by the molecular fingerprint is essentially arbitrary and might be adapted to the specific problem for which 2D-FPTs are to be tailored. For the sake of concise specification, basis triplets are named $T_1d_{23}-T_2d_{13}-T_3d_{12}$, where T_i are the corner pharmacophore-type labels mentioned above and d_{ij} are the lengths of edges opposing each corner. For example, Ar4−Hp5−PC8 stands for a triangle in which the hydrophobe is four bonds away from the cation and eight bonds from the aromatic, while the aromatic and cation are five bonds apart. Basis triplets in this work were generated by systematic nonredundant enumeration, looping over each corner type, and respectively over each edge length from a user-defined minimal value E_{min} to a maximal E_{max}, with an integer step E_{step}. A pseudocode depiction of this procedure is given in Figure 2. Fingerprint element i hence monitors the population level of the basis triangle coded by the ith enumerated name in the list. The choice of E_{min}, E_{max}, and E_{step} (see Table 1) controls the coverage and graininess of the triplet basis set.

With these prerequisites, 2D-FPT buildup starts by the enumeration of all atom triplets $\{a_1, a_2, a_3\}$ in a molecule

```
for each T1 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {        #loop over type of corner1
   for each T2 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {     # ... corner 2
      for each T3 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {  #... and corner 3

        # Visit all the edge lengths from Emin to Emax with Estep
        for (d12=Emin,d12<=Emax,d12+=Estep) {

          #For 2nd edge, no need to loop over lengths below d12
          for (d13=d12,d13<=Emax,d13+=Estep) {

            # Only length combinations that may represent a triangle are enumerated
            # - third length may take only values verifying triangle inequalities
            dmin=max(Emin,|d12-d13|);
            dmax=min(Emax,d12+d13);
            for (d23=dmin,d23<dmax,d23+=Estep) {

              # Generate triangle corner labels Lk by concatenating types and
              # opposed edge lenght
              L1=T1d23;L2=T2d13;L3=T3d12;

              # Sort triangle corner label strings into a sorted list S.
              sort(L,S);

              # Final basis triplet name is obtained by concatenating corner labels in
              # their sorted alphabetical order
              NAME=S1'-'S2'-'S3;

              # Check whether this name had been generated previously;
              # if not add it to the list of basis triplets BLIST
              if !(BLIST.containsElement(NAME)) BLIST.add(NAME)

            } # end third edge length loop
          } # end second edge length loop
        } # end first edge length loop
      } # end third corner type loop
   } # end second corner type loop
} # end first corner type loop
```

Figure 2. Pseudocode rendering of the basis triplet enumeration procedure.

m, such that (1) the shortest topological distance between any two atoms equals or exceeds the minimal edge length $E_{\min}$ in basis triplets and (2) the longest one does not exceed the maximal edge length $E_{\max}$ by more than a tunable excess parameter e (Table 1).

To avoid confusion, in the following, the notation $t(a_k,a_j)$ to denote the (shortest-path) topological distance between two atoms will replace the generic interatomic distance dist(a_k,a_j) used in the introductory discussion on pharmacophore triplets. An atom triplet [note that the atoms of a triplet must be ordered such as to conveniently assign atoms to triangle corners; {a_1, a_2, a_3} should not be understood as a list of three atoms taken according to their sequential ordering in the structure but the permuted list with the aromatic atom in position 1 if $T_1(i)$ = Ar etc.] is said to "potentially match" a basis triplet i if (1) each atom a_j features the pharmacophore type $T_j(i)$, in other terms, $\Phi_m[a_j,T_j(i)] > 0$ for each corner j, and (2) the topological distances $t(a_j,a_k)$ are close to the corresponding nominal edge lengths $d_{kj}(i)$, in the sense that $|t(a_j,a_k) - d_{kj}(i)| \leq \Delta$, the latter being a user-defined tolerance parameter (Table 1).

If a basis triangle is found to be a potential matcher of the triplet, their actual degree of similarity is calculated according to a simplified triangle overlay procedure related to the ComPharm[28] algorithm. Both the basis triplet i and the molecular triplet are represented as triangles of given (integer) edge lengths in the Euclidean plane. Each atom a_j in corner j is a source of a "pharmacophore field" $\psi_j(T,P)$ of type T. The intensity of such a pharmacophore field at any point P of space located at a distance d_{jP} from corner j is postulated to decrease according to a Gaussian function $\Phi(a_j,T)\exp(-\rho_{Tj}d_{jP}^2)$ of this distance, scaled by the extent $\Phi(a_j,T)$ to which atom a_j represents the pharmacophore type T. A 2D-superposition procedure translating and rotating the basis triangle with respect to the molecular triplet in order to achieve a relative alignment maximizing the covariance of these pharmacophore fields is launched after an initial triangle prealignment placing equivalent corners as closely together as possible. The fuzziness parameters ρ_T are treated as independent user-defined parameters of the method (Table 1).

Triplet-to-basis triangle overlay calculates a pharmacophore field covariance score ranging (in principle) between 0 (no match at all) and 1 (congruence). This score $O(i,\{a_k\})$ is an implicit function of the present pharmacophore types (and their intrinsic fuzziness parameters ρ_T), the nominal edge lengths of the basis triangle, and the actual topological distances within the atom triplet. In reality, covariance scores of 0 are never obtained, because the overlaid objects are filtered potential matchers. Actually, triangles sharing a common edge are guaranteed to score at least 0.67 (two conserved features out of three), no matter how far their third corners fall apart. Therefore, only covariance scores above the 2/3 threshold are considered:

$$O^*(i,\{a_k\}) = \max[0.0,O(i,\{a_k\}) - 2/3] \qquad (1)$$

The increment of the basis triplet population level due to the presence of a given atom triplet in m is proportional to $O^*(i,\{a_k\})$. Given the potentially large 2D-FPT fingerprint size, it is more practical to operate with integer rather than real population-level values. A scale-up factor of O^* has been introduced such that a basis triplet represented in a molecule by a single, perfectly congruent triplet reaches an arbitrary population level of 50. The ith 2D-FPT element $D_i(m)$, representing the total population level of a basis triplet i in species m, becomes

$$D_i(m) = \mathrm{int}[150 \times \sum_{\mathrm{atomtriplets}\{a_k\}\mathrm{in}m} O^*(i,\{a_k\})] \qquad (2)$$

2.2. Proteolytic Equilibrium-Dependent Fingerprint Buildup. The 2D-FPT generator uses ChemAxon's molecular reader classes[29] to input compounds in various formats and to standardize[30] connectivity and bond-order tables of compounds admitting several equivalent representations. Standardization rules were formally defined as chemical reactions in an XML configuration file read by the ChemAxon standardizer object (setup file in the Supporting Information).

On the basis of the standardized internal representations, the pharmacophore-type assignment procedure begins by submitting the current molecule to the ChemAxon pK_a plug-in.[31] This plug-in first predicts pK_a values for the ionizable groups of the molecule, then generates all of the possible conjugated acids and bases—the microspecies μ—together with their expected concentration $c\%(\mu)$, in percent, at the given pH (equal to 7.4 throughout this work). Next, the ChemAxon pharmacophore mapper tool (PMapper[15]) is used to flag the pharmacophore types within every microspecies. Specific pharmacophore flag matrices $F_\mu(a,T)$ and $\Phi_\mu(a,T)$ will be generated for each microspecies μ. PMapper is controlled by an XML file specifying flagging rules. A set of relevant substructures is specified as SMARTS[32] with labeled key atoms. Functional groups matching such sub-

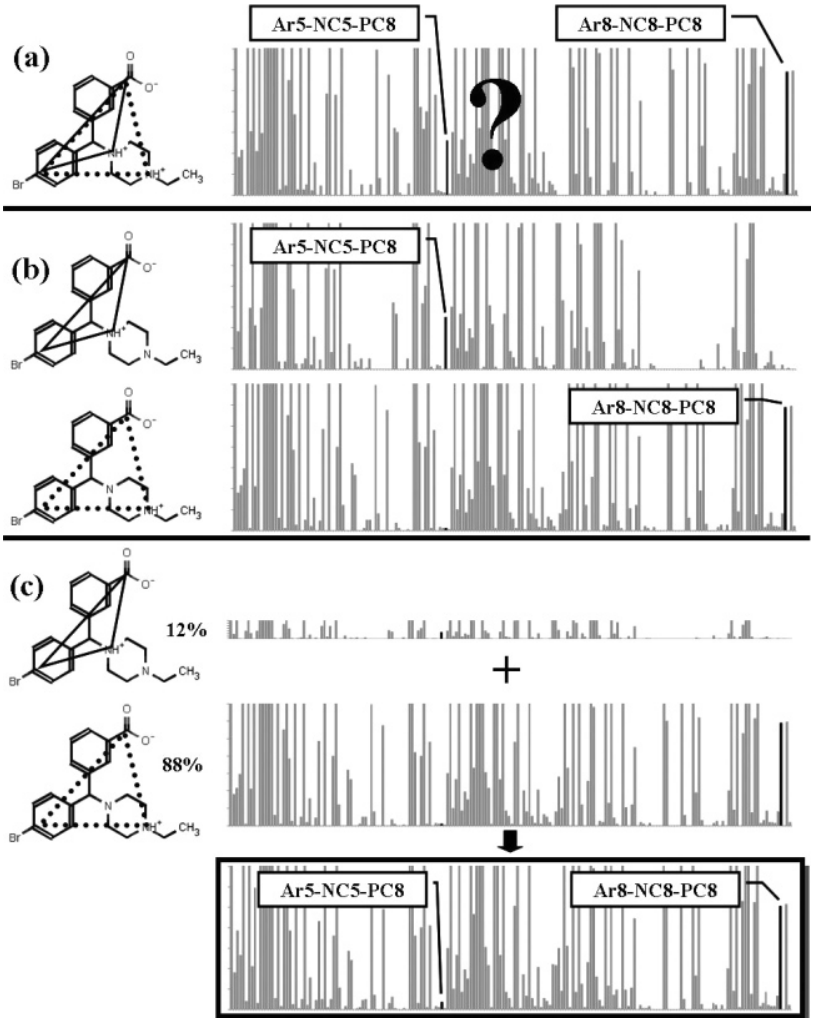

Figure 3. Graphical example of the principle of the construction of pK_a-sensitive 2D-FPT fingerprints: (a) rule-based pharmacophore flagging would assume three charged types in the molecule. Two triplets, both populated according to rule-based flagging, are localized in the sample fingerprint shown (bar sizes display population levels D_i, while the *x* axis enumerates the basis triplet counter *i*). Atom triplets that respectively contributed to each of the highlighted D_i's are marked in the structure. (b) The molecule actually appears at pH = 7 under the form of these two zwitterions. Each form carries only one of the triplets exemplified above. (c) The actual molecular fingerprint is obtained by weighed averaging of the microspecies fingerprints and, therefore, will resemble more the one of the zwitterionic forms predicted to occur at a concentration of 88% at equilibrium.

structures and the corresponding key atoms are detected in the molecule. An atom is assigned a given pharmacophore flag if it matches a certain substructure but not others. However, because formal charges are rigorously set in each microspecies, the assignment of PC and NC flags directly relies thereon. Any atom *a* carrying a positive formal charge (matching SMARTS "[*+]")—except for the nitrogen in nitro groups or nitrogen oxides—in the current microspecies μ will be assigned a flag $F_\mu(a,\mathrm{PC}) = 1$. By contrast, a classical flagging scheme would rely on the recognition of protonable group SMARTS and detect a potential cation even if it was not represented as such in the input molecule. Hydrogen-bond donor and acceptor flags are also set on the basis of specific rules pertaining to the microspecies. For example, a formally protonable N with a free electron pair, but not actually protonated in the current microspecies, will not be assigned an acceptor flag unless its pK_a value exceeds 5. Therefore, amide nitrogens will never be labeled as acceptors, but aniline nitrogens will unless they are strongly deactivated by electron-withdrawing groups. Oxygens always count as acceptors and −OH groups as donors. The recognition of aromatics is directly provided by ChemAxon's tools, while hydrophobes are defined as any carbon or halogen that is not aromatic and not charged.

The molecular fingerprint is thus obtained as a weighed average of microspecies fingerprints:

$$D_i(M) = \mathrm{int}\left[\sum_{\mathrm{microspecies}\,\mu\,\mathrm{of}\,M} \frac{c\%(\mu)}{100} D_i(\mu)\right] \qquad (3)$$

where $D_i(\mu)$'s are obtained for each microspecies μ, according to eq 2 using the specific pharmacophore flag matrix of the current microspecies for the estimation of the overlay score. The principle of proteolytic equilibrium-sensitive 2D-FPT buildup is illustrated in Figure 3. In the following, the notation D_i will, unless otherwise noted, implicitly refer to molecular average 2D-FPTs calculated according to eq 3.

2.3. FPT Similarity Scores. The appropriate choice of the similarity score $\Sigma(m,M) = f[\vec{D}(M),D(m)]$ comparing the 2D-FPT vectors of two molecules *m* and *M* is critical in order to ensure good NB. With classical metrics, such as the

Euclidean or Dice formulas, a first question is whether descriptors should be used as defined in eq 3 or after average/variance rescaling, leading to the set of normalized $\mathscr{D}_k(M)$: where $\alpha(D_k) = \langle D_k(m)\rangle_{\text{all } m}$ stands for the average of the

$$\mathscr{D}_k(M) = \frac{D_k(M) - \langle D_k(m)\rangle_{\text{all}m}}{\sqrt{\langle {D_k}^2(m)\rangle_{\text{all}m} - \langle D_k(m)\rangle_{\text{all}m}^2}} = \frac{D_k(M) - \alpha(D_k)}{\sigma(D_k)} \tag{4}$$

population level of triplet k over the BioPrint drugs and reference compounds[24] and $\sigma(D_k)$ stands for the corresponding variance. A further choice consisted in introducing a weighting scheme to specific triplets that are significantly populated in relatively few classes of compounds and absent from all of the others. These may be subject to an up to 10-fold increase of their relative importance with respect to ubiquitously present ones:

$$W_k = \min\left[10.0, \frac{\langle D_k(m)\rangle_{m\text{with}D_k(m)>0}}{\alpha(D_k)}\right] \tag{5}$$

Throughout this paper, structural dissimilarity metrics used with 2D-FPT will be denoted by the symbol Σ superscripted by the type of the metric, with an index informing on the use of normalized descriptors (N) as given in eq 4 or the weighting scheme (W) defined in eq 5. For example, the weighed Dice dissimilarity score using normalized descriptors is defined below, with N_T being the total number of basis triplets of the given 2D-FPT setup:

$$\Sigma^{\text{Dice}}_{N,W}(m,M) = 1 - \frac{2\sum_{k=1}^{N_T} W_{kk}(m)\mathscr{D}_k(M)}{\sum_{k=1}^{N_T} W_k {\mathscr{D}_k}^2(m) + \sum_{k=1}^{N_T} W_k {\mathscr{D}_k}^2(M)} \tag{6}$$

Indices N and W are omitted unless the metric explicitly relies on normalization and weighting and in cases of specific metrics (see below) or metrics from third-party software, whenever normalization and weighting options are no longer available.

The third, main, original contribution of this paper is the introduction of Σ^{FPT}, a specific metric of the dissimilarity of fuzzy pharmacophore triplets. Classical similarity scores, however, are generic metrics, applicable in arbitrary vector spaces, for example, independent of the actual nature of molecular descriptors associated with the degrees of freedom of the structure space. As this work will show, the specific design of a similarity scoring scheme based on an actual interpretation of the information in the fingerprint may significantly improve NB.

Concretely, the knowledge that $D_i(M)$ represents population levels of basis triplets, and that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence, will be actively exploited. A first prerequisite in this sense is the introduction of a measure of the significance $S_k(M)$ of a triplet k for a molecule M, with respect to the observed averages and variances of each triplet population level:

$$S_k(M) = \begin{cases} 0 & \text{if } D_k(M) < 0.7\alpha(D_k) \\ 1 & \text{if } D_k(M) > 0.7\alpha(D_k) + \sigma(D_k) \\ \dfrac{D_k(M) - 0.7\alpha(D_k)}{\sigma(D_k)} & \text{otherwise} \end{cases} \tag{7}$$

A triplet k in a pair of molecules (m,M) may fall into one of the following categories: shared (++), for example, significant—in the above-mentioned sense—for both m and M, null (− −), for example, not significant for either, and exclusive (+ −), for example, significant for either m or M but not for both.

Rather than assigning it to one and only one of these, its fuzzy levels τ of association to each of the categories are defined in order to always sum up to 1:

$$\tau_k^{++}(m,M) = \frac{S_k(M)\, S_k(m)}{\text{norm}}$$

$$\tau_k^{-\,-}(m,M) = \frac{[1 - S_k(M)][1 - S_k(m)]}{\text{norm}}$$

$$\tau_k^{+\,-}(m,M) = \frac{|S_k(m) - S_k(M)|}{\text{norm}}$$

$$\text{norm} = S_k(M)\, S_k(m) + [1 - S_k(M)][1 - S_k(m)] + |S_k(m) - S_k(M)| \tag{8}$$

The fraction of triplets in a category c therefore becomes

$$f^c(M,m) = \frac{1}{N_T}\sum_{k=1}^{N_T} \tau_k^c(M,m) \tag{9}$$

Classical distance functions are typically calculated on the basis of the differences observed for each component k of the molecular descriptors $\delta_k(m,M) = |\mathscr{D}_k(m) - \mathscr{D}_k(M)|$. The herein introduced originality consists of a separate monitoring of these contributions for the shared, exclusive, and null triplets. Rather than simply summing up all $\delta_k(m,M)$ contributions (leading to a Hamming-type dissimilarity score), weighed partial distances $\Pi^c(m,M)$ are estimated in order to monitor how much of the difference stems from triplets in each category:

$$\Pi^c_{W,N}(m,M) = \frac{\sum_{k=1}^{N_T} W_k \tau_k^c(m,M)\, \delta_k(m,M)}{\sum_{k=1}^{N_T} W_k} \tag{10}$$

The working hypothesis adopted here was that a meaningful dissimilarity score can be expressed as some linear combination involving certain of the three fractions defined in eq 9 as well as the three partial distances (eq 10). Successive trials monitoring the NB of the resulting metric with respect to a subset of the entire learning set (see the following section) led to the following expression:

$$\Sigma^{\text{FPT}}(m,M) = 0.1323{\Pi_{W,N}}^{+-}(m,M) + 0.6357{\Pi_{W,N}}^{++}(m,M) + 0.2795[1 - f^{++}(m,M)] \tag{11}$$

The NB of the herein proposed scoring scheme was benchmarked with respect to classical dissimilarity metrics in various validation studies.

2.4. Experimental Data and Validation Studies. The performance of 2D-FPT in similarity searches has been assessed and compared to that of other 2D and 3D pharmacophore descriptors, following the previously published methodology[16] for monitoring the NB of in silico similarity scores. In the current work, activity profiles of 2275 nonproprietary (commercial drugs and drug precursors) molecules from the BioPrint database of Cerep were used to calculate the activity dissimilarity scores $\Lambda(m,M) = f[\vec{p}(M),\vec{p}(m)]$ expressing the amount of difference between the response patterns of the two molecules with respect to the considered battery of targets. Profiles $p_t(m)$ report measured $pIC_{50} = -\log IC_{50}$ (mol/l) values of every molecule m against each of $N_{targets} = 154$ different biological targets t (enzymes, receptors). $p_t(m) = 9/6/3$ means that molecule m is a nano-/micro-/millimolar binder of t, respectively. The actual algorithm used for estimating the activity profile dissimilarity score $\Lambda(M,m)$ is outlined in Appendix A.

An alternative NB study has been conducted on the basis of an activity profile compiled from publicly available data sets[25−27] (see the Supporting Information). Unlike the highly diverse BioPrint data, this study features a compilation of 112 compounds tested on the angiotensin converting enzyme (ACE), 111 on acetylcholine esterase (AchE), 163 on the benzodiazepine receptor (BzR), 321 on cyclooxygenase-II (Cox2), 641 on dihydrofolate reductase (DHFR), 66 on glycogen phosphorylase B, 67 on thermolysin, and 88 on thrombin (THR)—a total of 1569 molecules from eight activity classes. Each activity class is represented by a typical QSAR set, featuring variations of one or a few central scaffolds and including both actives ($pIC_{50} > 6$) and inactives in roughly equal proportions. The actual compilation of 1569 compounds has been realized by standardizing[30] the structures of molecules from the cited sources, then merging the sets and discarding duplicate compounds with conflicting activity data (associated activity values for a same target differing by more than one pIC_{50} log). In the absence of experimental data about the affinity of a compound m with respect to a target t, inactivity was assumed and $pIC_{50}(m,t)$ set to 3.5 in order to fill up the structure−activity profile matrix. Under this assumption, activity dissimilarity scores $\Lambda(M,m)$ were calculated according to Appendix A, with the conversion function ψ in equation A6 modified so as to return 1.0 only if its argument exceeds 12.5% of the number of targets in the profile (that is, one difference with respect to eight targets—the 5% threshold used with the much larger BioPrint profile makes no sense when $N_{targets} = 8$). With these specifications, an active compound M appears as equally distanced—at $\Lambda(M,m) = 1$—from any confirmed inactive of its own class, as well as from all of the molecules belonging to different classes. $\Lambda(M,m) = 0$ only if m and M are both actives within the same class. An inactive is set at $\Lambda(M,m) = 0.1$ from any other inactive, within its own series or not, but such pairs were consistently discarded, like in the BioPrint study case.

In the comparative NB studies, the experimental activity dissimilarity $\Lambda(M,m)$ is confronted to various calculated molecular dissimilarity scores $\Sigma(M,m)$. The purpose of such a benchmark is assessing in how far molecules (m,M) that are predicted to be neighbors in a given "structure space"—low $\Sigma(M,m)$—are systematically found to also be neighbors in "activity space"—low $\Lambda(M,m)$. The statistical formalism used to quantitatively evaluate NB is briefly revisited in Appendix B. NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. For simplicity, the plots were truncated at $\chi = 0.4$—displaying only the high-consistency range. Therefore, the characteristic U shape of $\Omega-\chi$ plots[16] may not always be apparent, but this is of little relevance for the discussion: the rule of thumb for the interpretation of the obtained graphs is that low Ω at high χ signals good neighborhood behavior.

2.4.1. Benchmarked Descriptors and Metrics. The NB of the 2D-FPT has been compared to the ones of different two-point pharmacophore descriptors, including fuzzy bipolar pharmacophore autocorrellograms (FBPA),[9] a 3D descriptor, and ChemAxon's topological fuzzy pharmacophore fingerprints.[15] The latter were calculated using both the recommended standard configuration (PF) and employing the "-R/--ignore-rotamers" (PFR) option of the ChemAxon descriptor generation tool. This option suppresses the default hypothesis according to which more fuzziness is applied when generating descriptor elements corresponding to more distanced atom pairs, as these have more options to experience important relative movements in the real molecule subjected to thermal agitation. ChemAxon's Chemical Fingerprints[33] (CF) were also used for benchmarking, as a representative of fragment-based fingerprints. To explicitly monitor the benefit of the novel-type flagging technique used with 2D-FPT, an alternative FPT relying on the same rule-based procedures used with PF/PFR has been generated. Molecular dissimilarity scores based on third-party descriptors were calculated according to the metrics best adapted for each—the Tanimoto score with ChemAxon's PF and CF and the fuzzy FBPA metric, respectively. XML setup files used for PF and CF descriptor and dissimilarity score calculations (PF.xml and CF.xml respectively) are included in the Supporting Information.

2.4.2. Virtual Screening of Seeded Compound Collections. A set of 50 000 random compounds—excluding organometallic derivatives and compounds of molecular mass above 1000 g/mol—from the MayBridge[34] vendor catalog were used as a reference chemical space to which molecules of known activities were added: (1) 194 compounds with reported c-Met tyrosine kinase activities from the literature,[35−37] including 72 actives with $IC_{50} \leq 10^{-7}$ M and (2) 460 molecules that were tested against the dopamine D2 receptor[38] (219 with $IC_{50} \leq 10^{-7}$ M). Both sets covered activity ranges from nanomolar to low millimolar values of IC_{50}. For each, the pharmacophorically most diverse three representatives were picked out of the respective subsets of very potent inhibitors ($IC_{50} < 10^{-8}$ M) and used as lead compounds for virtual screening according to both the 2D-FPT (FPT-2) and the PF-based Tanimoto metrics. The numbers of both confirmed actives ($IC_{50} \leq 10^{-7}$ M) *and* confirmed inactives ($IC_{50} > 10^{-7}$ M) were monitored within the sets of 200 nearest neighbors from the seeded chemical space found by each metric around each of these six leads.

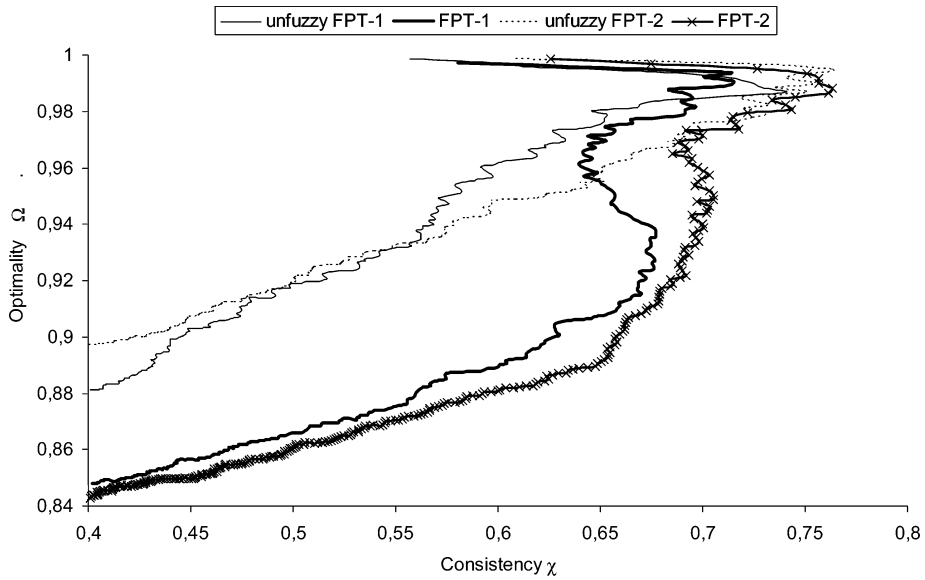

Figure 4. Comparative $\Omega-\chi$ plots illustrating the improvement of NB upon enabling the fuzzy mapping of atom triplets onto basis triplets, for both fingerprint versions FPT-1 and FPT-2, using the 2D-FPT specific similarity score Σ^{FPT} (BioPrint data set).

3. RESULTS AND DISCUSSIONS

3.1. The Importance of Fuzzy Mapping. To explicitly quantify the importance of fuzzy atom triplet mapping onto the basis triangles, the fuzziness factors ρ of considered FPT versions from Table 1 were temporarily set to 5.0 in order to generate comparative $\Omega-\chi$ plots for the corresponding unfuzzy fingerprints (the specific Σ^{FPT} score was used in all cases). At such high values of ρ, atom triplets will strictly highlight basis triplets of identical edge lengths. They will fail to highlight any basis triplet if the given combination of interatomic separations is not represented in the basis set. The corresponding curves in Figure 4 differ very little at their origins, where the selected pairs mostly include analogues with the same molecular scaffold and therefore are made of almost exactly the same atom triplets. However, the use of fuzzy logics is essential for extending the selection beyond these very first close analogues, to encompass pairs of compounds for which the underlying pharmacophore pattern similarity is not necessarily backed by a skeleton similarity. With fuzzy logics, many more activity-related compound pairs can be successfully picked without allowing pairs of different activities to enter the selection. Ω is observing a significant decrease without a loss of consistency, which is not seen when fuzzy mapping is turned off.

3.2. Importance of the pK_a-Dependent Fingerprint Buildup Strategy. The introduction of pK_a-dependent pharmacophore-type weights is expected to significantly contribute to the chemical meaningfulness of FPT. For example, a rule-based "educated guess" typically used to recognize potentially ionized groups in organic compounds would rely on the axiom that aliphatic amines are protonated, for example, must be flagged as cations and donors. Accordingly, N-alkylpiperazine-containing organic compounds will be assumed to harbor a cation−cation pair (see example in Figure 3). However, at pH = 7, only one of the two nitrogens is likely to carry a proton, its charge preventing the second one to do so. The cation−cation pair hence only appears in a minority of molecules, and its weight in the overall pharmacophore pattern should be adjusted accordingly. Piperazine may in reality be closer related to cyclohexylamine or morpholine than the rule-based pharmacophore pattern matching would suggest. Of course, rules can be tentatively optimized to avoid these kind of pitfalls: for example, the ChemAxon default pharmacophore mapping rules do not include tertiary amines into the cation category. This makes sense in medicinal chemistry, where the majority of amino groups in drugs are tertiary. The undue hypothesis of polycation patterns in the pharmacophore motif may hence be avoided, though at the cost of failing to perceive the similarity between secondary and tertiary amines.

An accurate prediction of the ionization status of protonable groups is a prerequisite for the success of the herein advocated flagging strategy. The NB of the fingerprints relying on ChemAxon's pK_a prediction plug-in outperforms the strategy of rule-based protonation state setup (Figure 5). This is thus an indirect proof of the accuracy of the pK_a prediction tool, offering an accurate estimation of expected protonation states. The rules used to build the alternative 2D-FPT (all other setup parameters being equal to FPT-1 values) were ChemAxon's default rules, the same used to construct the PF two-point pharmacophore fingerprints. A total of 59 pairs of compounds with identical activity profiles, ranking among the top 1000 most similar according to the pK_a-based approach, would lose their top-ranking positions and regress by more that 10000 ranks in the ordered pair list according to the rule-based method. Conversely, 50 activity-related pairs are perceived as similar by the rule-based metric, but not by the pK_a-based scoring scheme. The significant differences appear with respect to the distribution of activity-unrelated compound pairs. A total of 14 "violators" of the pK_a-based scheme (pairs with $\Lambda = 1$ but nevertheless ranked among the top 1000) are correctly reranked among the structurally dissimilar by the rule-based procedure. By contrast, 100 of the rule-based violators are successfully eliminated by the pK_a-based approach. Four typical examples of these latter ones are given in Figure 6. The similarity of compound pair a is clearly overstated by

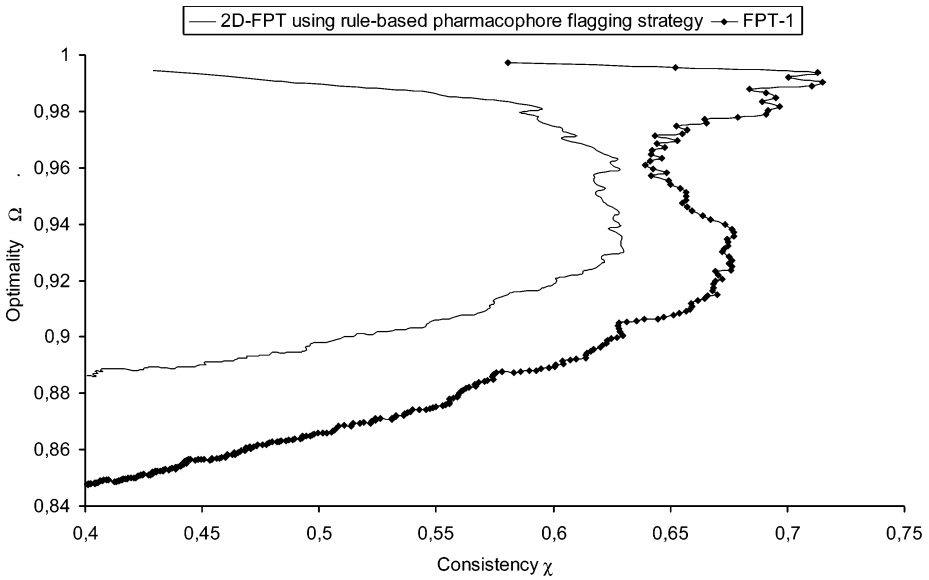

Figure 5. Standard rule-based flagging strategy of ionizable groups outperformed by the herein introduced pK_a-dependent fuzzy-type assignment procedure.

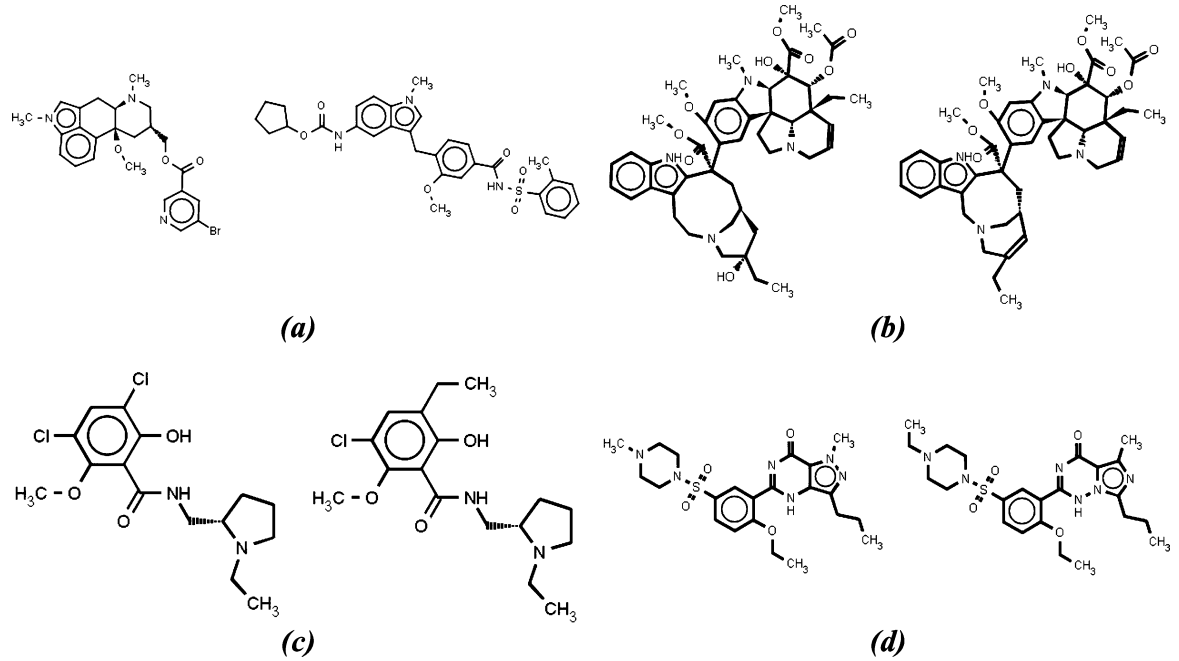

Figure 6. Examples of BioPrint compound pairs that look similar and are ranked among the top 1000 structurally closest pairs by the rule-based pharmacophore flagging scheme but, in reality, display radically different activity profiles and are correctly perceived as structurally different by the pK_a-based pharmacophore flagging scheme.

the rule-based scoring scheme, which regards both molecules as neutral species—acylsulfonamides are not declared as potential anions, and tertiary amines are not declared as cations in the ChemAxon default setup file pharma-frag.xml. Pair a stands thus for the numerous examples of activity-unrelated violator pairs that might have been avoided by redefining some of the flagging rules. In cases b, c, and d, however, pharmacophore dissimilarity cannot be accounted whatsoever by detailed flagging rule definitions: subtle substitution effects are seen to trigger relatively small pK_a shifts, but with dramatic impacts on the overall populations at proteolytic equilibrium. In compound pair c, the dissimilarity stems from the much more important ionization of the dichlorophenol compared to the monochlorophenol. While the left-hand compound mainly appears (according to the ChemAxon pK_a tool) under its zwitterionic form at pH = 7.4, the right-hand counterpart is predominantly positively charged. Even more dramatically, in example d, the addition of a simple methyl group enhances the protonation of the tertiary amine (70% cation at pH = 7.4 compared to 40% only in the left-hand molecule). Unless this effect is explicitly accounted for, a pharmacophore dissimilarity metric might never be able to explain the important activity differences observed upon the addition or deletion of a single hydrophobic center. Of course, the success of the approach relies on the precise pK_a estimation, or else the overestimated equilibrium population shifts that fortuitously explain observed activity differences might as well prevent the metric from recognizing the real pharmacophore similarity of activity-related pairs. As many com-

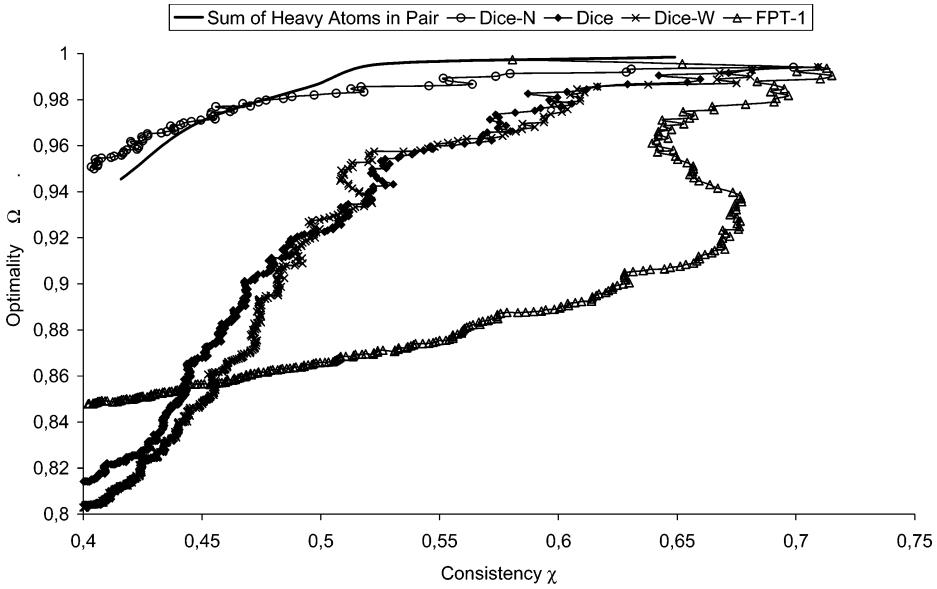

Figure 7. Comparative $\Omega-\chi$ plots of the NB (BioPrint data set) of various similarity scores with 2D-FPT (FPT-1 setup). Considered metrics are variants of the Dice formula: Σ^{Dice} ("Dice" in Figure legend), Σ_N^{Dice} ("Dice-N" in legend), and Σ_W^{Dice} ("Dice-W" in legend), as well as the 2D-FPT specific similarity score Σ^{FPT} ("FPT" in legend, eq 11).

pounds in this study are well-known drugs and reference molecules that are likely to have served for the pK_a tool calibration, further validation on the basis of original compound collections might be welcome. This notwithstanding, it can be concluded that one of the notorious limitations of pharmacophore-based similarity, the inability to explain activity shifts accompanying slight substitution pattern changes—a thorny issue raising fundamental questions about the validity of the neighborhood principle—might be successfully overcome in quite numerous cases of pK_a shift-related activity differences.

3.3. The Relative Performance of the Specific FPT Similarity Score. The NB of the various similarity scoring schemes using 2D-FPT (built according to setup 1 in Table 1) has been assessed, the results being shown in Figure 7.

The uppermost, solid curve represents the behavior of a fake dissimilarity score equaling the sum of heavy atoms in the molecule pair (*m*,*M*). It is nevertheless a well-shaped $\Omega-\chi$ plot, proving that activity-relatedness is statistically more likely to occur within subsets of small molecule pairs. This size effect is due to the fact that the smaller (~10 heavy atoms) of the employed molecules are unlikely to be strong binders to targets in the activity panel. Activity profiles of such compounds will be mostly empty, and their comparison returns low Λ scores (of about 0.1). Significant accumulation of such compound pairs at the top of the by-size sorted pair list ensures a significant consistency level of more than 60% within the top 20 lightest pairs (right-most point on the curve). Compound pairs with Λ scores of 0 (hitting common targets) are not contributing to these initial high consistency scores. The artifactual NB of size would have been even more marked if a bonus for binding to a same target would not have been included in Λ (results not shown).

Any rational pair selection strategy must therefore do better than (e.g., lay below) the size-driven NB curve. This is, unsurprisingly, not the case for the Dice metric based on normalized descriptors, which is quite sensitive to the complexity of the pharmacophore patterns of molecules, and implicitly to molecular size (see Figure 8). Small molecules with few populated triplets run an artificially high chance to be ranked as very similar: at $D_k(m) = 0$, $\mathscr{D}_k(m)$ simply relates to $-\alpha_k(m)$. The lesser the number of populated triplets is, the closer to the vector of average triplet populations—and the more correlated—the vectors $\mathscr{D}_k(m)$ and $\mathscr{D}_k(M)$ will be.

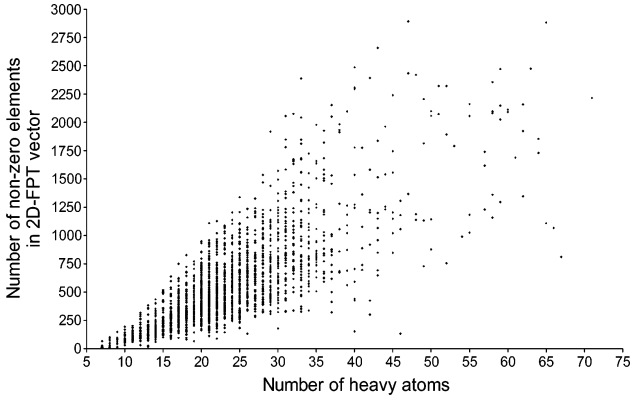

Figure 8. Dependence of the number of populated triplets on molecule size.

The same effect can be noticed with Euclidean scores (not shown). When $D_k(m) > 0$ and $D_k(M) > 0$, the chances that $D_k(m) = D_k(M)$ are quite small. Molecule pairs with a significant common set of populated basis triplets will, because of the summation of small but numerous residuals $\delta_k(m,M)$, typically end up at a higher Euclidean dissimilarity than pairs of small molecules with $D_k(m) = D_k(M) = 0$ for an overwhelming majority of triplets *k*. For example, the introduction of a methyl group in a large molecule *M* would trigger changes in the population levels of many more triplets *k* than the introduction of the same $-CH_3$ in a small compound *m*. Therefore, the calculated Euclidean distance score for a methyl/normethyl compound pair would counterintuitively increase with molecule size.

The Dice scores with or without the weighting of rare pharmacophore triplets can be successfully used to compare brute 2D-FPT, although they are clearly outperformed by the spe-

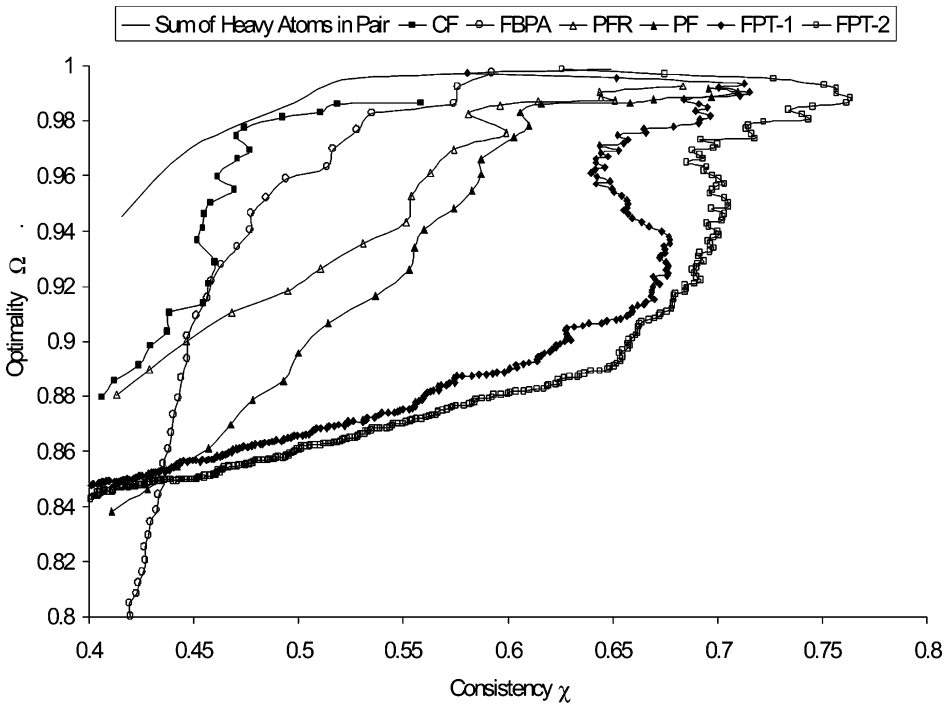

Figure 9. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (both setups, using the specific Σ^{FPT}) with respect to other descriptors and associated metrics (BioPrint data set).

cific FPT metric. In the Dice formula using 2D-FPT without any further norming or rescaling, the main criterion controlling dissimilarity is the number of common nonzero descriptor elements, as these are the only contributing to the sum of $D_k(m)D_k(M)$. Any molecules having no nonzero D_k values in common will be considered 100% dissimilar. However, two large molecules with less-sparse 2D-FPT vectors are much more likely to achieve some fortuitous overlap of their fingerprints than two small molecules. Even if an overwhelming number of exclusively populated D_k's exist, having $D_k(m)D_k(M) > 0$ for at least one k automatically ensures that such a molecule pair will nevertheless be ranked as more similar than any pair of small molecules with no shared triplets at all.

A general problem in molecular similarity scoring—be it molecular descriptor comparison or activity profile matching—appears to be the appropriate handling of the uncertain "null" situations describing the absence of an item (pharmacophore triplet, affinity with respect to a target) from both molecules. On one hand, it may be argued that the two compounds share the absence of an item, which makes them more similar. On the other, sharing the presence is clearly a stronger proof of similarity than sharing the absence, and the question is, how much stronger? Also, how can shared presence and shared absence be counterbalanced against the number of differences observed in the fingerprint, to achieve a meaningful final score?

The excellent NB of the dedicated dissimilarity score defined in eq 11 suggests an appropriate balancing of the contributions for the specific case of 2D-FPT. The dissimilarity score Σ^{FPT} is seen to increase in response to (a) observed differences between population levels of exclusively populated basis triplets and (b) observed differences between population levels of shared triplets. The coefficient of the latter is more important—however, it is the former that statistically contributes the most to the dissimilarity scores because situation a occurs more often.

Furthermore, Σ^{FPT} decreases as the total fraction of shared triplets increases—with the effect that $\Sigma^{FPT}(M,M)$ will decrease with molecule size: larger molecules (with richer pharmacophore patterns, strictly speaking) are "more similar to themselves" than smaller ones. This is not paradoxical if we give up considering Σ^{FPT} as a similarity metric, but consider it as a substitution score not unlike the ones used for sequence matching in bioinformatics:[39] the conservation of the rarer, larger, and functionally specific tryptophane in two sequences is seen as more significant and given a larger bonus than the conservation of a ubiquitous alanine.

3.4. Neighborhood Behavior of 2D-FPT, Compared to the Other Descriptors. Figure 9 compares the NB of 2D-FPT using Σ^{FPT} to that of other descriptor spaces and metrics. In can be seen that CF chemical fingerprints, which are tailored for (sub)structure recognition, do not fare better than size-driven artifacts. All of the pharmacophore descriptors, however, perform better than cumulated size. At low selection sizes (large Ω), PF outperform the fuzzy three-dimensional FBPA. However, although the latter metric tends to be too permissive (allowing compound pairs with different activities among its top-scoring pairs), it is nevertheless able to retrieve a maximum of existing activity-related pairs while maintaining a reasonable consistency of the selection (deep Ω minimum). Interestingly, applying higher fuzziness levels for more distant pharmacophore point pairs (default behavior in ChemAxon's pharmacophore fingerprint calculator) seems counterproductive in this benchmarking test: better results (PFR) are obtained when this approach is switched off.

It is remarkable that the 2D-FPT curves and notably the one obtained with the smaller triangle basis set (FPT-1) originate at relatively low consistency levels. As the selection is extended, the fraction of activity-related among the co-

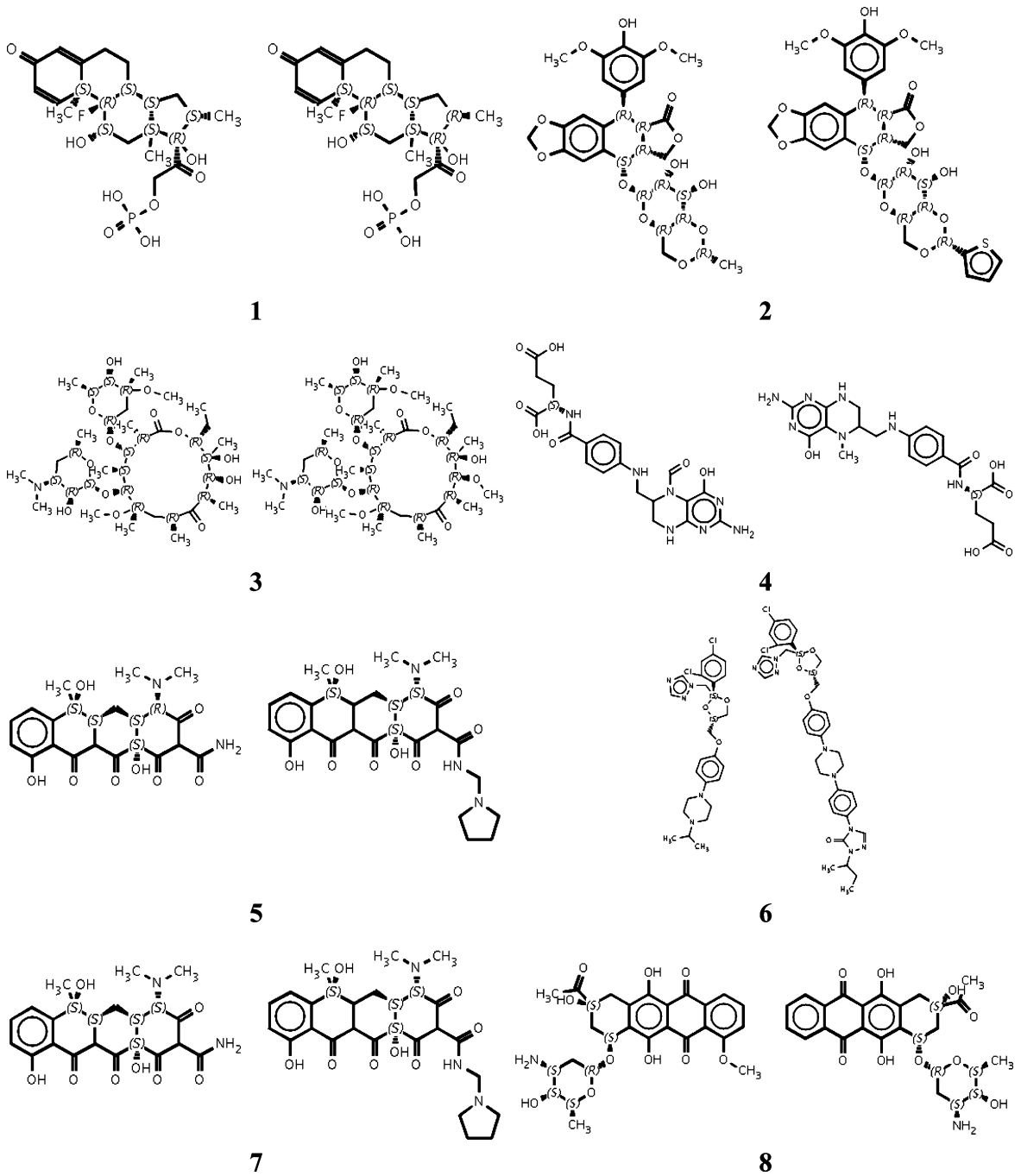

Figure 10. The eight pairs with highly dissimilar activity profiles found among the 50 most similar pairs according to 2D-FPT similarity scoring (FPT-1 setup).

opted pairs becomes much larger than that seen within the first top scorers. At high consistency values (0.5−0.7), significantly more activity-related compound pairs are retrieved by 2D-FPT than by any of the other scoring schemes.

Such behavior might be expected with topological descriptors such as 2D-FPT, because pairs of diastereomers (M,M*) score as much as a compound scores with respect to itself: $\Sigma^{FPT}(M,M^*) = \Sigma^{FPT}(M,M)$. The hypothesis that the initial inconsistency is due to the accumulation of activity-unrelated diastereomer and enantiomer pairs at the top of the similarity-sorted pair list must however be discarded. PFs, for example, are also topological distance-based and use a classical Tanimoto-based scoring scheme, so that $\Sigma^{PF}(M,M^*) = \Sigma^{PF}(M,M) = 0$ and diastereomers are always top scorers. However, the very high consistency of the right-most data point of the PFR curve proves that the 105 compound pairs with $0.00 \leq \Sigma^{PFR} < 0.01$, the herein included pairs of diastereomers, are not overwhelmingly activity-unrelated.

Actually, Σ^{FPT} no longer guarantees diastereomer pairs to rank among top scorers. $\Sigma^{FPT}(M,M) > 0$ decreases with the complexity of M, and pairs of slightly differently substituted analogues (M,M') sharing a highly complex pharmacophore pattern may score better than pairs of less complex molecules (m,m*) with identical fingerprints. Although $\Pi^{+-}(m,m^*) = \Pi^{++}(m,m^*) = 0$, having $f^{++}(M,M') > f^{++}(m,m^*)$ may eventually let the pair of close analogues score lower Σ^{FPT} values than the pair of diastereomers. The consistency inversion observed with 2D-FPT is, unexpectedly, not a

consequence of ignoring stereochemical information but actually stems from pairs of closely related analogues of very high molecular complexity. Among the best-ranked 100 pairs of compounds according to the FPT-1 setup of 2D-FPT scoring scheme, 66 have $\Lambda > 0.2$, 30 have $\Lambda > 0.5$, and 15 have $\Lambda > 0.8$. By contrast, in the pair subset ranked from 100 to 200, there are only 21 at $\Lambda > 0.2$, 13 at $\Lambda > 0.5$, and 6 at $\Lambda > 0.8$, for example, less than half as many NB violators than in the first 100 pairs. Violator pairs are, beyond doubt, chemically similar (to the point that finding the difference when looking at the structures is not always easy; Figure 10, except for examples 6 and 7, where substitution differences involve the introduction of a heterocycle and a cationic group, respectively). It is difficult to "blame" the 2D-FPT metric for having selected them. However, such "me-too" close analogue pairs are always among the top scorers of all of the similarity metrics, including PF and FBPA, but they are not seen to distort either of the herein-obtained NB curves. It can be safely assumed that, statistically speaking, closely related analogues differing in terms of either the stereochemistry or minor substituent changes tend to have similar biological activities, the exceptions to this rule being relatively rare (but widely publicized[40]). The previous section showed that 2D-FPTs are able to successfully explain some of these "activity cliffs" on the basis of predicted pK_a shifts. It appears however that they also tend to specifically pinpoint another subset of activity cliffs, pertaining to a specific series of close analogues that tend to score better than the ubiquitous activity-related "me-too" pairs. The 2D-FPT score-driven ranking of the BioPrint compound pairs evidenced a top-ranking subset of highly complex and very similar compound pairs with an increased propensity to form activity cliffs versus that of "typical me-too" pairs. At this point, it is however unclear whether this finding may be generalized to suggest that more-complex molecules are more likely to have their biological properties strongly affected by small chemical alterations. This is certainly not true with respect to overall physicochemical properties: methylation of a macrocycle like the third example in Figure 10 would hardly affect properties such as the octanol−water partition coefficient; by contrast, the methylation of methanol leads to the physicochemically different dimethyl ether. It is however important to remark that most of the compound pairs in Figure 10 are natural compounds or derivatives of natural compounds, optimized by Darwinian evolution to be perfect binders to a given target. From this viewpoint, it seems understandable that any small chemical alteration on the natural ligands may have dramatic changes in affinity. Synthetic drug molecules appear to be much less well-adapted to their targets and therefore, statistically spoken, much more tolerant to structural variations. 2D-FPT might provide a very useful metric for molecular complexity and implicit lead-likeness or drug-likeness—issues[41] that will be explored elsewhere.

The second parametrization attempt FPT-2 turned out to be more successful, but although the subsets of top scorers are significantly less marked by the accumulation of activity-unrelated pairs, the previously discussed consistency inversion does not vanish. Its better performance can be mainly ascribed to the shift of the minimal and maximal topological edge lengths from 2 to 4 and from 12 to 15, respectively. Monitoring triplets including directly bound, geminal or vicinal atoms does not enhance NB. This makes sense: binding pharmacophores typically include anchoring points from different parts of the ligand. Triplets involving, for example, both the carbonyl =O and the hydroxyl −OH in a hydroxamic acid RC(=O)−NH−OH are not accounted for in any of the versions—a specific fitting for metal enzyme inhibitors might prove necessary under these circumstances. The coverage of long-range molecular triplets seems to be very important: it also seems a good idea to extend the size of actually considered molecular triplets by $e = 2$ more bonds beyond E_{max}.

The initial choice of a grid of basis triplets having a mesh size (edge increment E_{step}) of 2 appears to be the good compromise. An E_{step} of 3 would have reduced the basis set size dramatically—however, molecular triangles with edge size values not appearing in the basis triplets would have been at risk to fall through the grid meshes, in failing to match any one of the basis triplets. Successful 2D-FPT setups with $E_{step} = 3$ may exist but must be actively searched for in the setup parameter space. $E_{step} = 1$ would, on the contrary, engender much larger grid sizes, thus causing significantly more practical problems with the handling of the resulting descriptors. Given the excellent behavior at $E_{step} = 2$, potential benefits of denser basis sets are unlikely to outweigh the descriptor size-related inconveniences.

A first key observation in Figure 11, monitoring the NB of various metrics with respect to the public data set obtained by merging eight independent QSAR series, is the much lower Ω values compared to what had been seen within the BioPrint set. Unsurprisingly, detecting structurally similar pairs of related activities is a much harder problem within the diverse set of drugs than within an artificially constructed set of series of analogues around a limited number of scaffolds. In this latter case, a simple discrimination between structural families—telling benzodiazepine-like chemotypes apart from acetycholine-like ligands and so forth—is sufficient to ensure significant NB. There are, for example, 65 active and 47 inactive ACE binders in the set; for example, $65/1569 = 4.14\%$ of ACE actives in the entire set. Any metric that would consistently score lower dissimilarity between any two ACE set members than between an ACE and a non-ACE compound pair effectively discriminates between the ACE set and the rest of compounds. Within the ACE set, the rate of actives is however $65/112 = 58\%$, which represents a $58/4.14 = 14$-fold enrichment in actives. Under these circumstances, dissimilarity scoring based on chemical fingerprints does display a significant NB, in sharp contrast to the observations made on the BioPrint set. The discrimination between the various chemical families that make up the public data set is readily achievable by all three metrics monitored in Figure 11: all of them avoided ranking any of the pairs of compounds from different series within the top 550 pairs corresponding to the checkpoints highlighted on the plots. All NB violators—in the sense of $\Lambda(m,M) > 0.5$—encountered at these checkpoints are intraseries activity cliffs regrouping an active and a structurally very close inactive. Within the top 550 pairs selected by the CF metric, the 128 observed NB violation instances break down into 15 ACE, 27 AchE, 5 BzR, 20 Cox2, 43 DHFR, and 18 THR compound pairs. Pharmacophore-based metrics should go beyond activity class recognition and successfully tell apart actives and inactives on the basis of a common scaffold. This

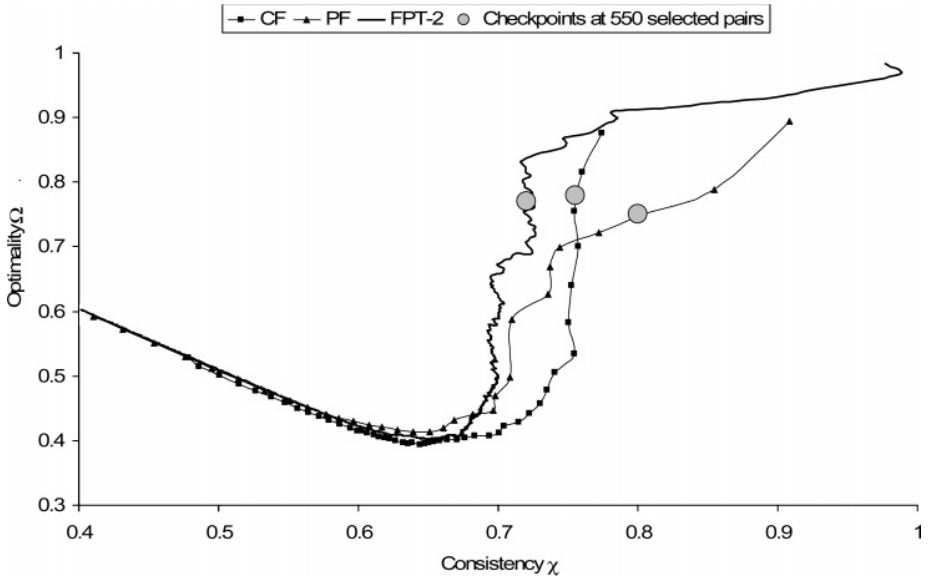

Figure 11. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (setup FPT-2, using Σ^{FPT}) with respect to ChemAxon chemical and pharmacophore descriptors and associated metrics (public data set regrouping 1569 compounds from eight QSAR series).

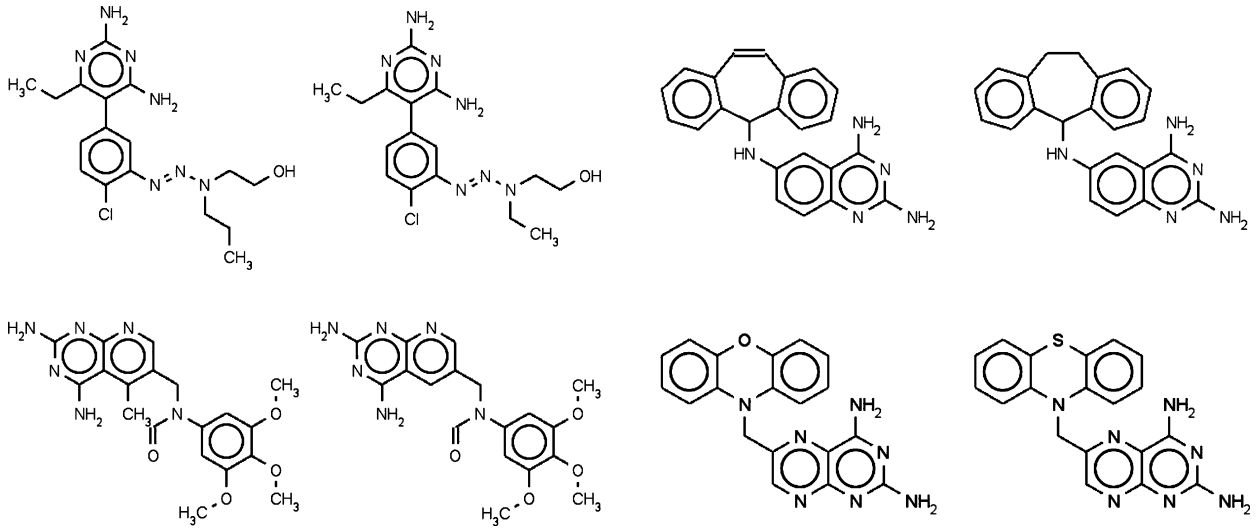

Figure 12. Typical "activity cliffs" of dihydrofolate reductase—very similar compound pairs with significantly differing DHFR activities ($\Lambda > 0.5$). Such compound pairs are consistently perceived as similar by all metrics—however, only the Σ^{FPT} formalism ranks these relatively complex compound pairs among the top 550.

is indeed observed with both PF and FPT metrics: both of these and particularly the latter reach out into higher consistency domains, not accessible to the CF approach. Unlike in the BioPrint study case, PF-driven NB reaches relatively better optimality scores at a same consistency or relatively higher consistencies at the same selection size (0.8 instead of 0.7 for the top 550 selected pairs, see checkpoints). An analysis of NB violators reveals that PF retrieved 92 such pairs within the top 550: 7ACE, 4 AchE, 3 BzR, 59 Cox2, and 19 DHFR, whereas FPT retrieved 138: 5 ACE, 48 Cox2, 83 DHFR, and 2 THR. The FPT approach thus experiences a sharp decrease of its NB criteria because of a local accumulation of DHFR activity cliffs, some typical examples of which are depicted in Figure 12. These are clearly structurally highly related compounds scoring very low dissimilarity values within both FPT and PF formalisms. However, only the former score includes a bonus for pharmacophore complexity, or it can be seen that DHFR ligands are among the most complex compounds in this set. DHFR pairs are therefore relatively better ranked than other intraset pairs when using FPT. Unfortunately, DHFR appears to display a rugged structure−activity landscape ridden by activity cliffs that cannot be conveniently explained by any of the herein explored metrics. This may be an illustration—but still no definite proof—of the possible correlation between ligand complexity and the propensity for activity cliffs, previously cited as an envisageable explanation for the observed consistency inversion of the FPT metric within the BioPrint set.

3.5. Virtual Screening Results of Seeded Compound Collections. Such simulations directly address the ability of the metrics to discover actives from databases but are less well-suited for rigorous benchmarking than the general NB analysis reported previously, insofar as the following are concerned:

• While a retrieval of a maximum of hidden actives among the top neighbors of each lead compound is desirable, it is not clear how many of the hidden actives are genuinely

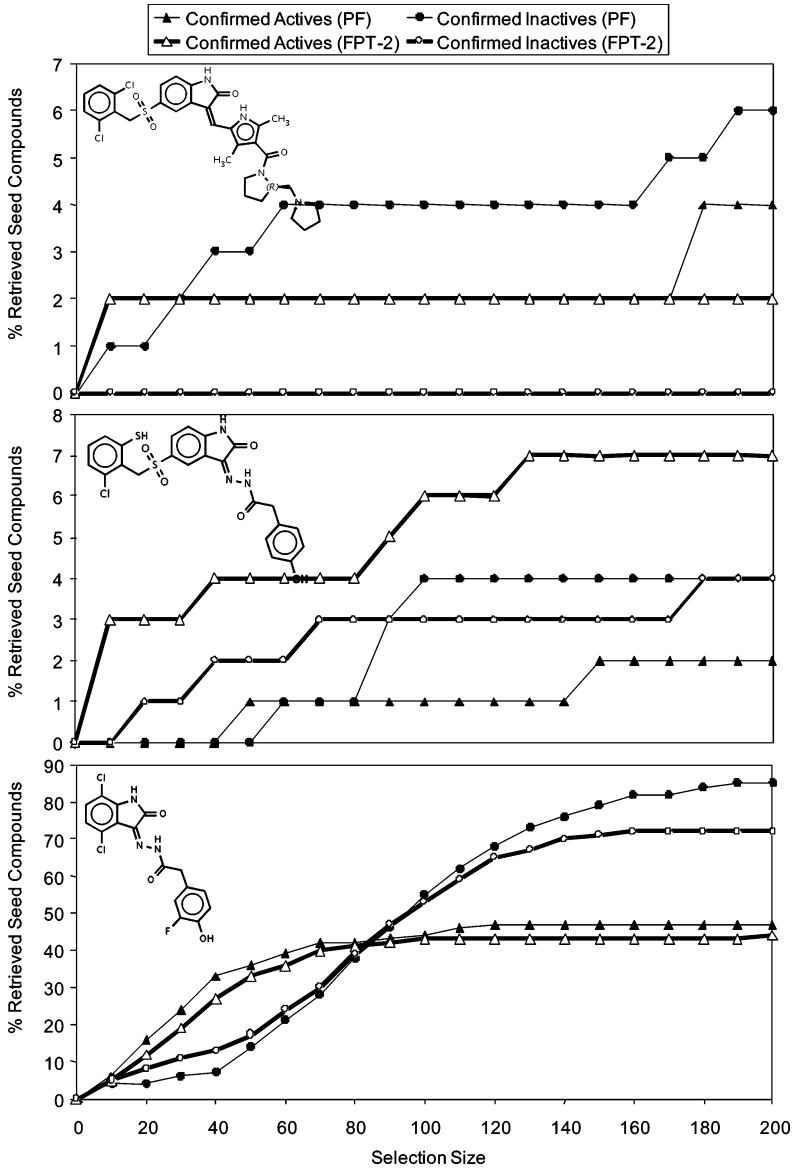

Figure 13. Results of virtual screening, probing each of the shown references against the MayBridge collection, seeded with compounds of known c-Met affinity (including actives with $pIC_{50} \geq 7$). Plots report the number of known actives and known inactives within subsets of nearest neighbors (subset size on the *x* axis) retrieved by the 2D-FPT (FPT-2 setup) and PF metrics, respectively.

similar to the lead and therefore eligible to be a virtual hit. Similarity to an active lead may be a sufficient but is clearly not a necessary condition. Unlike in virtual screening approaches based on QSAR or docking scores, successful similarity scoring is not expected to systematically score all of the actual active "ligands" better than the inactive "decoys"—if the set to be screened includes actives that are genuinely dissimilar to the reference, this subset of ligands might actually systematically score worse than decoys. The distributions of active ligands with respect to their similarity scores might actually be bi- or multimodal, complicating even more the statistical assessment of its robustness.[42] The selection criterion being the match of overall pharmacophore patterns—including those parts in which variability is not detrimental to binding—a search around a single lead may be too narrow.[43] In the present work, searches around single leads were performed with two different metrics (FPT and PF) and will be discussed in terms of relative retrieval rates.

- The key uncertainty in exploiting these results is the unknown activity status of the compounds from the bulk collection. The total number of actives present within the top neighbors is unknown, unless those compounds are ordered and tested against the target under study. Therefore, this study used both known actives and inactives for seeding. Selective enrichment in known actives, all while keeping the known inactives (often closely related analogues from the same series) out of the top neighbor set, is a strong indication of an increased probability to discover real actives among the hits from the bulk collection.

In the c-Met tyrosine kinase study case, the first two out of three lead compounds appear to be located at the rims of the cluster of the literature compounds of known activities. Both the PF and 2D-FPT-based metrics agree on the fact that the first lead (top plot in Figure 13) appears to have only two other known actives in its immediate neighborhood, with PF finding two more within the (arbitrary) limit of 200

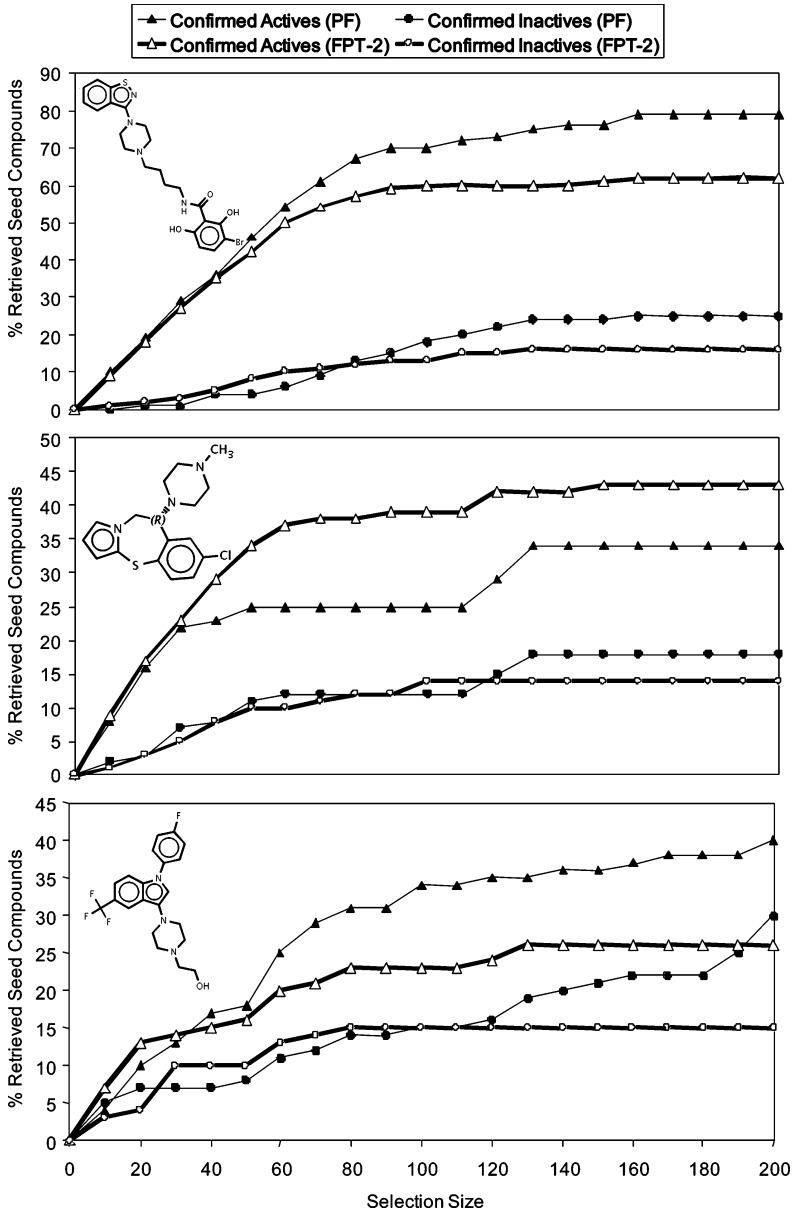

Figure 14. Virtual screening results for the D2 ligand study case (see legend of Figure 10 for details).

selected neighbors. However, the PF approach also co-opts four to six known inactives, which 2D-FPT successfully avoids. The results around the second lead compound are also clearly better with 2D-FPT, which recognizes roughly three times more known actives at basically equal numbers of co-opted inactives. The third c-Met lead appears, according to both metrics, to lay at the center of the c-Met compound cluster. Within the top 120 neighbors, retrieval levels closely match each other—with a slight advantage in favor of the PF approach, while at bigger selection sizes, the number of inactives co-opted by the PF significantly increases.

The study cases involving dopaminergic D2 compounds (Figure 14) showed that in all three situations lead molecules were well-surrounded by neighbors within the series. The first experiment may be considered a success of the PF approach—although it is still co-opting more inactives, it does better in known active retrieval by a clear margin. 2D-FPT clearly wins the second screening round, by simultaneously maximizing actives and minimizing co-opted inactives. The third experiment, eventually, is less clear-cut as the PF approach manages to retrieve more actives but only at the price of co-opting many more inactives than 2D-FPT.

Overall, the 2D-FPT-driven virtual screening appears to be more consistent—with respect to known actives and inactives—in the sense that higher active retrieval rates by PF are always accompanied by higher inactive retrieval rates as well. 2D-FPT systematically keeps the inactive retrieval rate equal or lower while nevertheless managing to improve the active retrieval rate in certain examples.

4. CONCLUSIONS

The insofar proven success of 2D-FPT-based similarity scoring compared to other fuzzy 2D and 3D pharmacophore descriptors is not surprising, as the three key innovations introduced here with respect to classical state-of-the-art descriptors and metrics are straightforward, chemically meaningful, and therefore expected to trigger improvements:

(1) The fuzzy mapping of molecular triplets on basis triplets is beneficial even in the context of topological distances (and assumed essential in a 3D context prone to conformational artifacts). It allows to accommodate the natural tolerance of receptors with respect to the number of bonds separating two binding groups and, from a practical point of view, allows a significant reduction of the descriptor dimension to a few thousands compared to >50 000 in binary fingerprints.

(2) The pK_a-dependent pharmacophore-type weighting scheme is able to correct many of the unavoidable inconsistencies that are introduced by rule-based flagging. Furthermore, local substituent swaps that, per se, would not translate to any significant pharmacophore pattern change as far as rule-based flagging is concerned may cause pK_a values to drift across the pH threshold and therefore trigger dramatic changes in the equilibrium population (and compound activity). Some of the "activity cliffs" in the structure−activity landscape of classical descriptor spaces are thus proven to be artifacts due to the failure of the latter to account for proteolytic equilibrium shifts. In the 2D-FPT space—for the first time, to our knowledge—this particular cause of landscape ruggedness has been successfully dealt with (insofar as the pK_a prediction tool is accurate, which appears to well be the case of the ChemAxon pK_a calculator employed in this work).

(3) The original similarity scoring scheme developed here recalls the simple truism that similarity due to the fact that a type is absent from both molecules is weaker than similarity due to the fact that both molecules contain the same type. As, in our hands, none of the classical scoring schemes managed to find the appropriate balance between contributions from shared, null, or exclusive triplets, such an optimal balance has been actively searched for—and found.

FPT as well as other pharmacophore-based descriptors have shown significant NB with respect to both diverse compound sets (BioPrint) and sets composed of several series of analogues. It is generally speaking much easier to demonstrate NB with respect to the latter situation, where simple discrimination between the main chemotypes at the basis of the various analogue series may suffice. The conclusions drawn on the basis of such studies may however be subject to different sources of bias due to relative size, chemical complexity, and other peculiarities of the considered analogue series. Mining for the underlying pharmacophore similarity in series with few representatives for each represented scaffold is much more challenging but successfully achieved by the FPT methodology. An interesting and recurring observation made in this work, requiring further investigation, is the possible correlation between the average pharmacophore complexity of the ligands of a target and its propensity for activity cliffs.

ACKNOWLEDGMENT

Special thanks to the ChemAxon (www.chemaxon.com) team, for allowing academics to freely use their software and for quick and effective hotline help. Sunset Molecular Inc. (http://sunsetmolecular.com/) and Tudor Oprea are acknowledged for providing the dopamine D2 data set. Nicole Dupont and Alexandre Barras (Institut de Biologie de Lille) are acknowledged for gathering the c-Met activity data from the literature. Thanks to Dr. Guy Lippens (University of Lille 1) for careful reading and important suggestions. ACCAMBA project members (http://accamba.imag.fr/) are acknowledged for encouraging this work.

APPENDIX A: THE ACTIVITY DISSIMILARITY SCORE

Similarity is an empirical concept, and there are no fundamental laws determining whether the activity profiles of two bioactive organic molecules are intrinsically similar or not. Like in the case of structural similarity, activity dissimilarity awaits for empirical definitions to be tried, validated, or discarded with respect to their usefulness in quantitative NB studies. Neighborhood behavior is necessarily a boot-strapping problem: its key assessment—that neighbors in a first (calculated) property space are likely to also be neighbors in a second (activity) property space—relies on two independent definitions of what "neighborhood" is supposed to mean in each one of the spaces.

For the above-mentioned reasons, this work postulates an activity dissimilarity score on the basis of plain medicinal chemistry common sense. Examples in which classical metrics (Euclidean, vector dot product, etc.) return counterintuitive dissimilarity measures will be discussed in order to highlight the need for a novel scoring scheme. Its implicit validation however comes from the fact that this definition of closeness in activity space respects the NB principle with respect to various molecular similarity metrics in structure space. In the following, the working hypotheses and parameters adopted in order to estimate the similarity of two activity profiles will be briefly outlined.

Profile similarity is determined by the behavior of a molecule pair (M,m) with respect to each target t. The target-specific response difference $\Delta_t(M,m)$ is defined as

$$\Delta_t(M,m) = \begin{cases} 0 & \text{if } |p_t(M) - p_t(m) \le 0.5 \\ 1 & \text{if } |p_t(M) - p_t(m) \ge 2.0 \\ \dfrac{|p_t(M) - p_t(m)| - 0.5}{1.5} & \text{otherwise} \end{cases} \quad \text{(A1)}$$

$\Delta_t(M,m)$ expresses a typical medicinal chemist's approach to activity comparison: two compounds with pIC_{50} values within 0.5 log units are said to have roughly the same activity; if however the pIC_{50} difference exceeds two log units, the molecules are beyond any doubt of different activity. In many situations, two log units is used as a landmark for selectivity: more than 2 orders of magnitude of affinity difference may not make any practical difference.

The activity index $\alpha_t(m)$ of a molecule m with respect to a target t is defined as a step function of the actual pIC_{50} value, such that compounds with affinities better than or equal to 1 μM count as active. A micromolar landmark for activity is widely used, especially in early stages of lead discovery.

$$\alpha_t(m) = \begin{cases} 0 \text{ if } p_t(m) < 6.0 \\ 1 \text{ otherwise} \end{cases} \quad \text{(A2)}$$

On the basis of definitions A1 and A2, $N_{diff}(m,M)$ and $f_{diff}(m,M)$—the index and respectivel fraction of significant differences in the profiles of molecules M and m are defined

as

$$N_{\text{diff}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} [\alpha_t(m) + \alpha_t(M) - 2\alpha_t(m)\,\alpha_t(M)]\,\Delta_t(m,M)$$

$$f_{\text{diff}}(m,M) = \frac{N_{\text{diff}}(m,M)}{N_{\text{targets}}} \quad \text{(A3)}$$

In the N_{diff} index, the first factor plays the role of logical exclusive or it equals 1 if and only if either $\alpha_t(m) = 1$ or $\alpha_t(M) = 1$. If so, N_{diff} is incremented by the amount of the target-specific response difference $\Delta_t(M,m)$: a pair (M,m) of approximately micromolar affinities on opposite sides of the 1 μM threshold will not contribute. Intuitively, N_{diff} is a fuzzy counter of the obvious activity differences in the profile.

The index and respective fraction of similarities $N_{\text{sim}}(m,M)$ and $f_{\text{sim}}(m,M)$ observed in the activity profiles of the two molecules are defined as

$$N_{\text{sim}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} \alpha_t(m)\,\alpha_t(M) \times [1 - \Delta_t(m,M)]$$

$$f_{\text{sim}}(m,M) = \frac{N_{\text{sim}}(m,M)}{N_{\text{targets}}} \quad \text{(A4)}$$

N_{sim} is the fuzzy counter of targets with respect to the two compounds having both strong $[\alpha_t(m) = \alpha_t(M) = 1]$ and similar $[\Delta_t(M,m) < 1]$ activities. Positive N_{sim} signals that the two compounds both interact with the same active site(s) and are therefore likely to include some common pharmacophore elements—insofar as most receptors tend to display a set of key interaction points that are always used in ligand binding, next to less important specific anchoring groups that form specific interactions with specific ligands. It is important to note that N_{diff} and N_{sim} do however not sum up to the total number N_{targets}. With respect to a pair of molecules, the set of targets making up the activity profile can be split into three domains: similarity, difference, and uncertainty, of sizes N_{sim}, N_{diff}, and $N_{\text{targets}} - N_{\text{diff}} - N_{\text{sim}}$, respectively. The uncertainty domain regroups targets for which molecules m and M display neither clear-cut different nor obviously similar behaviors. These include the (few) cases when compounds display significant potency differences despite both being active and the (ubiquitous) targets with respect to which m and M similarly fail to bind. A mutual lack of activity brings little information: molecules may be both inactive because of their similarity, or they may be each inactive in their own way.

The final activity dissimilarity score $\Lambda(m,M)$ associated with the activity profiles of molecules m and M is defined according to the following equation:

$$\Lambda(m,M) = \psi[f_{\text{diff}}(m,M) - \lambda \times f_{\text{sim}}(m,M)] \quad \text{(A5)}$$

with the conversion function $\psi(x)$ defined below:

$$\psi(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \le 0.05 \\ 0.1 + 18x & \text{if } 0 \le x < 0.05 \end{cases} \quad \text{(A6)}$$

In our opinion, this piecewise context-depending similarity scoring scheme returns a calculated profile activity score in agreement with medicinal chemistry and pharmaceutical know-how. Λ is a compromise between the sizes of the difference and similarity domains, with an empirical $\lambda = 5$ empirically chosen to emphasize the importance of observing actual similarities. The role of the conversion function $\psi(x)$ is to ensure the following:

- Only compound pairs sharing at least one significant (better than 1 μM) common hit in the profile may qualify to score top profile similarity (e.g., minimal $\Lambda = 0$), provided that the number of observed differences is low enough.
- If difference compensates for similarity, or if neither differences nor similarities could be evidenced (fully "uncertain" profiles, in the above-mentioned sense), a compromise score of 0.1 is returned. This value was chosen such as to signal that such profiles are clearly not different but should nevertheless not be allowed to compete in ranking with doubtlessly similar profiles at $\Lambda = 0$.
- Clearly different profiles, with $N_{\text{diff}} > \lambda N_{\text{sim}}$ score Λ values above 0.1, reach an upper limit of 1.0 if the excess differences make up more than 5% of the total number of targets in the profile.

It must be noted that Λ is not, strictly speaking, a metric: $\Lambda(M,M) = 0$ only if M binds at least to one target, with more than 1 μM of affinity. It is important to note that the conception of the Λ score ensures, unlike Euclidean or block distance metrics, a context-dependent activity difference interpretation. For example, the situation $p(m,t) = 5.0$ and $p(M,t) = 7.0$ marks an important difference between m and M, in the sense that selecting m from a database by means of a similarity screening experiment with respect to M might count as a failure. However, if $p(m,t) = 7.0$ and $p(M,t) = 9.0$, the discovery of m starting from M typically goes as a success, although the same 2 orders of magnitude of activity were lost. In the former case, target t contributes +1 to $N_{\text{diff}}(m,M)$, while in the latter, t contributes zero to both N_{diff} and N_{sim}. Eventually, if $p(m,t) > 7.0$ and $p(M,t) = 9.0$, target t becomes a contributor to N_{sim}. The Λ score therefore ranks a compound pair of activities (8,9) as more similar than a pair of activities (7,9) with respect to the target in question—like any Euclidean or Hamming score. Unlike these latter, however, Λ also meaningfully prioritizes the (7,9) pair over the (5,7) pair.

The failure of classical similarity metrics to respond differently to compound pairs that are both active and respectively both inactive often leads to an inappropriate, counterintuitive estimation of activity dissimilarity, as exemplified in Figure 15. The two bar plots represent comparative activity profiles—biological targets are aligned along the x axis, while the empty and filled bars respectively represent the pIC_{50} values of the compared molecules with respect to each target. Practically, IC_{50} values are only measurable starting from a certain activity threshold of the ligand—for compounds that are not active enough, a baseline pIC_{50} value of 3.0 is assumed (this also applies to BioPrint data). The left-hand graph displays a pair of molecules which have measurable pIC_{50} values with respect to a single target in the profile, and only one of them binds strongly enough to qualify as a potential hit or lead. A significant activity difference of three log units can be observed—obviously,

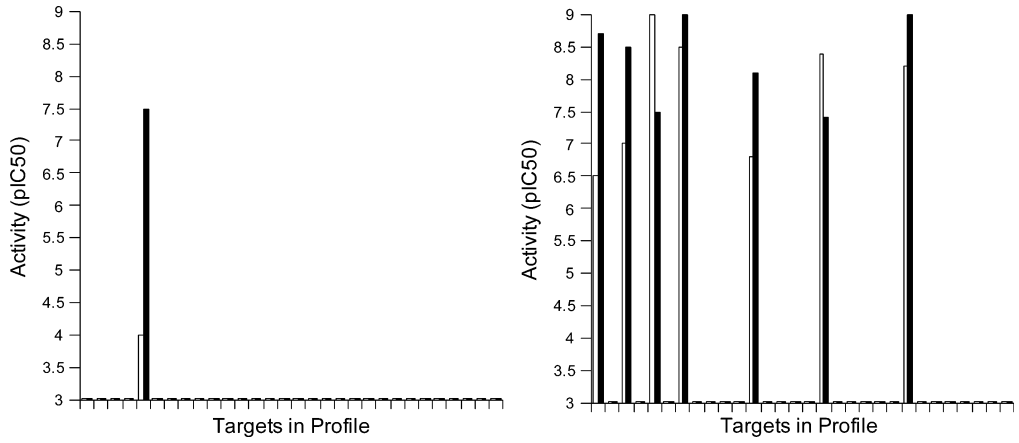

Figure 15. Two bar plots representing comparative activity profiles.

these molecules have different activity profiles. No other targets contribute to the Euclidean activity dissimilarity score, which therefore equals 3. The right-hand plot displays, by contrast, a pair of molecules with almost ideally covariant activities: they bind to the same targets, with comparable and significant—although not identical—affinities. However, every such target, rather than counting as a bonus in the profile similarity scoring, actually contributes some increment to the Euclidean profile dissimilarity score, which exceeds the dissimilarity level of the left-hand "different" compound pair and reaches 3.68. It is highly unlikely to expect identical activity values from binders to a same target, but it is guaranteed to get identical entries in the profile vector if none of the compounds have measurable pIC_{50} values—therefore, compound pairs with low hit rates in the profile will be spuriously favored by Euclidean scoring. A vector dot-product-based scoring metric would hardly perform better—as, in the left-hand plot, the only signals above the basis level stem from the same target; scores close to 1.0 (maximum similarity) are expected no matter what precise formula is used to calculate the profile correlation coefficient.

APPENDIX B: NEIGHBORHOOD BEHAVIOR CRITERIA.

NB analysis relies on monitoring activity dissimilarity within the subset $P(s)$ of molecule pairs (m,M) having calculated structural dissimilarity scores $\Sigma(M,m)$ below a variable dissimilarity threshold s. Let $N(s)$ represent the number of pairs retrieved by the selection $P(s)$ and which represent a fraction $f(s) = N(s)/N_{\text{all}}$ out of the total number of molecule pairs in the study. The consistency score $\chi(s)$ is defined in eq B1 by situating the average activity dissimilarity $\langle\Lambda(m,M)\rangle_{P(s)}$ of the $N(s)$ pairs in the actual selection at threshold s, in the context of (1) its upper baseline, the global average $\langle\Lambda(m,M)\rangle_{\text{all}}$ of all of the pairs in the study, which $\langle\Lambda(m,M)\rangle_{P(s)}$ approaches if selection at threshold s leads to a subset $P(s)$ as poor in activity-related pairs as a randomly picked one, and (2) its lower, ideal baseline, representing $\langle\Lambda(m,M)\rangle_{N(s)}^{\text{MIN}}$, the average Λ of the $N(s)$ compound pairs with the lowest Λ among the given N_{all} pairs.

$$\chi(s) = \frac{\langle\Lambda(m,M)\rangle_{\text{all}} - \langle\Lambda(m,M)\rangle_{P(s)}}{\langle\Lambda(m,M)\rangle_{\text{all}} - \langle\Lambda(m,M)\rangle_{N(s)}^{\text{MIN}}} \qquad \text{(B1)}$$

The overall optimality criterion $\Omega(s)$ renders a weighted account of two molecule pair counts in the actual selection of pairs $P(s)$ and randomly picked pairs:

- The first is the number of false similar pairs N_{FS} [structurally similar pairs with dissimilar activity profiles: $\Sigma(M,m) \le s$ and $\Lambda(M,m) > \kappa$]. A scaling factor $K > 1$ is applied to N_{FS} in order to take into account that, in virtual screening applied to drug discovery, the selection of pairs with diverging activity profiles is more penalizing than a failure to select all of the activity-related pairs (see below). In this work, $K = 100$.
- The second is the number of potentially false dissimilar pairs N_{PFD} [activity-related molecule pairs, apparently not structurally similar enough to be selected: $\Sigma(M,m) > s$ and $\Lambda(M,m) \le \kappa$].

The determination of N_{FS} and N_{PFD} requires in principle[16] a choice of the tolerated activity dissimilarity threshold κ—in the current context, however, every selected molecule pair (M,m) in $P(s)$ is fuzzily contributing an increment of $\Lambda(m,M)$ to N_{FS} and $1 - \Lambda(M,m)$ to N_{PFD}. In a random selection process, a set of size $N(s)$ would include activity-related and activity-unrelated pairs in a proportion equal to their overall occurrence in the total pair set and therefore

$$\Omega(s) = \frac{KN_{\text{FS}} + N_{\text{PFD}}}{KN_{\text{FS}}^{\text{rand}} + N_{\text{PFD}}^{\text{rand}}} = \frac{K\sum_{P(s)}\Lambda(M,m) + \sum_{All-P(s)}[1 - \Lambda(m,M)]}{K\frac{N(s)}{N_{\text{all}}}\sum_{\text{all}}\Lambda(m,M) + \left[1 - \frac{N(s)}{N_{\text{all}}}\right]\sum_{\text{all}}[1 - \Lambda(M,m)]} \qquad \text{(B2)}$$

NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. Low Ω at high χ signals good neighborhood behavior.

Supporting Information Available: The public data set complied from eight QSAR series, including calculated FPT descriptors (FPT-2) and the .xml setup files controlling compound standardization and generation of ChemAxon PF and CF descriptors. This material is available free of charge via the Internet at http://pubs.acs.org. Activity dissimilarity $\Lambda(M,m)$ and FPT dissimilarity scores $\Sigma^{\text{FTP}}(M,m)$—not shared via

pubs.acs.org for technical reasons (files too large)—are available upon request (dragos.horvath@univ-lille1.fr).

REFERENCES AND NOTES

(1) Adam, M. Integrating Research and Development: The Emergence of Rational Drug Design in the Pharmaceutical Industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(2) Geney, R.; Sun, L.; Pera, P.; Bernacki, R. J.; Xia, S.; Horwitz, S. B.; Simmerling, C. L.; Ojima, I. Use of the Tubulin Bound Paclitaxel Conformation for Structure-Based Rational Drug Design. *Chem. Biol.* **2005**, *12*, 339−48.

(3) Ivanov, A. A.; Baskin, I. I.; Palyulin, V. A.; Piccagli, L.; Baraldi, P. G.; Zefirov, N. S. Molecular Modeling and Molecular Dynamics Simulation of the Human A2B Adenosine Receptor. The Study of the Possible Binding Modes of the A2B Receptor Antagonists. *J. Med. Chem.* **2005**, *48*, 6813−20.

(4) Bernacki, K.; Kalyanaraman, C.; Jacobson, M. P. Virtual Ligand Screening against Escherichia coli Dihydrofolate Reductase: Improving Docking Enrichment Using Physics-Based Methods. *J. Biomol. Screening* **2005**, *10*, 675−81.

(5) Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A. M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. Pharmacophore-Based Design of HIV-1 Integrase Strand-Transfer Inhibitors. *J. Med. Chem.* **2005**, *48*, 7084−8.

(6) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.

(7) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold Hopping with Molecular Field Points: Identification of a Cholecystokinin-2 (CCK2) Receptor Pharmacophore and Its Use in the Design of a Prototypical Series of Pyrrole- and Imidazole-Based CCK2 Antagonists. *J. Med. Chem.* **2005**, *48*, 6790−802.

(8) Güner, O. F. *Pharmacophore Perception, Use and Development in Drug Design*; International University Line: La Jolla, CA, 2000.

(9) Horvath, D. High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and its Role in the Drug Discovery Laboratory. In *Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 429−472.

(10) Makara, M. G. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(12) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of Validated QSAR Models of d(1) Dopaminergic Antagonists for Database Mining. *J. Med. Chem.* **2005**, *48*, 7322−32.

(13) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-Protein-Coupled Receptor Affinity Prediction Based on the Use of a Profiling Dataset: QSAR Design, Synthesis, and Experimental Validation. *J. Med. Chem.* **2005**, *48*, 6563−74.

(14) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004.

(15) For details on the two-point topological pharmacophore descriptors developed by ChemAxon, see http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Sept 2006).

(16) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with respect to In Vitro Activity Spaces − A Benchmark for Neighborhood Behavior Assessment of Different in Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691−698.

(17) Horvath, D.; Mao, B. Neighborhood Behavior − Fuzzy Molecular Descriptors and their Influence on the Relationship between Structural Similarity and Property Similarity. *QSAR Comb. Sci.* **2003**, *22*, 498−509; special issue "Machine Learning Methods in QSAR Modeling".

(18) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−23.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1995**, *38*, 144−150.

(20) Menard, J. P.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−13.

(21) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865−71.

(22) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces − A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.

(23) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470−80.

(24) http://www.cerep.fr/cerep/users/pages/Collaborations/Bioprint.asp (accessed Sept 2006).

(25) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(26) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(27) The above-mentioned data sets are also available via http://www.cheminformatics.org/ (accessed Sept 2006).

(28) Horvath, D. ComPharm − Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design. In *QSPR/QSAR Studies by Molecular Descriptors*; Diudea, M., Ed.; Nova Science Publishers: New York, 2001; pp 395−439.

(29) http://www.chemaxon.com/jchem/doc/api/ (accessed Sept 2006).

(30) http://www.chemaxon.com/jchem/index.html?content=doc/user/Standardizer.html (accessed Sept 2006).

(31) http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (accessed Sept 2006).

(32) http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Sept 2006).

(33) http://www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed Sept 2006).

(34) http://www.maybridge.com/ (accessed Sept 2006).

(35) Christensena, J. G.; Burrows, J.; Salgiab R. c-Met as a Target for Human Cancer and Characterization of Inhibitors for Therapeutic Intervention. *Cancer Lett.* **2005**, *225*, 1−26.

(36) Vojkovsky, T.; Koenig, M.; Zhang, F.-J.; Cui, J. Tetracyclic Compounds as c-Met inhibitors. Patent WO2005004808, 2005.

(37) Koenig, M. Indolinonehydrazides as c-Met Inhibitors. Patent WO200500-5378, 2005.

(38) Compounds and activity data taken from the WOMBAT database of Sunset Molecular, Inc. (http://sunsetmolecular.com/products/?id=4) courtesy of Tudor I. Oprea, 2005.

(39) Altschul, S. F. Amino Acid Substitution Matrices from an Information Theoretic Perspective. *J. Mol. Biol.* **1991**, *219*, 555−65.

(40) Kubiny, H. Structure-Based Design of Enzyme Inhibitors and Receptor Ligands. Second European Workshop in Drug Design, Certosa di Pontignano, May 17−24, 1998; oral presentation.

(41) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−63.

(42) Seifert, M. H. J. Assessing the Discriminatory Power of Scoring Functions for Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.

(43) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049−54.

CI6002416

Cinquième partie

La stratégie stochastique par rapport à la stratégie point par point (*stepwise*) dans la recherche de Relations Structure-Activité

L'une des phases d'une étude de Relation Structure-Activité (*QSAR*) consiste à sélectionner les descripteurs à utiliser. Généralement, un ensemble d'hypothèses de travail visent à ramener l'ensemble de descripteurs de départ, dont la dimensionalité peut être très grande, à une taille gérable.

Habituellement, la régression point par point, ou *stepwise regression* est utilisée, malgré ses défauts potentiels, à cause de son coût en temps machine assez bas. La mise en place et le succès des 2D-FPTs, des descripteurs de haute dimension, représente une occasion de développer un nouvel outil puissant de sélection.

Dans ce chapitre, nous allons explorer une autre approche qui pourrait nous permettre de sélectionner les descripteurs pertinents à utiliser dans des études de Relation Structure-Activité. L'échantillonneur *QSAR* que nous proposons est basé sur un algorithme génétique original, *SQS*, qui favorise la recherche de modèles non biaisés par rapport au gain en temps de calcul. Cet algorithme est indépendant du filtrage *a priori* des descripteurs et n'est pas limité aux modèles linéaires.

Le *SQS* a été testé sous différentes conditions et comparé à l'outil de régression point par point d'ISIDA *SR*. Les données utilisées pour les tests (trois ensembles de composés anti-VIH) ont été décrites par trois types de descripteurs moléculaires différents : les descripteurs ISIDA (des fragments sous-structuraux incluant des séquences d'atomes et de liaisons ainsi que les atomes avec leur environnement proche), les descripteurs CAX (des empreintes pharmacophoriques à deux points associés à d'autres descripteurs variés proposés par Chemaxon), et les 2D-FPTs présentés dans le chapitre précédent. La comparaison entre le *SQS* et le *SR* a été faite en variant la manière de découper l'ensemble de données en sets d'entrainement et de validation, la sélection des descripteurs et la politique de non-linéarité.

Le *SQS* est une approche stochastique, ce qui implique qu'une relance du programme sur un même jeu de données n'amènera pas forcément à la découverte des mêmes équations. Néanmoins, même si les équations trouvées peuvent être différentes, le niveau de robustesse est constant. De plus, si le *SQS* est autorisé à appliquer des transformations non-linéaires au descripteurs, l'espace du problème augmente. Les modèles non-linéaires tendent cependant à être plus robustes pour la validation.

Il est montré que le *SQS* donne d'aussi bonnes performances que le *SR* lors de l'évaluation du modèle de validation. Les simplifications du *SR* ont parfois des conséquences négatives, ce qui entraine un dépassement de ses performances par le *SQS* dans ces cas précis. De plus, les modèles consensus provenant de larges modèles *SQS* se valident bien, mais pas beaucoup mieux que les équations consensus de *SR*.

Il ressort donc de ces travaux que le *SQS* est un outil de construction de *QSAR* robuste d'après les tests de validation standards sur des ensembles de composés externes (provenant des mêmes familles que celles utilisées pour l'entrainement). Interpréter les résultats du *SQS* reste cependant un défi par rapport à la manière traditionnelle d'interprétation des résultat du *QSAR* et ses apports et inconvénients ne sont sans doute pas tous révélés par ces tests. Le *SQS* génère des milliers de modèles validés, chacun apportant des domaines d'applicabilité ainsi que des valeurs prédites potentiellement divergentes pour les composés externes. Comment gèrer tous ces modèles ? Par rapport au *SR*, qui n'impose pas autant de choix à l'utilisateur, mais qui parie sur une équation pouvant potentiellement se comporter correctement lors d'un criblage virtuel, quelle est la meilleure stratégie ?

Sixième partie

Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generations. How Much Effort May the Mining for Successful QSAR Models Take ?

Stochastic versus Stepwise Strategies for Quantitative Structure−Activity Relationship Generation—How Much Effort May the Mining for Successful QSAR Models Take?†

Dragos Horvath,*,‡ Fanny Bonachera,‡ Vitaly Solov'ev,§,|| Cédric Gaudin,§,⊥ and Alexander Varnek§

UGSF-UMR 8576 CNRS/USTL, Université de Lille 1, Bât C9., 59650 Villeneuve d'Ascq, France, Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France, Technologies Servier, 25-27 rue E. Vignat, 45000 Orléans, France, and Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a, 119991 Moscow, Russia

Received October 31, 2006

Descriptor selection in QSAR typically relies on a set of upfront working hypotheses in order to boil down the initial descriptor set to a tractable size. Stepwise regression, computationally cheap and therefore widely used in spite of its potential caveats, is most aggressive in reducing the effectively explored problem space by adopting a greedy variable pick strategy. This work explores an antipodal approach, incarnated by an original Genetic Algorithm (GA)-based Stochastic QSAR Sampler (SQS) that favors unbiased model search over computational cost. Independent of a priori descriptor filtering and, most important, not limited to linear models only, it was benchmarked against the ISIDA Stepwise Regression (SR) tool. SQS was run under various premises, varying the training/validation set splitting scheme, the nonlinearity policy, and the used descriptors. With the considered three anti-HIV compound sets, repeated SQS runs generate sometimes poorly overlapping but nevertheless equally well validating model sets. Enabling SQS to apply nonlinear descriptor transformations increases the problem space: nevertheless, nonlinear models tend to be more robust validators. Model validation benchmarking showed SQS to match the performance of SR or outperform it in cases when the upfront simplifications of SR "backfire", even though the robust SR got trapped in local minima only once in six cases. Consensus models from large SQS model sets validate well—but not outstandingly better than SR consensus equations. SQS is thus a robust QSAR building tool according to standard validation tests against external sets of compounds (of same families as used for training), but many of its benefits/drawbacks may yet not be revealed by such tests. SQS results are a challenge to the traditional way to interpret and exploit QSAR: how to deal with thousands of well validating models, nonetheless providing potentially diverging applicability ranges and predicted values for external compounds. SR does not impose such burden on the user, but is "betting" on a single equation or a narrow consensus model to behave properly in virtual screening a sound strategy? By posing these questions, this article will hopefully act as an incentive for the long-haul studies needed to get them answered.

1. INTRODUCTION

Quantitative Structure−Activity Relationships[1−3] (QSAR) are empirical mathematical models (equations) returning an estimate of the activity level of a given molecule as a function of descriptors. Such models are obtained by calibration against a training set (TS) opposing activity values A_m of already tested molecules m (the explained variable) to their descriptor values (the explaining variables $D^1_m, D^2_m, ..., D^N_m$ where each element i of the vector stands for a certain structural or physicochemical aspect. Let N be the total number of available descriptors). Various data mining procedures may be used to complete the three main calibration steps:

I. Select a minimal number of explaining variables for actual use in model building. In the following, let n be the number of selected descriptors. Formally, the N-dimensional binary phase space associated with descriptor selection has each axis i associated with a binary variable $\delta_i = 0$ if the ith descriptor is be ignored and $\delta_i = 1$ if it enters the model.

II. Choose a functional form for the equation expected to optimally estimate activities as a function of the above-selected descriptors $Y_m = f(D^i_m;c^k)$—the simplest choice being a linear expression $Y_m = \sum_{\delta i>0}^{i=1,N} c^i D_m{}^i + c^0$.

III. Fit the coefficients c^k, $k = 1...n$, to obtain an equation minimizing the residual sum of squares $\Sigma(Y_m - A_m)^2$ between measured and predicted activities. If Y_m is a linear expression, c^k may be found by linear regression.

From a point of view of problem complexity, the descriptor selection step (I.) theoretically requires the exhaustive exploration of 2^N possible schemes. The function selection step (II.) offers virtually endless possibilities once nonlinear expressions are envisaged. Step (III.) may be quite time-consuming even with linear models: regression requires a $(n+1) \times (n+1)$ matrix inversion step scaling as $O(n^3)$ in terms of computational effort. Furthermore, successful completion of steps I.−III. is necessary but not sufficient: cross-validation[4,5] and external testing procedures have to be integrated into model buildup. However, typically,

† Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

* Corresponding author phone +33.320.43.49.97; e-mail: dragos.horvath@univ-lille1.fr.

‡ Université de Lille 1.

§ Université Louis Pasteur.

|| Technologies Servier.

⊥ Russian Academy of Sciences.

10.1021/ci600476r CCC: $37.00
Published on Web 05/05/2007

relatively little time and computer effort is invested in QSAR buildup, for two main reasons:

(1) Traditionally, the QSAR building problem is declared "solved" once acceptable models were found—the exploration of the entire phase space of the problem is not, at this time, seen as a goal per se—although this may be arguably wrong.

(2) Problem space pruning strategies were successfully developed. These emerged under the pressure of limited computer power in the early years of QSAR, and were not systematically challenged when RAM capacity and CPU time ceased to represent bottlenecks. Descriptors are routinely discarded for being "intercorrelated" although this correlation may (a) not extend beyond the training set or (b) hide an independent variable that happens to be a small difference of two large terms. Descriptor selection is routinely performed on hand of linear models, with selected variables being subsequently used as input for neural networks. They are not necessarily the best choice for nonlinear models, but the high cost of simultaneous descriptor selection coupled to nonlinear modeling justifies the "short-cut" as far as some valid equation is being obtained.

The goal of this publication is to explore whether a more computationally intensive (days on a typical dual processor PC) approach to QSAR buildup, involving distributed computation, may provide an in-depth exploration of the problem space and provide a deeper understanding of QSAR methodology. In particular, it may allow to situate typical equations from stepwise procedures in this broader context.

The recent development of high-dimensional (N=10^3...10^4) Fuzzy Pharmacophore Triplet (FPT) descriptors[6] represented an additional incentive to develop a powerful selection tool. Furthermore, we extended the selection phase space to include additional states encoding the choice of nonlinear functional forms. To this purpose, the degrees of freedom were allowed to take values beyond the two binary options $\delta_i = 0/1$. While $\delta_i = 0$ maintains its original meaning "ignore descriptor i", $\delta_i = 1,2,...,N_T$ selects one of the predefined N_T nonlinear transformation functions and enters the transformed descriptor i, according to rule $T_{\delta i}(D^i)$, in the model. The final functional form of the QSAR model will thus be

$$Y_m = c^0 + \sum_{i=1,N}^{\delta i>0} c^i T_{\delta(i)}(D_m^i) \qquad (1)$$

e.g. a linear combination of nonlinearly transformed descriptors, which is equivalent to a single-layer neural network. This does not cover all the possible nonlinear expressions but benefits from the relative facility of fitting the c^i by linear regression while nevertheless allowing for nonlinear treatment (the coefficients appearing within the transformation functions $T(D)$ are constant; they may be partially optimized by entering a same functional form with different coefficient values as independent choices). Any rules $T(D)$ may in principle be envisaged—including $T_1(D) = D$ and $T_2(D) = D^2$ (the linear and square functions being now particular "nonlinear" transformations)—at the cost of an "exploding" phase space volume of $(N_T+1)^N$.

In response to this challenge, the Stochastic QSAR Sampler (SQS) has been developed, based on a distributed, hybrid genetic algorithm-based descriptor selection procedure, the Model Builder (MB). SQS is inspired from an analogous conformational sampling tool[7] designed to handle folding problems of miniproteins and docking problems with full site side-chain flexibility. The evolutionist approach central to the descriptor and functional form selection strategies has been hybridized with alternative optimization techniques and driven by a meta-optimization loop choosing the MB control parameters.

Three different data sets of molecules of known anti-HIV activities have been used for benchmarking the potential benefits of the more aggressive SQS strategy. They are relatively small (~100 molecules each) and are known to permit successful stepwise QSAR model buildup. For each set, five different splitting schemes into Training (TS_k) and Validation (VS_k) Sets, k = 1...5, were considered such that to ensure that every compound in a series is once being kept for validation. This should avoid any potential bias due to any peculiar choice of validation molecules. Five independent molecular descriptor sets were used in this work, including Fuzzy Pharmacophore Triplets (FPT), ISIDA[8] fragment descriptors, and ChemAxon[9] (CAX) descriptors (including two-point pharmacophore fingerprint, BCUT descriptors etc.). Among the 75 combinations of 3 data sets × 5 splitting schemes × 5 molecular descriptor choices, the following QSAR build-up simulations were performed (and duplicated, in order to assess reproducibility): (a) stochastic searches for linear (N_T=1) and polynomial (including the squared transformation, N_T=2) models based on FPT, ISIDA, and ChemAxon descriptors, (b) stochastic searches of fully nonlinear FPT-based models, and (c) deterministic model buildup of linear QSARs using Stepwise Regression (SR), with FPT and ISIDA descriptors

The stochastic procedure typically generates ~10^5 models, out of which diverse representatives are selected among the best cross-validating ones. These are then systematically confronted to their respective validation sets in order to obtain validation correlation coefficients R^2_V. Out of the selected models, not all validate successfully, and the ones that do so cannot be known a priori.[10] Successful validation of SQS representative models will thus be treated like a probabilistic event. Density distribution histograms monitoring the probability to discover a model with R^2_V within a given range $r \le R^2_V < r + \epsilon$ will be traced and compared for various simulations—using different sets of descriptors, different approaches to nonlinearity, etc. Average validation scores $<R^2_V>$ taken over all selected models can thus serve as success criterion related to the specific setup(s) of the considered simulation(s).

The first subject of this paper is the study of the intrinsic behavior of SQS:

1. Reproducibility: to what extent is the overall performance of the final models subject to fluctuations? Will models be found again when repeating a stochastic search?

2. SQS response with respect to a phase space volume increase: How does the introduction of nonlinearity impact model validation propensity? Is SQS sensitive to the size of the initial descriptor pool?

Next, linear SQS models will be compared to equations obtained by a deterministic QSAR build-up procedure: the Stepwise Regression tool (SR) of the ISIDA package.

3. Comparing individual SQS and SR models: how many individual models with R^2_V above a given threshold were

Table 1. Considered Transformation Functions[a]

code	function	remark
0	$T_0(D) = 0$	null function (ignore D)
1	$T_1(D) = D$	identity function
2	$T_2(D) = D^2$	squared descriptor
3	$T_3(D) = \exp\{-[(D - \langle D\rangle)/\sigma(D)]^2\}$	broad Gaussian (zexp)
4	$T_4(D) = \exp\{-3[(D - \langle D\rangle)/\sigma(D)]^2\}$	sharp Gaussian (zexp3)
5	$T_5(D) = \{1 + \exp[(D - \langle D\rangle)/\sigma(D)]\}^{-1}$	flat sigmoid (zsig)
6	$T_6(D) = \{1 + \exp[3(D - \langle D\rangle)/\sigma(D)]\}^{-1}$	steep sigmoid (zsig3)

[a] Nonlinear functions 3−6 require the input of the expectation values (averages $<D>$) and variances $\sigma(D)$ for the concerned descriptors calculated on hand of 2200 currently marketed drugs and reference compounds.

produced by every approach—and what fraction do these make out of the respective sets of models? Are there study cases in which SR fails to discover well validating models, while SQS succeeds? Reversely, can the fluctuation-prone SQS be seen to miss valid models picked up by the deterministic approach?

4. Comparing SQS and SR consensus models[11] (CM), reportedly a safer choice, in terms of validation propensity, than individual QSAR equations: does this still apply with extensive averaging of thousands of sampled parent equations?

This work is structured as follows: in the Methods section, a description of SQS and SR procedures precedes an introduction of the different descriptors and compound sets followed by a presentation of the statistical tools utilized to tackle the four above-mentioned key questions. The Results section will sequentially address these questions and lead to the Conclusion paragraph.

2. METHODS

2.1. The Stochastic QSAR Sampler (SQS). SQS is aimed to provide an effective, combined descriptor and nonlinear function selection procedure for the fitting of QSAR models according to eq 1. Table 1 enumerates the $N_T = 6$ currently implemented predefined transformation functions. The use of sigmoids and Gaussians requires the input of nominal expectation values $\langle D_i\rangle$ and variances $\sigma(D_i)$ for each descriptor, which in the present work were sought to be representative of the "universe" of drugs and druglike molecules. $\langle D_i\rangle$ and $\sigma(D_i)$ were calculated on hand of an independent set of 2200 drugs and druglike molecules and are constants throughout this work, e.g., independent of the processed training/validation sets.

2.1.1. Chromosome Definition and Prefiltering of Allowed Transformations. The Genetic Algorithm (GA)-driven Model Builder (MB) features chromosomes of size N, where every locus i codes $\delta(i)$—a natural number standing for the function $T_{\delta(i)}$ to be applied to D_i. Setting an upper threshold for acceptable δ values limits the allowed transformations to 1 (linear), 2 (polynomial, involving descriptors and/or their squares), or 6 (fully nonlinear regime). At input of training set molecules, the appropriateness of using a function T_k in conjunction with D_i is evaluated: the minimal variance of the vector of transformed descriptor values $T_k(D_i^m)$ over selected subsets including 2/3 of the training set molecules is required to exceed a user-defined threshold. If this test fails, then k will be deleted from the list of integer values chromosome locus i is allowed to adopt—the opposite would likely have caused a cross-validation failure anyway. Also, if the transformed descriptor is too strongly correlated with an already accepted transformation p—e.g. $T_k(D_i^m) \approx aT_p(D_i^m) + b$ for all m—then k will be rejected. For example, consider binary descriptor $D_i = 0/1$. A Gaussian transformation $T_3(D) = \exp[-(D-0.5)^2]$ makes no sense, since $T_3(0) = T_3(1)$. Also, using $T_1(D) = D$ is sufficient—replacing 0 by $T(0)$ and 1 by $T(1) \neq T(0)$ will have no other effect but adjusting of the associated linear coefficient.

No attempt is made at this stage to discard descriptor columns D_i that are (themselves or via their transformations) correlated to other potential descriptors.

2.1.2. Hybrid Genetic Algorithm. The following is a brief description of the Model Builder (MB). Its tunable operational and control parameters are given in capital italics.

Parallelization. A tunable number of NCONT parallel GA processes are started simultaneously on different CPUs, in order to simulate distinct "continents" in which Darwinian evolution may explore diverging paths. If a new fittest chromosome is found by a process, it is allowed to "migrate" to another of the parallel runs.

Population Initialization. After having built, for each locus i of the chromosome, the list of allowed functions $\delta(i)$ to be used in conjunction with D_i, an initial population of NPOP random chromosomes is used as departure point of Darwinian evolution. Due to the obvious interest in models with a minimal number of descriptors, an explicit upper threshold for the allowed number of variables, $\mu = 5$, is set at this point. μ is an adaptive parameter, gradually incremented during the evolution process (see later on). Two possible chromosome initialization schemes are alternatively considered: (a) Random pick: first, a random value for the effective number n of selected descriptors is drawn between 1 and μ. All the loci of the new chromosome are set to 0, then a number of n loci between 1 and N are randomly picked, and for each such locus i one of the available non-null options for $\delta(i)$ is randomly chosen. (b) Random ancestor crossovers: if a set of fit chromosomes was provided by previous runs, new ones may be produced by random crossovers of these "ancestors", while ensuring that (b1) the result is original (not among the ancestors) and (b2) it has less than μ selected descriptors.

A tunable probability PANCEST to apply ancestor crossover instead of random picking controls the stochastically alternating use of above-mentioned options.

Fitness Estimation. The fitness (see Figure 1) of chromosomes with more than μ selected descriptors is set to an arbitrary low level. Otherwise, it is calculated as follows: the matrix of the transformed descriptor values (n columns) is submitted to leave-a-third-out cross-validated MLR, returning a cross-validated correlation coefficient Q^2 and the predicted activity Y_m^{XV} for each molecule in the training set. Fitness is related to Q^2 amended by a model size-dependent penalty, with a tunable VARPENALTY parameter and $\bar{A}_{TS}$ being the average TS compound activity.

$$\mathrm{Fit}(C) = Q^2(C) - n \times \mathrm{VARPENALTY}$$
$$Q^2(C) = 1 - \frac{\sum_{m\in \mathrm{TS}} [Y_m^{XV}(C) - A_m]^2}{\sum_{m\in \mathrm{TS}} [\bar{A}_{\mathrm{TS}} - A_m]^2} \qquad (2)$$

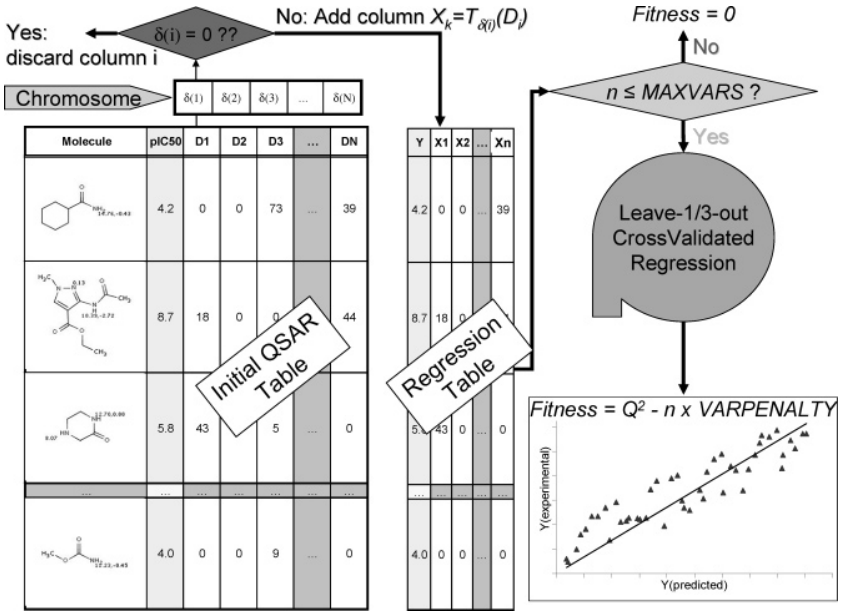

Figure 1. Evaluation of the fitness of a chromosome encoding a general nonlinear QSAR model. The MAXVARS control parameter is referred to as "μ" in the text.

Reproduction. Given a current generation of NPOP chromosomes, Darwinian evolution is simulated by generating a buffer population of offspring from crossovers and mutations. Parent chromosomes are randomly paired. Crossovers are performed at a randomly chosen positions, while checking that offspring differs from parents. Mutations, occurring at a tunable rate MUTFRQ, are random changes of the content of randomly picked loci i: $\delta(i) > 0$ mutate to 0, while $\delta(i) = 0$ will be toggled to one of the allowed transformer function codes.

Selection. The algorithm chooses either the *global* or the *intrafamily* scheme to pick the NPOP chromosomes making up the next generation out of the extended population emerging from reproduction.

In the *global* selection scheme, all the chromosomes of the extended population are sorted by decreasing fitness and then subjected to diversity filtering. The similarity of two chromosomes C and c is defined by the error pattern correlation score $S(C,c)$, where $Err_m(C) = Y_m(C) - A_m$ is the prediction error of the activity A calculated according to the model coded by chromosome C, for the molecule m:

$$S(C,c) = \frac{\sum_{m\in \mathrm{TS}} \mathrm{Err}_m(C)\mathrm{Err}_m(c)}{\sqrt{\sum_{m\in \mathrm{TS}} \mathrm{Err}_m{}^2(C) \times \sum_{m\in \mathrm{TS}} \mathrm{Err}_m{}^2(c)}} \qquad (3)$$

If $S(C,c) >$ MAXSIM, the less fit of C and c will be discarded. This selection strategy is favoring convergence and is therefore applied only once every NGLOBAL generations.

If the *intrafamily* selection scheme is chosen, then offspring issued from crossovers will compete against its parents only: if the fittest child is fitter than the best of parents, then both parent chromosomes are replaced by the children. Mutants may only replace the "wild type" if they are fitter than the latter. This selection scheme favors population diversity and is therefore the default procedure.

The fittest NPOP chromosomes passing the similarity filter will form the next generation. If less than NPOP passed, random ones will be added to restore nominal population size.

Deterministic (*Lamarckian) Chromosome Optimization.* Occasional use of "Lamarckian" approaches (back-copying into the chromosome the knowledge about locally fitter configurations, obtained by exploring the neighborhood of a solution) has been shown[7] to enhance Darwinian evolution. The herein used local optimizer tool alternatively discards each of the n selected descriptors and reassesses the fitness of all the models of size $n-1$. If any of the latter is found to be fitter than the parent, it will take its place.

Population Refreshment Strategies. Failure to find a new fittest chromosome during NWAIT successive generations triggers a reinitialization ("apocalypse") of all but the fittest chromosome (elitist strategy). At this point, μ is incremented by 5 unless it already had reached its (tunable) maximum. Larger models are thus gradually being allowed for, as soon as no more progress can be obtained with fewer variables.

MB Termination. Eventually, if a series of NAPOC successive apocalypses did not lead to fitter solutions, a run stops.

Triplicate Runs and MB Success Score (*Meta-Fitness).* The SQS meta-optimization loop (Figure 2) pilots the search for the most appropriate MB operational parameters. The success of the triplicate MB run (3 successive runs with same operational parameters) defined by the quality of the solution set L produced. A large L, richer in fit models, means that the current choice of operational parameters has been judicious. The meta-fitness score, a measure of sampling success, is defined as

$$\mu\mathrm{Fitness} = \log \sum_{c\in L} \exp[-\beta \times \mathrm{Fit}(c)] \qquad (4)$$

μFitness is an implicit function of the operational parameters that leads to sampling of the local solution set L. The temperature factor was empirically set to $\beta = 30$.

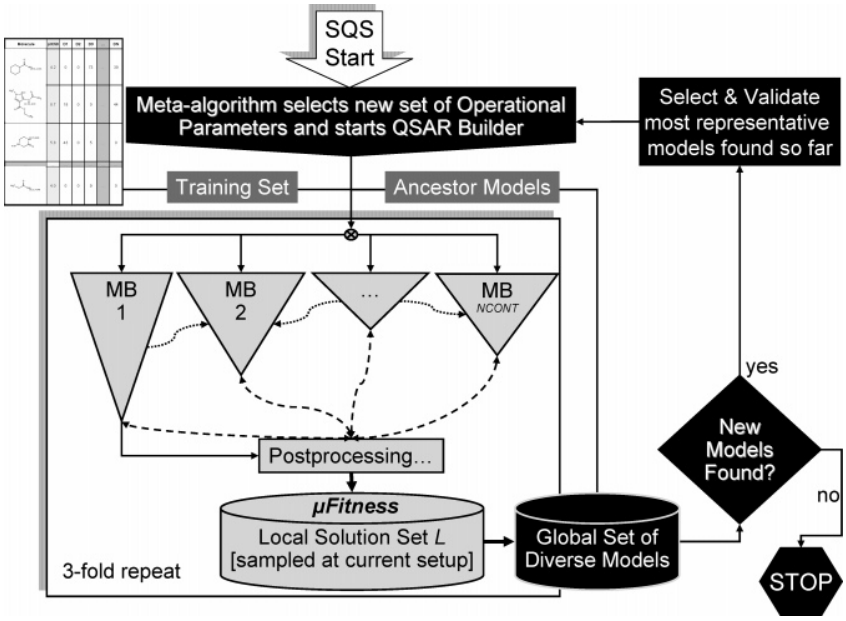

Figure 2. Overview of the stochastic QSAR build-up procedure, featuring the distributed hybrid GA-based QSAR builder, driven by a meta-optimization loop in search of the most appropriate operational parameter setup.

SQS Flowchart. After completion of a triplicate MB run, the local solution set L is merged with the global database of visited models, which is sorted by fitness. Redundant models at MAXSIM > 0.9 are rejected. Meta-optimization continues, according to a basic GA scheme, as long as the latest triplicate runs continued to add new valuable solutions to the global database.

2.1.3. SQS Run Validation and Postprocessing. After each triplicate run, a current set of most representative models is extracted from the up-to-date global set: chromosomes with Q^2 within 0.2 units of the current Q^2 maximum are classified with respect to the number of selected descriptors n, and, for each of these size classes, the top 10 fittest representatives are selected. Each selected model C is subjected to external validation in order to calculate their validation correlation score, on hand of the predicted activity values $Y_m(C)$ for all the m of the validation set, where $\bar{A}_{VS}$ represents the average activity of VS compounds (for models with prediction errors exceeding the ones of the "null" model $Y_m = \bar{A}_{VS}$, negative R^2_V values are truncated to 0):

$$R^2_V(C) = \max\left\{0,1 - \frac{\sum_{m\in VS}[Y_m(C) - A_m]^2}{\sum_{m\in VS}[A_m - \bar{A}_{VS}]^2}\right\} \qquad (5)$$

R^2_V truncation is required because very low negative values which, per se, have no quantitative significance beyond the one of a flag for validation failure will skew the average $<R^2_V>$ values used to compare the relative proficiency of various sampling strategies (see further on). For example, a QSAR method leading to 4 models of $R^2_V = 0.9$ and one of $R^2_V = -5.0$ is obviously to be preferred to one producing 5 models of $R^2_V = 0.0$, although averages of untruncated R^2_V suggest the contrary.

The set of representative models extracted after each triplicate MB run is steadily evolving, as newly visited configurations are added to the global pool of found models. All the models selected as "representative" at any instance of the meta-optimization loop will be considered in the final analysis, if their Q^2 lies within 0.2 units of the latest, absolute Q^2 maximum. Any SQS simulation, defined by (1) the used compound set **C** and the considered TS/VS splitting scheme **S** = 1..5, (2) the used descriptors **D**, and (3) the nonlinearity policy **N** = "L"(linear), "P"(polynomial), or "N"(nonlinear), will be thus associated with its set **R(C,S,D,N)** of representative model chromosomes. However, for most practical purposes, it makes sense to merge the representative sets of the 5 SQS runs corresponding to different splitting schemes. **R(C,D,N),** the merger of all **R(C,S,D,N)** sets, will serve as a representative sample of built models under given "premises" (e.g., compound set, descriptors and nonlinearity policy).

All the SQS simulations were repeated once—let the corresponding duplicate sets of models be denoted **R′**. SQS reproducibility is assessed by comparing the "twin" sets **R** and **R′**. In order to compare results of SQS run under different premises, the merged sets **R*** = **R**U**R′** from both twin runs will be used.

2.1.4. SQS Consensus Models. SQS CM were built, for each splitting scheme **S**, as the plain averages of the equations in the corresponding merged sets **R***. CM coefficients are set to the average of coefficients in each individual equation from **R*(C,S,D,N)**. Only CM concerning linear ISIDA and FPT based models will be discussed here.

2.2. Stepwise Model Build-Up Procedure. The QSPR/MLRA and Variable Selection Suite (VSS) modules of the ISIDA software have been used to build QSAR models using multilinear regression analysis.

2.2.1. Variable Selection. Two different strategies of stepwise variables selection have been used. The first (SR-1) is based on three steps procedure involving filtering, forward stepwise, and backward stepwise stages.

(1) Filtering Stage. The program eliminates variables D_i which have small correlation coefficient with the activity,

$R_{Y,i} < R^0_{Y,i}$ and those highly correlated with other variables D_j ($R_{i,j} > R^0_{i,j}$), which were already selected for the model. In this work, the values $R^0_{Y,i} = 0.001$ and $R^0_{i,j} = 0.99$ were used. Fragments always occurring in the same combination in each compound of the training set (concatenated fragments) are treated as one extended fragment. "Rare" fragments (i.e., found in less than m molecules, here $m \geq 2$) were excluded from the training set.

(2) Forward Stepwise Preselection Stage. This is an iterative procedure, on each step of which the program selects two variables D_i and D_j maximizing the correlation coefficient $R_{Y,ij} = (R^2_{Y,i} + R^2_{Y,j} - 2R_{Y,i}R_{Y,j}R_{ij})/(1 - R^2_{ij})$ between D_i and D_j and dependent variable Y. At the first step ($p = 1$), the modeled activity for each compound is taken as its experimental one $Y_1 = A_m$. At each next step p, as the activity values Y_p were used residuals $Y_p = Y_{p-1} - Y_{calc}$, where $Y_{calc} = c_0 + c_iD_i + c_jD_j$ is calculated activity by the two-variables model with selected variables D_i and D_j. This loop is repeated until the number of variables k reaches a user-defined value; in this work, k is set to half of the molecule number in the full set.

(3) Backward Stepwise Selection Stage. The final selection is performed using backward stepwise variable selection procedure based on the t statistic criterion.[12−14] Here, the program eliminates the variables with low $t_i = c_i/s_i$ values, where s_i is standard deviation for the coefficient c_i at the ith variable in the model. First, the program selects the variable with the smallest $t < t_0$, then it performs a new fitting excluding that variable. This procedure is repeated until $t \geq t_0$ for selected variables or if the number of variables reaches the user's defined value. Here, t_0, the tabulated value of Student's criterion is a function of the number of data points, the number of variables, and the significance level.

Selected descriptors are used by ISIDA to build the multilinear correlation equation $Y = c_0 + \Sigma c_iD_i$. The Singular Value Decomposition method[15] (SVD) is used to fit the coefficients. The most robust models are selected at the training stage according to statistical criteria.

The SR-1 calculations were initiated from the initial pools of ISIDA and FPT descriptors. Forty-six (CU), 42 (HEPT), and 36 (TIBO) descriptors of both types have been preselected by the forward stepwise preselection procedure. Initial pools of descriptors for CU, HEPT, and TIBO contained, respectively, 1586, 1114, and 576 ISIDA fragments and 1328, 1347, and 1201 of FPT. Further reduction of the number of variables was performed using backward stepwise variable selection procedure based on t statistic criterion allowing building the QSPR models containing desirable number of variables. Eventually, 5−7 models were selected for each splitting scheme **S**. These models included 10-33, 28-39, and 5-27 ISIDA descriptors and 16-23, 14-24, and 10-23 FPT descriptors for CU, HEPT, and TIBO training sets, respectively. The second strategy of variables selection (SR-2) involves splitting of the initial pool of fragment descriptors into subsets, followed by filtering and backward stepwise selection stages. ISIDA generates 319 subsets of fragment descriptors corresponding either to sequences of particular length from n_{min} to n_{max} atoms containing atoms and bonds, atoms only or bonds only, or to augmented atoms. Three or four models per subset possessing reasonable statistical criteria at the training stage have been selected, for each splitting scheme **S**, from the SR-2 calculations. They involve from 9 to 21 descriptors representing the sequences of atoms, bonds, and atoms/bonds containing up to 8 atoms. SR models were subjected to the same external validation procedure using the respective validation sets associated with the considered splitting schemes **S**, with respect to which the validation correlation coefficient R^2_V was calculated according to eq 5.

2.2.2. SR Consensus Models. The ISIDA software may generate CM combining the information issued from several individual models obtained in SR-2 calculations.[16,17] The idea is to use simultaneously a set of best models, for which the values of cross-validation correlation coefficient $Q^2 \geq Q^2_{lim}$, where Q^2_{lim} is a user-defined threshold. Thus, for each compound from the test set, the program computes the activity as an arithmetic mean of values obtained with these models, excluding those leading to outlying values according to Grubbs's test.[18] Our experience shows[10,17,19] that such an ensemble modeling allows one to smooth inaccuracies of individual models. Three CM were prepared for CU, HEPT, and TIBO derivatives ($Q^2_{lim} = 0.7$, 0.85, and 0.65, respectively) based on 11-32, 11-29, and 22-60 individual models for training sets of splitting schemes **S**.

2.3. Compound Sets. The QSAR modeling has been performed on different types of anti-HIV activity for three families of compounds: cyclic ureas (CU), 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymines (HEPT), and tetrahydroimidazobenzodiazepinones (TIBO).

CU Derivatives. The HIV-1 protease inhibition constants K_i for 118 compounds were selected from refs 20−22. Activities $A = \log(1/K_i)$ vary between 5 and 11.

HEPT Derivatives. Effective concentrations of compounds required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (EC_{50}) for 93 molecules have been collected.[23−30] Modeling was performed with respect to the activities $A = \log(1/EC_{50})$, which vary from 3.9 to 9.2.

TIBO Derivatives. The concentration required to 50% the HIV-1 reverse transcriptase enzyme inhibition (IC_{50}) for 84 TIBO derivatives has been critically selected.[31−33] Modeling was performed for the $\log(1/IC_{50})$ values which vary from 3.1 to 8.5.

The data sets (2D structures and activities) are also used in ref 34.

2.4. Molecular Descriptors. Three main categories of descriptors have been considered in this work: (1) **ISIDA:** Substructural molecular fragments including sequences of atoms and bonds (from 2 to 15 atoms per sequence) as well as atoms with their closest environment ("augmented atoms") were generated by ISIDA.[8,12,14] (2) **CAX**: Two-point topological pharmacophore fingerprints and various other descriptors provided by ChemAxon's *generatemd* utility (including calculated logP, logD, the Topological Polar Surface Area, and four BCUT descriptors). All these were obtained with default ChemAxon setups. (3) **FPT**: Three-point topological fuzzy pharmacophore triplets were generated according to setup number one discussed in the above-cited publication.

2.5. Specific Statistical Approaches Used To Compare Model Build-Up Results. The steps undertaken to specifically address the key questions formulated in the Introduction will be briefly underlined in the following:

2.5.1. Reproducibility of SQS Simulations. Two different aspects of SQS reproducibility were addressed:

1. Were the model equations rediscovered when repeating the procedure? A Rediscovery Rate (RR) was estimated by reporting the number of equations visited by both instances of the SQS run to the total number of visited models ("visited" refers to any chromosome for which a trace has been kept, irrespective of fitness and validation score).

2. Do repeated, "twin" SQS runs generate models with similar validation behavior? In this sense, the validation behavior (R^2_V) of models from **R** and **R′** is considered to be a random variable of unknown distribution law. Assuming that **R** and **R′** contain respectively $N_{\mathbf{R}}$ and $N_{\mathbf{R}'}$ models distributed around average values $A(\mathbf{R}) = <R^2_V>_{\mathbf{R}}$ and $A'(\mathbf{R}') = <R^2_V>_{\mathbf{R}'}$ with variances $s(\mathbf{R})$ and $s'(\mathbf{R}')$ respectively, a statistical criterion t rejecting the "null hypothesis" that the two averages A and A' are identical (all the differences being attributable to sampling fluctuations) can be calculated:

$$t = \frac{|A(R) - A'(R')|}{\sqrt{\frac{s^2(R)}{N_R} + \frac{s'^2(R')}{N_{R'}}}} \quad (6)$$

Equation 6 is applicable for arbitrary distribution laws,[35] provided that the numbers of instances N_R and $N_{R'}$ of the random variables are much larger than 30. A large t value signals low SQS reproducibility (repeats lead to model sets of significantly different validation propensities).

2.5.2. Comparison of Model Sets Issued from Different SQS Runs. Several of the key questions addressed in introduction require a methodology to compare the validation propensity distributions of models obtained under different premises. In order to decide whether the differences are significant, the typical stochastic shift observed in respective the twin runs will serve as baseline.

Let $\mathbf{R}_P$ and $\mathbf{R}_Q$ be representative model sets obtained under different premises. For example, $\mathbf{R}_P = \mathbf{R}(\mathbf{C},ISIDA,L)\ \mathbf{U}\ \mathbf{R}(\mathbf{C},ISIDA,P)$ and $\mathbf{R}_Q = \mathbf{R}(\mathbf{C},CAX,L)\ \mathbf{U}\ \mathbf{R}(\mathbf{C},CAX,P)$ to compare linear and polynomial models built on hand of set **C** with ISIDA and CAX descriptors, respectively, etc. The merged sets from the twin SQS calculations, $\mathbf{R}^*_P = \mathbf{R}_P \mathrm{U} \mathbf{R}'_P$ and $\mathbf{R}^*_Q = \mathbf{R}_Q \mathrm{U} \mathbf{R}'_Q$ contain models with average validation propensities $A(\mathbf{R}^*_P) = <R^2_V>_{\mathbf{R}^*P}$ and $A(\mathbf{R}^*_Q) = <R^2_V>_{\mathbf{R}^*Q}$, respectively. According to eq 6 (under consideration of the variances of R^2_V within the two sets $\mathbf{R}^*_P$ and $\mathbf{R}^*_Q$), let t_{P-Q} be the statistical criterion rejecting the hypothesis that the observed shift

$$\Delta_{P-Q} = |A(R^*_P) - A(R^*_Q)| \quad (7)$$

has no statistical significance (is due to normal fluctuations):

$$t_{P-Q} = \frac{\Delta_{P-Q}}{\sqrt{\frac{s^2(R^*_P)}{N_{R^*_P}} + \frac{s^2(R^*_Q)}{N_{R^*_Q}}}} \quad (7a)$$

However, only part of Δ_{P-Q} may be ascribed to the differential impact of working premises P and Q on the model building process, as this magnitude is also influenced by the stochastic noise affecting the calculated averages. In the following, it will be assumed that the noise in Δ_{P-Q} can be related to the t factor of the least reproducible simulations, $max(t_{P-P'},t_{Q-Q'})$, where $t_{P-P'}$ and $t_{Q-Q'}$ are measures of sampling reproducibility under premises P and Q, respectively. Under this conservative assumption, the minimal guaranteed shift δ_{P-Q} that can be directly attributed to the change in sampling premises is

$$\delta_{P-Q} = \Delta_{P-Q} - \max(t_{P-P'},t_{Q-Q'}) \times \sqrt{\frac{s^2(R^*_P)}{N_{R^*_P}} + \frac{s^2(R^*_Q)}{N_{R^*_Q}}} \quad (8)$$

The second term in eq 8 evaluates the random shift attributable to imperfect sampling (s and N representing the validation score variances and the number of models in the merged sets). The above equation amounts to a relative interpretation of t_{P-Q}: the observed shift of average validation propensity Δ_{P-Q} is considered relevant if $t_{P-Q} > \max(t_{P-P'},t_{Q-Q'})$. If $\delta_{PQ} > 0$, one of the premises P and Q is significantly more appropriate for QSAR modeling than the other. Density distribution histograms monitoring the percentage of models in $\mathbf{R}^*$ having R^2_V within each of the ten bins spanning the range 0−1 may provide a detailed illustration of observed differences.

2.5.3. Comparison of the Stepwise and SQS Model Building Strategies. Stepwise regression (SR) leads to a limited set of models that cannot be considered as randomly spread in the problem phase space. Therefore, the statistical treatment envisaged to compare various SQS runs cannot be extended to include models built by the stepwise approach. For this reason, we compared the relative numbers of successfully validating SR and SQS models for which $R^2_V \geq R^2_{V,\mathrm{lim}}$, where $R^2_{V,\mathrm{lim}}$ has been allowed to vary between 0.6 and 0.8.

3. RESULTS AND DISCUSSIONS

3.1. Reproducibility of SQS Runs. The $\sim 2 \times 10^4$ locally fit and diverse chromosomes typically encountered in the global database after SQS completion represent only about 0.1% of the estimated 2×10^7 effectively visited states (from typically 600 MB runs—featuring ~10 meta-generations × 10 triplicate runs/meta-generation, involving two or three parallel executions of the GA sampler—each processing ~300 generations, with ~100 chromosomes/generation). 99.6% of chromosomes were therefore either not fit enough (not even according to early, less constraining fitness standards) or redundant—it is thus irrelevant to include them in the calculation of Rediscovery Rates (RR). The probability that any model chromosome visited and considered for storage by a SQS run will be again encountered upon restarting a sampling process, under identical premises, is of 6% at best and may be as low as 0.03%. A SQS run visits a very limited fraction of the phase space volume of the problem, and the larger the volume, the less the chances to revisit any given chromosome. This is observed upon increasing of the descriptor set size N: for linear models, at $N = 217$, the CAX descriptor space offers by far the smallest sampling volume and scores the highest RRs (between 4% for CU and 6% for TIBO compounds), followed by ISIDA (1%, CU to 3.4%, TIBO) and eventually by FPT (0.5%, CU to 2.7%, TIBO). Phase space volume increases when enabling nonlinearity (with FPT) also triggers a RR decrease

Table 2. Average Validation Score Differences Observed When Repeating the SQS Simulations under Specified Premises[a]

set: CU	ISIDA	CAX	FPT
L	0.017 (3.80)	0.022 (5.79)	**0.053 (13.37)**
P	**0.068 (14.83)**	0.012 (3.75)	0.008 (2.43)
N	N/A	N/A	0.012 (4.68)
set: HEPT	**ISIDA**	**CAX**	**FPT**
L	0.016 (2.52)	*0.034 (7.19)*	0.008 (1.60)
P	0.024 (5.09)	0.021 (3.95)	0.022 (4.60)
N	N/A	N/A	*0.032 (7.28)*
set: TIBO	**ISIDA**	**CAX**	**FPT**
L	0.006 (1.20)	0.016 (2.31)	**0.062 (8.71)**
P	0.015 (3.08)	0.021 (2.59)	0.007 (1.00)
N	N/A	N/A	***0.107 (16.89)***

[a] Compound set − associated with individual tables, nonlinearity policy − in rows, used descriptors − in columns). Calculated *t* factors are given in parentheses. Coding concerns the size of observed shifts: $|A(\mathbf{R})-A'(\mathbf{R}')| < 0.025$: no fill; <0.05: italics; <0.075: boldface; >0.1: italics and boldface.

(0.2%, CU to 2.2%, TIBO for polynomial, but only 0.04%, all sets, for fully nonlinear models).

SQS runs therefore appear to barely "scratch the surface" of the phase space to explore to the point of questioning their usefulness altogether. It is therefore not surprising that the high *t* values reported for the large majority of SQS premises monitored in Table 2 clearly support the hypotheses that observed average validation propensity shifts cannot be ascribed to fluctuations expectable on behalf of two random walks exploring a common phase space zone. Observed shifts are deemed relevant—for example, at $t = 1.64$ there is only a 10% chance that observed shifts are due to fluctuations, while at $t = 3.29$ this chance drops to 0.1%. Statistical significance notwithstanding, shifts are, in general, quite small on an absolute scale: repeated SQS runs return in 15 out of 21 cases $<R^2_V>$ values within 0.025. Shifts do not appear to be related to the total phase space volume, except for the fact that CAX descriptors show, again, the best stability. All the FPT-based nonlinear model sets for CU showed an excellent reproducibility of the average R^2_V values in spite of lowest RRs. While nonlinearity seems to enhance stability of CU models, the HEPT set is characterized by overall low average shifts.

By contrast, the sampling of FPT-based TIBO models in general and of nonlinear models in particular shows significant reproducibility problems. An in-depth analysis of FPT-based nonlinear TIBO models has evidenced a surprisingly strong dependence of the validation propensities on the TS/VS splitting schemes. Only splitting schemes #3 and #5 lead to model sets with very high—and highly reproducible—average validation propensities—for split #2, $A(\mathbf{R}) = 0.52$, $A(\mathbf{R}') = 0.61$, for split #5, $A(\mathbf{R}) = 0.66$, $A(\mathbf{R}') = 0.67$, while for all other splits $0.12 < A < 0.32$. TIBO average validation propensity values are thus very sensitive to the sizes of representative model sets. These sizes may actually vary by a factor of 2, as they basically depend on how quickly the meta-optimization termination criterion is fulfilled. Insofar validation is not strongly splitting scheme-dependent, doubling the size of the solution pool for a specific splitting scheme upon SQS repeat will not impact on the average. This is not the case with TIBO—specifically, the repeat of the TIBO run with the well validating splitting scheme #5 lead to 1193 solutions, compared to only 626 found initially, which is more than sufficient in order to bias the second twin run toward significantly higher global averages.

The splitting scheme dependence of the TIBO model validation propensity is strongest with the pharmacophore triplets. This probably signals the existence of a small subset (<<1/5 of the 73 compounds) featuring a specific but relevant triplet. Validation failures may arise if splitting schemes group most of these examples in VS, so that learning does not have enough examples at hand to pick the key triplet. Switching from triplets to pharmacophore pairs is a radical solution to avoid sparsely populated descriptor matrices. CAX two-point pharmacophore models are expectedly more robust with respect to splitting scheme choice and therefore more reproducible. Unfortunately, this is not a solution to the QSAR problem: data compression upon resolution decrease causes various specific pharmacophore signatures to lose their identities when merged into coarser categories. This may well enhance reproducibility, in the sense of having models reproducibly *failing* to validate. Although fluctuation-prone, FPT TIBO models are as successful in validation tests as their CAX counterparts (see later on).

As a general conclusion, in most situations a single SQS run—or perhaps even fewer MB repeats—may produce a set of models that is representative in terms of validation propensities. Repeated SQS runs will typically discover novel equations but—in most cases—hardly any with radically improved validation behaviors. It may thus be concluded that these QSAR problem phase spaces feature many different attraction pools with roughly the same validation propensities. Several repeats are however mandatory if a complete mapping of the problem space is envisaged, and in view of measured RRs, the cost for enumerating all the properly cross-validating models may easily become prohibitive if $N > 200$.

3.2. Effect of Nonlinearity and Descriptor Choice on Model Validation Propensities. A nearly systematic increase from linear to polynomial to nonlinear models was expectedly witnessed for both Q^2 values and associated training set correlation coefficients (results not shown). The benchmarking study in Table 3 however confirms that overfitting does not occur: no significant validation propensity loss was observed in spite of phase space volume

Table 3. Benchmark of the Relative Average Validation Propensities of Models with Respect to the Nonlinearity Policy and Descriptor Choice[a]

		CU	HEPT	TIBO
N vs L	best	0.677(=)	0.581 (N)	0.420 (=)
	Δ	0.024	0.093	0.051
	δ	–	0.069	–
	t	10.11	28.48	10.65
N vs P	best	0.677 (N)	0.581 (N)	0.420 (=)
	Δ	0.049	0.076	0.080
	δ	0.039	0.051	–
	t	22.91	22.36	16.66
P vs L	best	0.653 (=)	0.505 (=)	0.369 (=)
	Δ	0.025	0.017	0.029
	δ	–	0.003	–
	t	9.48	4.92	5.72
FPT vs ISIDA	best	0.639 (FPT-1)	0.627 (ISIDA)	0.403 (ISIDA)
	Δ	0.071	0.131t	0.049
	δ	0.040	0.126	0.032
	t	33.26	50.60	15.88
FPT vs CAX	best	0.639 (FPT-1)	0.548 (CAX)	0.355 (=)
	Δ	0.048	0.052	0.009
	δ	0.035	0.032	–
	t	26.30	20.62	2.32
ISIDA vs CAX	best	0.591 (=)	0.627 (ISIDA)	0.403 (ISIDA)
	Δ	0.023	0.079	0.057
	δ	–	0.059	0.046
	t	10.91	30.61	17.36

[a] Top half, the pairwise comparisons of the three explored nonlinearity premises (N,P,L), based on FPT descriptors, report the average R^2_V of the winning nonlinearity policy (shown in parentheses, = if no statistically relevant differences exist) the observed average shift Δ − eq 7, the minimal guaranteed average shift δ − eq 8, and the *t* factor of the observed shift − eq 7a. Bottom half, benchmarking of the validation propensities of the joined linear and polynomial model pools obtained with the three different classes of descriptors.

increase. Nonlinear models actually outperform their linear counterparts. HEPT nonlinear models are significantly better validators. With CU and TIBO, they also fare better, but shifts fall short from reaching statistical significance. Nonlinear approaches furthermore outperform polynomial models, being significantly better in two out of three cases. Polynomial and linear models have no significantly differing behaviors. The tendencies underlined by the statistical studies can be intuitively illustrated by density distribution histograms with respect to the validation scores of representative models: in Figure 3, HEPT/FPT-based nonlinear models (grid fill pattern) preferentially accumulate at the top end of the R^2_V scale (X axis, middle plot), unlike the equivalent CU models (upper plot), In the lower plot, it can be seen that the relative proportion of models failing to validate appears to be lower within the nonlinear TIBO models.

According to the bottom half of Table 3, validation propensities—of merged linear and polynomial model sets—in different descriptor spaces are clearly uncorrelated with descriptor set size, or else CAX-based models should have been top performers. Descriptor spaces of dimension $N \leq 1600$ are thus not a prohibitively large "haystack" for SQS to find well validating models. The observed validation propensity differences are thus the expression of the different nature of the chemical information encoded by the descriptors. ISIDA fragment descriptors are outperforming FPT on the HEPT and TIBO sets, while FPT is the most successful in building CU models. This could have been expected, as the latter family of cyclic ureas is built around a single common and large scaffold and is therefore less diverse in terms of represented substructures (out of the $N = 1586$ populated ISIDA fragments, many are common substructures of the cyclic urea scaffold itself). FPT descriptors, encoding specific information about the pharmacophore "ornaments" around the scaffold, are thus better suited for CU QSAR model buildup. CU is furthermore the only set for which CAX descriptors are not outperformed by ISIDA, this underlining the CU preference for a pharmacophore-oriented approach. In this context, the pharmacophore triplets also outperform the CAX pharmacophore pairs, but this trend may vanish (TIBO) or reverse (HEPT) with compound sets of increased internal topological diversity. The validation propensity distribution histograms of Figure 4 clearly illustrate the above-mentioned QSAR-ability differences for HEPT (ISIDA > CAX > FPT) and CU (FPT > ISIDA > CAX). Histograms of winning ISIDA and respectively FPT descriptors witness a clear global shift (lower participations at low R^2_V compensated by higher ones at the high end). With TIBO, however, the situation is less clear-cut: although FPT-based models have an extremely high rate of complete validation failure (30% have $R^2_V < 0.1$), they also have the best rate of strong validation success (at $R^2_V \geq 0.7$). ISIDA-based models are failsafe but not outstanding—rarely completely failing to validate but rarely scoring excellent validation scores—and nevertheless better in terms of average validation propensities.

3.3. Relative Performance of Stochastic QSAR Sampling and Stepwise Regression. Relative validation success rates serve as comparators of SR vs SQS performance. They express the probability that a model picked at random out of the pool of equations provided by the method validates beyond a specified $R^2_{V,\mathrm{lim}}$ criterion. The multiple successful models produced by SQS may be useless if "drowned" in a much larger collection of bad validators.

Table 4 shows that for CU fragment-based approaches the density of well validating SQS models is systematically twice as big as with SR, but both approaches do produce valuable

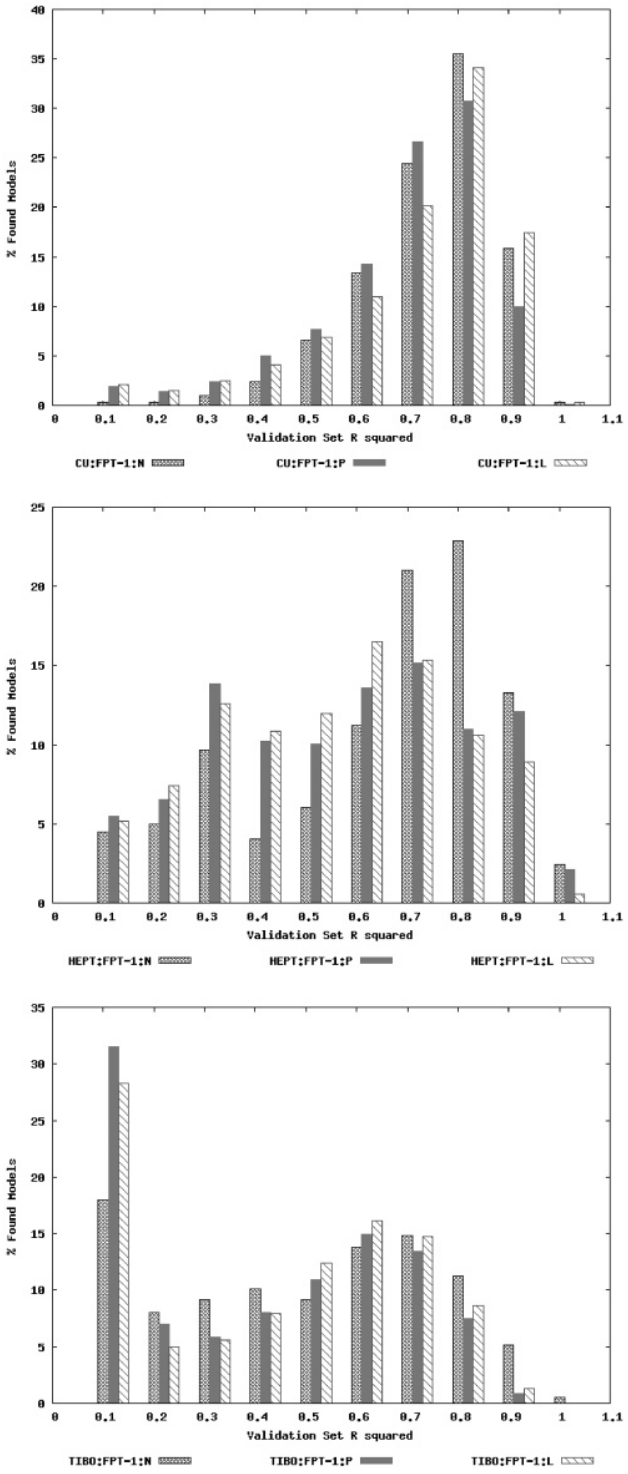

Figure 3. Comparative density distribution histograms of representative nonlinear (grid filling), polynomial (solid gray), and linear (hashed) FPT based models, representing on Y the percentage of models of ***R*** having validation correlation coefficients within each of the 10 bins listed on X (label represents upper bin threshold).

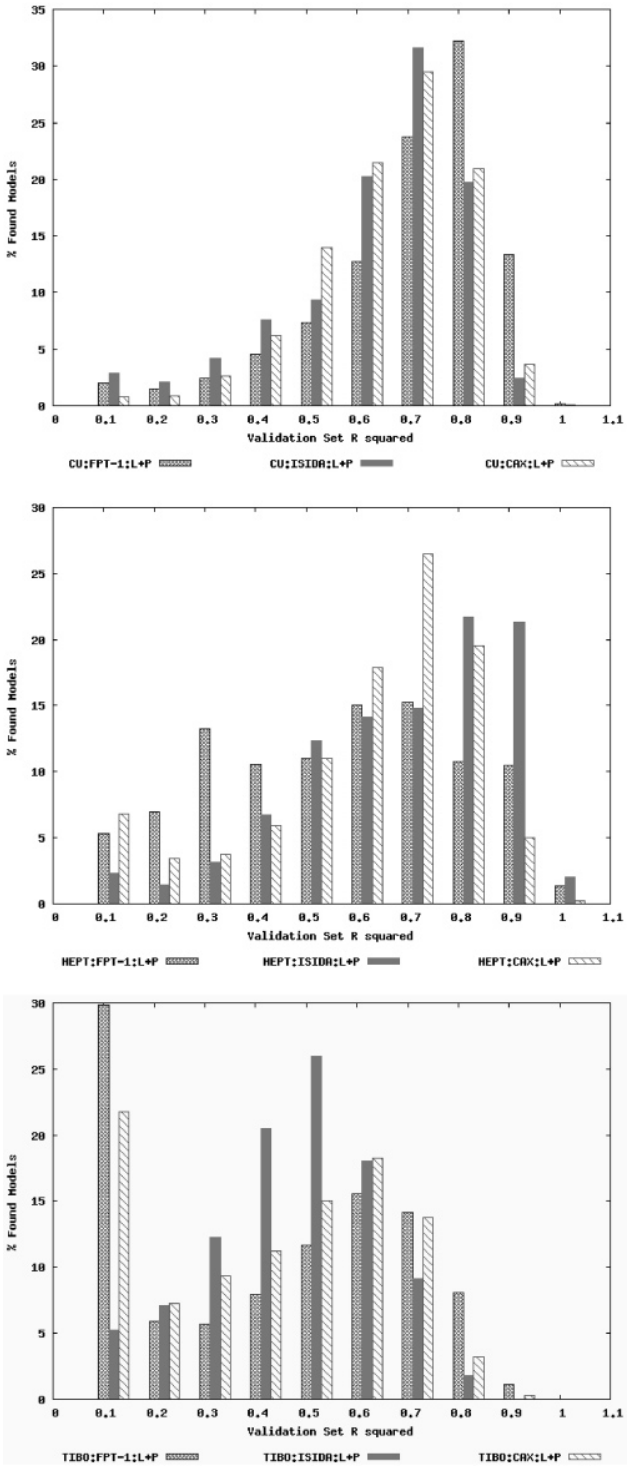

Figure 4. Comparative density distribution histograms of merged linear and polynomial SQS model sets obtained with FPT (grid filling), ISIDA (solid gray), and CAX (hashed) descriptors (check Figure 3 for additional information).

models at any $R^2_{V,\mathrm{lim}}$. For HEPT fragment-based approaches, the percentage of robust models as well as its trend as a function of $R^2_{V,\mathrm{lim}}$ are on the whole quite similar. SQS models are, at the strictest threshold of $R^2_{V,\mathrm{lim}} = 0.8$, at least as likely—and sometimes (HEPT/FPT) significantly more likely—to validate than SR models. The percentages of models exceeding a validation score of 0.6 are also quite satisfactory.

The situation with respect to the TIBO set is clearly the most intriguing: with ISIDA descriptors, the SQS tool is seen to consistently generate many well validating models that are unfortunately outnumbered by the validation failures. The SR approach, by contrast, picks a set of models with excellent validation propensities. When using FPT descriptors, the model family proposed by the SR tool is a family of nonvalidating equations. SQS, by contrast, discovers a

Table 4. Number of Models (and the Percentage They Represent out of the Entire Pool of Found Equations, %) for Which $R^2_V \geq R^2_{V,\mathrm{lim}}$[a]

	$R^2_{V,\mathrm{lim}}$	CU		HEPT		TIBO	
		SR	SQS	SR	SQS	SR	SQS
ISIDA	0.6	51 (33%)	3381 (64%)	124 (78%)	2886 (55%)	74 (31%)	538 (13%)
	0.7	12 (8%)	1624 (31%)	105 (65%)	2152 (41%)	25 (10%)	102 (2%)
	0.8	1 (1%)	207 (4%)	50 (31%)	1220 (23%)	9 (4%)	2 (0.05%)
FPT	0.6	13 (52%)	5971 (72%)	12 (40%)	3254 (36%)	3 (10%)	1437 (25%)
	0.7	10 (40%)	4278 (52%)	7 (23%)	1847 (20%)	0	579 (10%)
	0.8	5 (20%)	1492 (18%)	0	880 (10%)	0	79 (1%)

[a] SR models involving ISIDA descriptors include both the ones issued from SR-1 and SR-2 strategies, whereas those based on the FPT descriptors were all obtained with SR-1. Situations in which the percentages of successful validators are at least twice as large as those seen by the alternative procedure were shaded (7 in favor of SQS, 3 in favor of SR).

Table 5. Validation (R^2_V) Correlation Coefficients and Fraction of Worse Performing Individual Models (*%W*) for SQS and Respectively SR (in Parentheses) Consensus Equations for Each Splitting Scheme **S** and ISIDA Descriptors[a]

	CU		HEPT		TIBO	
S	R^2_V	%W	R^2_V	%W	R^2_V	%W
1	0.42 (0.46)	73 (63)	0.82 (0.81)	96 (88)	0.59 (0.57)	86 (65)
2	0.69 (0.65)	74 (83)	**0.56 (0.70)**	75 (79)	**0.44 (0.67)**	90 (100)
3	*0.80 (0.59)*	90 (64)	0.83 (0.87)	86 (91)	**0.60 (0.74)**	72 (80)
4	*0.76 (0.63)*	87 (83)	0.94 (0.90)	98 (82)	0.46 (0.55)	77 (90)
5	0.78 (0.70)	96 (100)	**0.51 (0.85)**	75 (94)	*0.63 (0.44)*	90 (48)

[a] Cells with either of the SQS or SR consensus equation outperforming the other by more than 0.1 R^2_V units are highlighted: italics when SQS wins, boldface when SR is better (3 in favor of SQS and 4 in favor of SR). "Nonredundancy" is indeed desirable, but a fail-safe definition for nonredundancy is difficult to give.

significant number of successful validators which, furthermore, are less "diluted" by nonvalidators. With the reserve that more studies on different sets should be run in order to reinforce the herein drawn conclusions, it can be said that the SQS approach offers the best guarantees to discover well validating models, if they exist. If SQS were to be used as a generator of numerous equations, only in order to pick one out at random and use it instead of a SR model, then the advantage of SQS is the virtual certainty that well validating models will be *present* in the initial pool of choices. However, SR also lived up to this expectation in 5 challenges out of 6, failing only under the TIBO/FPT premises. In light of this, the advantage of SQS is small and overshadowed by the higher computer cost. The disadvantage of SQS is that the well validating models, although potentially numerous, might yet represent only a small fraction of the entire solution pool. In such cases, picking a single SQS model from the large equation pool may prove a riskier approach than selecting any of the related SR models. However, SQS was only once affected by this caveat in six runs.

3.4. Consensus Strategy. The alternative to random picking single models among the representative ones is the use of consensus equations (Table 5). The R^2_V values of linear SQS and SR (in parentheses) CM built for each of 5 splitting schemes of the 3 compound sets with ISIDA descriptors. Individual models with validation scores below the consensus approach were accounted for as percentages of worse performers *%W*. Typically, SQS CM validate better than 80% of single equations. With certain splitting schemes, they may score low R^2_V values but nevertheless outperform most of individual models. Plain average consensus modeling remains a valid strategy even in the context of the many diverse equations from the representative SQS sets.

SR and SQS performances are again quite similar, in spite of the more sophisticated build-up procedure of SR CM, using outlier detection:[18] a count of situations in which either of the methods outperforms the other by more than 0.1 R^2_V score units reveals 4 in favor of SR, 3 in favor of SQS, and 8 draws.

3.5. Beyond Statistics: What Can Be Learned from Problem Space Mapping? If we take the ultimate goal of QSAR studies to be the discovery of at least one well-validating equation, then SQS does not appear be worth the excess computer effort (days of work on X86 biprocessor workstations, rather than hours). Both methods perform quite well—but then, *all* the QSAR models that were ever published do perform well in terms of validation tests, though few were reported to discover actives in actual virtual screening. In light of the insights gained from this study, this is not surprising. In a QSAR problem space with few well validating models, SR would be the ideal modeling tool—in as far as it succeeds to discover them. Such sets may exist, but none was encountered here—these feature *several broad zones populated by models of comparable validation propensity*. In there, SR equations have no special status—they are just typical models among many thousands. Similar results might be obtained with less sophisticated stochastic samplers[36] also relying on upfront descriptor candidate filtering. That any two well validating models may nevertheless have different application ranges and return diverging predictions when applied to external compounds is a fact as obvious as the "Kubinyi paradox".[16] A *Gedanken experiment* suffices to explain why: consider a mixed set of pharmacophore and electronic effect descriptors on a series in which a substituted phenyl ring is a key feature. It was found that -OH, -OR, -SR, $-NR_2$, or -NHR substituted compounds tend to be active, while -halogen, -H, and -alkyl substituted ones are not. There are two alternative explanations: "hydrogen bond acceptor required" (pharmacophore descriptor entering the model) or "electron-enriched phenyl required" (electronic effect descriptor chosen). They are of comparable predictive power and validation propensity: since the phenyls carrying an acceptor are—as far as this set goes—electron-enriched, pharmacophore and electronic effect descriptors are strongly correlated. SR discards one of the two and comes up with one single alternative. Therefore, correct prediction of the say carbonyl-substituted analog is a matter of luck: only the pharmacophore model returns "active" (−CR=O is an acceptor). As both training and test sets fail to include such an example—for good reason, perhaps: electron-withdrawing effects may cause synthesis

problems—SR does not have any means to issue a warning about this intrinsic degeneracy of the chemical information in the training set. By contrast, SQS may correctly enumerate both alternatives. There is still no way to know which is mechanistically correct, but having predictions carried out with both of these apparently indiscernible models may at last evidence the inherent limitations of the training set and suggest how to enrich it.

In the early days of QSAR, aggressive pruning of the set of initial descriptors was a technical constraint, and eliminating correlated terms was the less worse of arbitrary choices forced upon the user. Descriptor correlations are often training/test set specific (even with thousands of compounds[37]). Moreover, the small difference between two large and correlated terms may nevertheless "hide" a genuine independent variable—a well-known example being the free energy, not related to either enthalpy or entropy although the latter two are often correlated.[38] In spite of such sources of potential pitfalls, the paradigm of the nonredundant descriptor set became enshrined in QSAR building protocols, although the technical bottleneck at its origin has long since vanished. In our opinion, the argument that SR approaches are superior to stochastic sampling because they return few models and thus avoid the question of which one to actually use in drug design is fallacious. An extensive mapping of the QSAR problem space may be the key to reducing the overall failure rate of QSAR-driven virtual screening, but classical validation tests cannot shed light on how to best use the wealth of SQS-generated information. This work proves that, with hundreds of compounds and thousands of candidate descriptors, such a complete mapping would be feasible in a matter of weeks using state-of-the-art desktop PCs. Further effort will be dedicated to understand how to best use problem space maps in virtual screening.

4. CONCLUSIONS

The SQS model build-up procedure reported and tested in this paper successfully discovered multiple, successfully validating models under various working premises. This section sums up the observations with respect to the key questions in the Introduction and ends with a general debate of the insights gained due to this work.

1. Reproducibility. Given the huge problem space volumes it is meant to sample, SQS will never list all possible QSAR models nor rediscover the same if repeated. This notwithstanding, the model sets it actually produces display comparable validation performances. Sets of SQS equations close to optimal cross-validated Q^2 were found to be rich in successfully validating models.

2. Dependence on Phase Space Volume: Nonlinearity Policy and Descriptor Choice. Average validation propensities of SQS models are independent of problem space volumes: neither the introduction of preset nonlinear transformations nor moving from smaller to larger descriptor sets triggered overfitting artifacts. On the contrary, the introduction of nonlinearity actually had a positive impact. This proves that appropriate model building—pressure to minimize the number of entering variables and systematic cross-validation as part of the model fitness estimation—may avoid overfitting.

3. Stochastic vs Stepwise Model Building: Comparison of Individual Models. While SR may occasionally fail to produce any validating models when SQS succeeds (this has been observed once in six cases covered by the study), the latter approach may occasionally "hide" the numerous, valuable equations within an even larger set of nonvalidators (seen once in six cases, as well). Pharmacophore descriptors are more likely to cause SR failure with the current sets.

4. Consensus Strategy. Consensus SQS models were found to display an extremely robust behavior, virtually always showing better characteristics than 70−90% of individual models. The scale-up from tens of related SR models to 10^4 randomly picked SQS equations does not impede on the validity of the CM paradigm. SQS CM are however not outstandingly better, nor worse, than SR consensus equations as far as the validation exercise may tell, but they might have decisive advantages when used in virtual screening of large databases.

Finding more models does not automatically imply finding much better models—in the sense of standard intrafamily validation tests. Such (necessary, but hardly sufficient) tests are not the ultimate QSAR validity criterion and cannot tell which is the mechanistically sound equation out of many "apparently" equivalent forms (equivalent as far as training and validation sets go but not necessarily throughout the space of druglike compounds). The benefits of a global analysis of QSAR problem space must therefore be addressed from the more general point of view of actual utility in drug design. SQS-driven model sampling might be used to get an idea on the degree of degeneracy of the chemical information in the training set and on the novel compounds to add to training in order to lift some of these degeneracies.

Abbreviations: **QSAR** − quantitative structure−activity relationships, **TS/VS** − training/validation set, **FPT** − fuzzy pharmacophore triplets, **ISIDA** − fragment descriptors, **CAX** − ChemAxon descriptors, **CM** − consensus models, **MLR** − multilinear regression, **SR** − stepwise regression, **GA** − genetic algorithm, **MB** − Model Builder: distributed, GA-based QSAR model generator, **SQS** − stochastic QSAR sampler, managing the parameter control and centralizing the output of repeated MB runs, **RR** − rediscovery rate

Supporting Information Available: Data sets, splitting schemes, and molecular descriptors plus an information file describing the exact contents of each data file—everything under Unix text format. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Towards generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and related methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.
(2) Milicevic, A.; Nikolic, S.; Trinajstic, N. Toxicity of aliphatic ethers: A comparative study. *Mol. Diversity* **2006**, *10*, 95−99.
(3) Adam, M. Integrating research and development: the emergence of rational drug design in the pharmaceutical industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.
(4) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(5) Baumann, K. Cross-validation as the objective function for variable selection techniques. *TrAC, Trends Anal. Chem.* **2003**, *22*, 395−406.
(6) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, published on Web 10.21.2006.

(7) Parent, B.; Kökösy, A.; Horvath, D. Optimized Evolutionary Strategies in Conformational Sampling. *Soft Computing* **2007**, *11*, 63−79.
(8) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847−858.
(9) http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Feb 2, 2006).
(10) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.
(11) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.
(12) Varnek, A.; Fourches, D.; Solov'ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D.; Katritzky, A. R. "In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests. *J. Chem. Inf. Comput. Sci.* **2005**, *44*, 1365−1382.
(13) Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. QSAR modeling of blood:air and tissue:air partition coefficients using theoretical descriptors. *Bioorg. Med. Chem.* **2005**, *13*, 6450−6463.
(14) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693−703.
(15) Golub, G. H.; Reinsch, C. Singular value decompositions and least squares solutions. *Numer. Math.* **1970**, *14*, 403−420.
(16) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Dobchev, D. A.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. Correlation of blood-brain penetration using structural descriptors. *Bioorg. Med. Chem.* **2006**, , *14*, Jul 15, 4888−4917.
(17) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Ionophores. *J. Chem. Inf. Model.* **2006**, *46*, 808−819.
(18) Grubbs, F. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, *11*, 1−21.
(19) Varnek, A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F. Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 812.
(20) Wilkerson, W. W.; Akamike, E.; Cheatham, W. W.; Hollis, A. Y.; Collins, R. D.; DeLucca, I.; Lam, P. Y.; Ru, Y. HIV Protease Inhibitory Bis-benzamide Cyclic Ureas: A Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1996**, *39*, 4299−4312.
(21) Wilkerson, W. W.; Dax, S.; Cheatham, W. W. Nonsymmetrically Substituted Cyclic Urea HIV Protease Inhibitors. *J. Med. Chem.* **1997**, *40*, 4079−4088.
(22) Lam, P. Y.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLucca, G. V.; Eyermann, C. J.; Chang, C. H.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Hodge, C. N. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. *J. Med. Chem.* **1996**, *39*, 14−3525.
(23) Miyasaka, T.; Tanaka, H.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1 Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1989**, *32*, 2507−2509.
(24) Tanaka, H.; Baba, M.; Hayakawa, H.; Haraguchi, K.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Walker, R. T.; De Clercq, E. Lithiation of uracil nucleosides and its application to the synthesis of a new class of anti-HIV-1 acyclonucleosides. *Nucleosides Nucleotides* **1991**, *10*, 397−400.
(25) Tanaka, H.; Baba, M.; Saito, S.; Miyasaka, T.; Takashima, H.; Sekiya, K.; Ubasawa, M.; Nitta, I.; Walker, R. T.; Nakashima, H.; De Clercq, E. Specific anti-HIV-1 acyclonucleosides which cannot be phosphorylated: synthesis of some deoxy analogs of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine. *J. Med. Chem.* **1991**, *34*, 1508−1511.
(26) Tanaka, H.; Baba, M.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 1394−1399.
(27) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubawasa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV-1 Specific 6-Substituted Acyclouridine Derivatives: Synthesis and Anti-HIV-1 Activity of 5- or 6-Substituted Analogs of 1-[(2-Hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349−357.
(28) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and antiviral activity of deoxy analogs of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) as potent and selective anti-HIV-1 agents. *J. Med. Chem.* **1992**, *35*, 4713−4719.
(29) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure-activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)-thymine analogs: effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity. *J. Med. Chem.* **1992**, *35*, 337−345.
(30) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of 6-Benzyl Analogs of 1-[(2-Hydroxyethoxy)methyl]-5-(phenylthio)thymine (HEPT) as Potent and Selective Anti-HIV-1 Agents. *J. Med. Chem.* **1995**, *38*, 2860−2865.
(31) Kukla, M. J.; Breslin, H. J.; Pauwels, R.; Fedde, C. L.; Miranda, M.; Scott, M. K.; Sherrill, R. G.; Raeymaekers, A.; van Gelder, J.; Andries, K.; Moens, L. J.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo[4,5,1-jk]-[1,4]benzodiazepin- 2(1H)-one (TIBO) derivatives. *J. Med. Chem.* **1991**, *34*, 746−751.
(32) Ho, W.; Kukla, M. J.; Breslin, H. J.; Ludovici, D. W.; Grous, P. P.; Diamond, C. J.; Miranda, M.; Rodgers, J. D.; Ho, C. Y.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo-[4,5,1-jk][1,4]benzodiazepin- 2(1H)-one (TlBO) derivatives. 4. *J. Med. Chem.* **1995**, *38*, 794−802.
(33) Breslin, H. J.; Kukla, M. J.; Ludovici, D. W.; Mohrbacher, R.; Ho, W.; Miranda, M.; Rodgers, J. D.; Hitchens, T. K.; Leo, G.; Gauthier, D. A.; Ho, C. Y.; Scott, M. K.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo [4,5,1-jk][1,4]-benzodiazepin-2(1H)-one (TIBO) derivatives. 3. *J. Med. Chem.* **1995**, *38*, 771−793.
(34) Solov'ev, V. P.; Varnek, A. Anti-HIV activity of HEPT, TIBO, and cyclic urea derivatives: structure-property studies, focused combinatorial library generation, and hits selection using substructural molecular fragments method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1703−19.
(35) Grais, B. L'interprétation des sondages aléatoires : problèmes d'estimation et de comparaison. In *Méthodes Statistiques*, 3rd ed.; Dunod Eds.; Dunod: Paris, France, 1992; pp 288−296.
(36) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J. Med. Chem.* **2005**, *48*, 6563−74.
(37) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Cheminformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Ed.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; pp 117−137.
(38) Ford, D. M. Enthalpy-Entropy Compensation is Not a General Feature of Weak Association. *J. Am. Chem. Soc.* **2005**, *127*, 16167−16170.

CI600476R

Septième partie

Les empreintes pharmacophoriques tricentriques floues – 2eme partie : Utilisation des triplets pharmacophoriques flous dans des études de Relation Structure-Activité (*QSAR*).

Les bons comportements au voisinage des triplets pharmacophoriques flous (2D-FPTs) observés dans les travaux précédents nous ont amené à vouloir les tester dans un contexte de mise en place et de validation d'études de Relation Structure-Activité ($QSAR$). En effet, leur capacité à traduire des informations chimiquement pertinentes ainsi que leur faible coût en temps machine étaient des atouts qui pouvait donner lieu à la mise en place de modèles puissants.

Dans les travaux suivants, la capacité des 2D-FPTs à mettre en relation la structure moléculaire à l'activité biologique a été évaluée. Les données de nos tests proviennent de treize ensembles déjà utilisés dans la littérature pour des études de $QSAR$. Ceci nous a permis de pouvoir comparer les performances des 2D-FPTs par rapport à d'autres descripteurs variés, utilisés pour la création de modèles $QSAR$. Afin de conserver la cohérence de nos études, la répartition faite par les auteurs en ensembles d'entrainement et de validation a été conservée.

Des modèles $QSAR$ linéaires et non-linéaires ont été construits grâce au SQS présenté précédemment (voir Chapitre IV), pour chaque série de composés. Trois versions différentes des 2D-FPTs ont été utilisées. De plus, une variante des 2D-FPTs ne prenant pas en compte le pK_a des composés a été prise en compte dans l'étude. Enfin, l'impact de la projection floue a été étudié, par rapport à la projection stricte des triplets.

Dans cette étude, nous verrons que les modèles basés sur les 2D-FPTs montrent à nouveau de bons résultats. Dans la majorité des cas (10 fois sur 13), les modèles sont tout aussi bien voire mieux validés que les meilleurs modèles décrits dans la littérature. Les triplets montrent à nouveau leur efficacité : la plupart des séries d'analogues ont été bien décrites, que ce soit par les 2D-FPTs basés sur des règles ou par les 2D-FPTs dépendants du pK_a. Une exception notable a été mise en avant dans la série des inhibiteurs de la thermolysine. En effet, bien qu'il n'y ait pas d'effet apporté par l'équilibre protéolytique, les 2D-FPTs basés sur le pK_a ont augmenté la qualité du modèle.

Il est à noter que le degré optimal de flou à utiliser dans les 2D-FPTs est dépendant du set de composés. Cette remarque est cependant cohérente par rapport à nos autres observations indiquant qu'il est important de bien choisir son ensemble de descripteurs par rapport aux donnés d'entrée.

.

Cependant, malgré la richesse des ensembles de composés étudiés, les évaluations effectuées dans cette étude sont mises en défaut par la diversité trop basse au sein des ensembles. Toute une série d'artefacts dus à cette non-diversité a pu être mise en évidence.

Afin d'illustrer cette observation, une investigation en profondeur du modèle mis en place pour les inhibiteurs de la thrombine a été mise en place. Elle a révélé que la sélection de certains des triplets a effectivement du sens (certains d'entre eux représentent une bonne couverture pharmacologique et topologique des poches de liaison P1 et P2). Malgré cela, les équations ont été incapables de prédire l'activité de ligands structuralement différents, ou ont prédit de façon non-discriminante que n'importe quel composé hors de l'ensemble d'entrainement était actif.

Les modèles *QSARs* basés sur les 2D-FPTs ne dépendent pas d'une structure commune qui serait requise pour des superpositions de molécules. En principe, ces modèles devraient être entrainés sur divers ensembles, ce qui est indispensable si l'on souhaite pouvoir obtenir des modèles applicables sur un grand nombre de problèmes. Afin de pallier au problème de la non-diversité, il est nécessaire d'ajouter à l'ensemble d'entrainement des (supposés) composés inactifs appartenant à des familles diverses. Ceci permettrait la découverte de modèles capables de reconnaître spécifiquement les actifs ayant une structure différente.

Les FPTs ont été adoptés par des partenaires industriels dans le cadre de collaborations. Ils ont permis la découverte de nouveaux composés bioactifs. Ainsi, un modèle d'inhibition des cytochromes a été développé avec Servier Orléans. Avec Servier Croissy-sur-Seine, des inhibiteurs ont été conçus à partir des modèles *QSAR* par *SQS* pour 3 projets différents. Deux projets ont rencontré du succès, le troisième est actuellement en attente. Sur 3014 molécules testées, 9 ont été trouvées avec un $K_i <= 1$nM, 34 entre 1 et 10nM et 71 entre 10 et 99nM.

Huitième partie

Fuzzy Tricentric pharmacophore Fingerprints. 2. Application of Topological Fuzzy pharmacophore Triplets in Quantitative Structure-Activity Relationships

Reprinted with permission from F. Bonachera and D. Horvath. Fuzzy Tricentric pharmacophore Fingerprints. 2. Application of Topological Fuzzy pharmacophore Triplets in Quantitative Structure-Activity Relationships *J. Chem. Inf. Model.*, 48 :409-425, **2008**.

Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure−Activity Relationships

Fanny Bonachéra and Dragos Horvath*

UMR 8576 CNRS − Unité de Glycobiologie Structurale & Fonctionnelle, Université des Sciences et Technologies de Lille, Bât. C9-59655 Villeneuve d'Ascq CEDEX, France

Received August 30, 2007

Topological fuzzy pharmacophore triplets (2D-FPT), using the number of interposed bonds to measure separation between the atoms representing pharmacophore types, were employed to establish and validate quantitative structure−activity relationships (QSAR). Thirteen data sets for which state-of-the-art QSAR models were reported in literature were revisited in order to benchmark 2D-FPT biological activity-explaining propensities. Linear and nonlinear QSAR models were constructed for each compound series (following the original author's splitting into training/validation subsets) with three different 2D-FPT versions, using the genetic algorithm-driven Stochastic QSAR sampler (SQS) to pick relevant triplets and fit their coefficients. 2D-FPT QSARs are computationally cheap, interpretable, and perform well in benchmarking. In a majority of cases (10/13), default 2D-FPT models validated better than or as well as the best among those reported, including 3D overlay-dependent approaches. Most of the analogues series, either unaffected by protonation equilibria or unambiguously adopting expected protonation states, were equally well described by rule- or pK_a-based pharmacophore flagging. Thermolysin inhibitors represent a notable exception: pK_a-based flagging boosts model quality, although—surprisingly—not due to proteolytic equilibrium effects. The optimal degree of 2D-FPT fuzziness is compound set dependent. This work further confirmed the higher robustness of nonlinear over linear SQS models. In spite of the wealth of studied sets, benchmarking is nevertheless flawed by low intraset diversity: a whole series of thereby caused artifacts were evidenced, implicitly raising questions about the way QSAR studies are conducted nowadays. An in-depth investigation of thrombin inhibition models revealed that some of the selected triplets make sense (one of these stands for a topological pharmacophore covering the P_1 and P_2 binding pockets). Nevertheless, equations were either unable to predict the activity of the structurally different ligands or tended to indiscriminately predict any compound outside the training family to be active. 2D-FPT QSARs do however not depend on any common scaffold required for molecule superimposition and may in principle be trained on hand of diverse sets, which is a must in order to obtain widely applicable models. Adding (assumed) inactives of various families for training enabled discovery of models that specifically recognize the structurally different actives.

1. INTRODUCTION

The recent development of topological fuzzy pharmacophore fingerprints[1] 2D-FPT, shown to display excellent neighborhood behavior,[2] naturally raised the question of their potential applications in quantitative structure−activity relationships[3−5] (QSARs), empirical mathematical models returning an estimate of the molecular activity as a function of structural descriptors. Relationships between activity and pharmacophore feature distribution descriptors have been intensely studied in chemoinformatics, either in terms of (a) QSAR model buildup or (b) binding pharmacophore[6−8] elucidation attempts. There is however no fundamental distinction between (a) and (b)—selecting and weighing specific elements of the vector describing the overall pharmacophore pattern of a molecule, as in (a), may in principle allow the backtracking of the important, activity-enhancing variables to the actual pharmacophore features in the molecules and thus translate a QSAR model into a pharmacophore hypothesis in the sense of (b). With molecular[9] or pharmacophore field[10] maps of the space surrounding the studied ligands, the space zones corresponding to the relevant field terms may be readily assimilated to the hypothesized binding site regions involved in interactions. This very tempting and straightforward interpretation of CoMFA[9] models has largely contributed to the success of the approach, albeit authors sometimes tend to forget that a statistically valid correlation is not enough evidence for a cause-to-effect relationship between correlating magnitudes.[11] Molecular field maps do however require the construction and alignment of one or several conformer(s) for each compound. Computer-effective overlay-independent descriptors of the pharmacophore patterns typically rely on autocorrellograms[12,13] and pair density distributions.[14] These 'encrypt' the pharmacophore/field pattern information, providing a less straightforward link between descriptors and structural elements in the molecules, but are nevertheless successful in QSAR.[15,16] Pharmacophore triplets[17] or quadruplets[18] provide an even more detailed description, but large size and strong conformer-dependence of 3D triplets dissuaded scientists to use these otherwise than in similarity searches, until recently.[19]

* Corresponding author phone: +333.20.43.49.97; fax: +333.20.43.65.55; e-mail: dragos.horvath@univ-lille1.fr, d.horvath@wanadoo.fr.

10.1021/ci7003237 CCC: $40.75
Published on Web 02/07/2008

Coding for population levels of specified (setup-dependent) pharmacophore triplets, with topological distances used as a metric of the relative positions of the atoms representing pharmacophore features[20] (hydrophobicity, aromaticity, hydrogen bond donors/acceptors, cations, anions), 2D-FPT contain in principle all the chemical information required to elucidate a binding pharmacophore. They should be able to model compound recognition by an active site, when fed into a QSAR building tool (descriptor selection and weighing procedure) set to detect (i) pharmacophore triplets selectively populated in actives, which are supposed to enhance binding, and (ii) triplets selectively populated in inactives, which are supposed to block it. However, the topological nature of 2D-FTP, only implicitly and imprecisely accounting for the actual 3D interfeature distances perceived by the receptor, is a further obstacle in the way of straightforward mechanistic interpretation. The assumption that triplets (i) actually stand for ligand atoms favorably interacting with the site, while (ii) include atoms clashing with the site, may be far-fetched.

This notwithstanding, 2D-FPT models may still convey more physicochemically meaningful information than equations based on abstract topological indices. Success of 2D-FPT descriptors in QSAR would be good news, because they are computationally cheap (no 3D structure generation and alignment required). This benchmarking study revisits 13 ligand/inhibitor sets concerning various receptors, from various literature sources[13,21−24] where various QSARs, including high-end, 3D overlay-dependent CoMFA models, were proposed. Linear models were generated for all these sets, using the following:

1. Three differently parametrized versions of 2D-FPT, differing in terms of the 'grid mesh' size (the step used to enumerate edge lengths of the considered basis triplets), the minimal and maximal considered edge lengths, etc. These include the 'default' (D) and the 'optimal' (O) setups reported in the original paper, plus a 'coarse' version (C).

2. Two variants based on the default 2D-FPT scheme but using a rule-based pharmacophore feature assignment procedure instead of the one based on calculated pK_a values for ionizable groups (D-R, for 'R'ule-based) and, respectively, abandoning the fuzzy mapping of molecular triplets onto basis triplets in favor of strict matching (D-S, for 'S'trict matching).

Using the stochastic QSAR sampler (SQS),[25] a set of relevant QSAR equations (having cross-validation scores within the upper end of the spectrum generated at model training) was kept and confronted with validation compounds, for each compound set and 2D-FPT version. SQS-generated relevant model sets typically feature thousands of independent equations selected due to their high cross-validation scores, but literature studies only present a few individual models. Therefore, the benchmarking only reports the statistical criteria of the best validating model from each representative set. As the initial splitting schemes into training/validation sets have been scrupulously followed here, the root-mean-squared prediction errors with respect to validation set compounds (not excluding any outliers) are directly comparable, irrespective of the nature of the reference models (regression, PLS, neural network) and their original calibration procedures (stepwise, stochastic, PLS, etc.).

This study continues with benchmarking the relative performance of different 2D-FPT versions against each other, in order to shed some light on the question of how to optimally generate QSAR-proficient 2D-FPT. Both the impact of explicit modeling of proteolytic equilibria vs rule-based pharmacophore feature assignment and the influence of fuzzy triplet mapping were assessed. Duplicate SQS runs were performed with both default 2D-FPT, the rule-based version D-R, and the nonfuzzy version D-S, in order to generate extended representative model sets, allowing a comparison of average validation propensities according to a previously described approach.[25] Comparison of QSARs based on triplet versions D, O, and C relies on the validation statistics of best validating models, like in literature model benchmarking.

Next, a comparison of best validating linear vs nonlinear SQS models is undertaken in order to further confirm the previously observed trend[25] of improving validation propensities when allowing for preset nonlinear transformations to enter the models.

Eventually, thrombin inhibition models are challenged to predict the affinity of chemically different, cocrystallized thrombin ligands, including two amidine[26,27] derivatives and a pyrazinone[28] adopting a different binding mode. The groups seen to directly interact with the thrombin site will be matched against the atom triplets corresponding to selected 2D-FPT elements.

This paper is structured as follows: in Methods, a brief revisiting of 2D-FPT and of the SQS model building methodology will continue with an outline of the statistical criteria used for benchmarking. An introduction of the employed data sets, followed by the details on the assessment of structural interpretability of 2D-FPT models, complete this section. Results and Discussions will first address the various benchmarking aspects: comparison with literature models, comparative assessment of pharmacophore flagging strategies, of the use of fuzzy logic for triplet generation and of the nonlinearity policy for SQS model buildup. The next addressed key point will be the extrapolability of the trained QSAR models to compounds of different topology. The section will close with the discussion of selected triplets as 'topological pharmacophores' matched against the actual pharmacophore points in the structures of cocrystallized ligands. The Conclusions paragraph, concerning the usefulness and interpretability of 2D-FPT triplets as QSAR descriptors, will be extended to a general discussion about the limitations of QSAR buildup and benchmarking caused by restricted training/validation set diversity, in light of the artifacts and chemically meaningless terms seen to enter some of the nonetheless well validating QSARs.

2. METHODS

2.1. 2D-FPT Buildup. 2D-FPT buildup has been described in detail elsewhere.[1] A basis set of reference pharmacophore triplets is chosen, enumerating all possible combinations of pharmacophore features (Hp-hydrophobic, Ar-aromatic, HA-hydrogen bond acceptor, HD-donor, PC-positive charge, NC-negative charge) of the corners, times all the considered integer edge lengths obeying triangle inequalities, within a finite range [E_{min}, E_{max}] and sampled by an E_{step} controlling the graininess of 2D-FPT. Next, all triplets of features

Table 1. Parameters Controlling 2D-FPT Buildup − Three Considered Setups: D − Default Setup and O − Optimal Setup (Maximizing NB) from Previous Work[1] and C − Coarse Setup[29]

parameter	description	D	O	C
E_{min}	minimal edge length of basis triangles (number of bonds between two pharmacophore types)	2	4	5
E_{max}	maximal triangle edge length of basis triangles	12	15	15
E_{step}	edge length increment for enumeration of basis triangles	2	2	3
E	edge length excess parameter: in a molecule, triplets with edge length > E_{max}+e are ignored	0	2	2
Δ	maximal edge length discrepancy tolerated when attempting to overlay a molecular triplet atop of a basis triangle	2	2	3
$\rho_{Hp} = \rho_{Ar}$	Gaussian fuzziness parameter for apolar (hydrophobic and aromatic) types	0.6	0.9	0.7
$\rho_{PC} = \rho_{NC}$	Gaussian fuzziness parameter for charged (positive and negative charge) types	0.6	0.8	0.3
$\rho_{HA} = \rho_{HD}$	Gaussian fuzziness parameter for polar (hydrogen bond donor and acceptor) types	0.6	0.7	0.2
L	aromatic-hydrophobic interchangeability level	0.6	0.5	0.7
	number of basis triplets at given setup	4494	7155	2625

represented in a molecule are analyzed, following a protonation state-dependent pharmacophore typing of the atoms, using shortest-path topological interatomic distances as actual edge lengths. Molecular triplets are then mapped onto basis triplets, using fuzzy logic (each molecular triplet may contribute to the population levels of several similar basis triplets, by increments directly related to their degree of similarity). Total population levels of basis triplets form a sparse vector, the 2D-FPT descriptor, with nonzero elements corresponding to the basis triangles that are either present per se or are represented by similar triplets in the molecule. Table 1 reports the specific setups used to generate the 2D-FPT versions used in this paper, where 'D' and 'O' correspond to the two setups already discussed in the original publication (therein called FPT-1 and FPT-2, respectively; labels 'D' and 'O' recall that the former is a default setup, while the latter was shown to optimize NB). An additional 'C'oarse fingerprint 'C' has been considered here, using a larger E_{step} of 3 and thus relying upon a significantly smaller basis triplet set, while still preserving excellent NB.[29] Two additional variants of the default 2D-FPT were also considered: D-R using a 'R'ule-based pharmacophore feature assignment strategy rather than the one based on predicted pK_a values for ionizable groups, and D-S, the 'S'trict fingerprint mapping molecular triplets strictly onto the identical basis triplets or ignoring them as no such triplet is listed within the reference basis set. Both had their NB tested in the original 2D-FPT publication.

2.2. The Stochastic QSAR Sampler (SQS). SQS is based on a hybrid parallelized genetic algorithm-driven engine for selecting both relevant descriptors and their optimal nonlinear transformation rules to enter a model. It uses randomized leave-1/3-out cross-validation to let the in silico Darwinian selection process pick the most robust models. SQS has been shown able to typically retrieve thousands of not overfitted QSAR equations, which successfully passed the subsequent external validation tests.[25] It may, upon request, exclusively mine for linear equations or try to select the most appropriate among a set of predefined nonlinear transformation functions to be used in conjunction with any given descriptor. Linear models have been generated for benchmarking purposes against literature equations, most of which were linear as well. Nonlinear QSARs were also systematically built for all the data sets, in quest of equations potentially outperforming the linear models. SQS proceeds by successively running a Model Builder (MB) with varying operational control parameters which are tuned on-the-fly to maximize MB sampling performance. After each MB run, a set of most relevant models found up-to-date are extracted, using error pattern similarity to decide which models are redundant. At the end, these sets of locally most representative equations are merged, and all models having their leave-1/3-out cross-validated correlation coefficient within a window of 0.2 units at the top end enter the final 'representative' model pool of that simulation.

For each of the 13 compound sets, SQS mining for linear models was systematically performed with each of the 5 considered 2D-FPT versions. Linear simulations with D, D-R, and D-S descriptors have been duplicated, and representative model pools were merged in order to allow the estimation of the average validation correlation coefficient shifts attributable to switching from one descriptor variant to the other, according to a previously outlined formalism[25] (also see below). Eventually, a second round of SQS simulations mining for fully nonlinear models was also carried out for all compound sets using all descriptor versions, leading to a total of 5 × 13 (first round, linear) + 3 × 13 (linear model generation duplicates: D, D-R, and D-S descriptors only) + 5 × 13 (nonlinear) = 169 different SQS runs using various Linux and IRIX workstations of the laboratory. This effort lead to a total of 236 852 individual QSAR equations, members of the respective representative sets, all compound series confounded.

2.3. Statistical Criteria Used for Benchmarking. This work only reports statistical criteria with respect to the external validation sets used in literature and taken over as such in the present work. Training set and cross-validation criteria are either uninteresting (some general information concerning training set R^2_T values of the selected best validating models will be given in the Results section) or not directly comparable to literature values (cross-validation schemes differ from author to author). The key benchmarking criterion used here is the root-mean-squared prediction error RMSPE of a model μ with respect to the N_{VS} molecules m of the external validation sets (VS), where their predicted activities $Y^{\mu}(m)$ are directly compared to experimental values $A(m)$:

$$\mathrm{RMSPE}(\mu) = \sqrt{\frac{\sum_{m \in VS} [Y^{\mu}(m) - A(m)]^2}{N_{VS}}} \qquad (1)$$

This prediction error might, if desired, be compared to the

Table 2. List of the 13 Considered Data Sets[a]

ID	symbol	description and references	D	D-S	D-R	O	C	training set size	validation set size	no. of inactives
ACE	+	angiotensin converting enzyme inhibitors[21]	3062	1498	3421	3948	1683	106	38	-
AChE	×	acetylcholinesterase inhibitors[21]	1535	808	1619	1884	761	74	37	-
AT1	*	angiotensin type-1 receptor activators[22]	1971	1131	1900	2490	948	122	122	-
AT2	□	angiotensin type-2 receptor activators[22]	1971	1131	1900	2490	948	122	122	-
Art	■	artemisinin analogues[23]	1492	803	1685	1697	692	142	37	-
BZR	○	benzodiazepine receptor inhibitors[21]	1491	674	1955	1195	482	98	49	16
Cox2	●	cyclooxygenase-2 inhibitors[21]	1479	653	1645	1518	627	188	94	40
DhfR	△	dihydrofolate reductase inhibitors[21]	2380	1447	1807	2705	880	237	124	36
GPB	▲	glycogen Phosphorylase B inhibitors[21]	1403	664	1462	1144	427	44	22	-
FXa	▽	factor Xa inhibitors[13]	3642	2487	3639	5822	2333	290	145	-
Ster	▼	original CoMFA steroids data set[24]	907	382	907	849	362	21	10	-
Ther	◊	thermolysin inhibitors[21]	2942	1693	3016	3718	1497	51	25	-
Thr	♦	thrombin inhibitors[21]	2790	1498	2950	3475	1508	59	29	-

[a] Featuring their ID used in this work, associated symbols used in plots such as Figure 2, a brief description, and referencing plus the total number of pharmacophore triplets populated in at least one of the molecules, depending on the fingerprint version as defined in Table 1.

variance of the experimental property witnessed within the validation set, to calculate the validation set correlation coefficient R^2_V:

$$R^2_V(\mu) = \max\left(0{,}1 - \frac{\sum_{m\in VS}[Y^\mu(m) - A(m)]^2}{\sum_{m\in VS}[\langle A\rangle_{VS} - A(m)]^2}\right) \quad (2)$$

As the denominator in eq 2 simply serves to provide an order of magnitude to serve as a reference for the sum of squared residuals, its actual choice may vary from author to author: some use the average over learning set molecules $<A>_{TS}$ rather than the one over validation set compounds (which should make no difference if the VS contains a representative sample of the entire data set—but this is not always the case with the herein adopted compound series and splitting schemes). Also, certain authors report R^2_V values after linearly refitting predicted to experimental values. This amounts to accepting a model if its predictions obey an arbitrary linear relationship to the experiment ($A \approx \alpha Y + \beta$), rather than expecting predictions to equal actual values. Therefore, validation correlation coefficients as reported here may, unlike the average prediction error, *not be comparable to literature values.*

R^2_V is truncated at 0, signaling that any model with prediction errors larger than the ones of a 'null' model will simply count as failing to validate. This only affects benchmarks monitoring average validation propensities over the representative sets of SQS models μ, $<R^2_V(\mu)>_\mu$. Sets of SQS models including few equations that fail to validate with strongly negative R^2_V untruncated values should not be overtly penalized with respect to sets of equations with many but unspectacular validation failures.[25] The benchmarking studies, monitoring the impact of pK_a-dependent pharmacophore flagging (or fuzzy mapping) on the average validation propensities of models $<R^2_V(\mu)>_\mu$ rely on a 'minimal guaranteed shift' criterion δ_R (or δ_S, respectively), which expresses the drift of averages obtained with different descriptor versions ($<R^2_V(\mu)>_{(D-based\ \mu)}$ vs $<R^2_V(\mu)>_{(D-R-based\ \mu)}$ and $<R^2_V(\mu)>_{(D-S-based\ \mu)}$, respectively), corrected by the amount of drift that may affect these averages due to imperfect SQS sampling[25]—see eq 8 in that publication. The larger δ_R (or δ_S, respectively), the more significant the advantage of using pK_a-dependent pharmacophore flagging (or fuzzy triplet mapping, respectively).

Some literature studies also provided classification scores with respect to external subsets—either percentages of inactives correctly classified[21] as such or the percentage of correctly classified molecules[13] (actives as actives and inactives as inactives, respectively). In either case, these criteria were recalculated following the original author's procedures—please refer to cited papers for details.

2.4. Experimental Data. Table 2 shows the considered data sets, with their IDs in the present work, the references to the publications[13,21,22,23,24] reporting the previous QSAR studies, the numbers of populated triplets for each 2D-FPT version, and the set sizes. Please refer to the original publications and the Supporting Information for compound set sizes and training/validation set definitions. Except for the artemisinin[23] set, where the explained variable is a global score of antimalarial activity, all the considered studies refer to in vitro binding tests, the explained variables being in all cases dose-dependent indicators of inhibitory potency (IC_{50}, K_i) on a logarithmic scale. No metabolism-related or pharmacokinetic properties were included, as pharmacophore descriptors are primarily aimed at explaining the affinity of reversible noncovalent site/ligand interactions, while fragment descriptors are better suited to capture reactivity-related properties. Also, although 2D-FPT contain all the chemical information needed to estimate physicochemical properties such as the lipophilic character and derived indices (LogP/LogD, solubility, permeability, etc.) they may be too fine-grained in this respect. Global descriptors such as the total polar surface area may be more useful to predict LogP, rather than allowing for all the possible triplets including polar features to enter a very long and therefore statistically less robust QSAR equation. However, 2D-FPT may prove very helpful to pinpoint specific effects (such as the impact of intramolecular hydrogen bond formation on LogP) in completion to overall polarity indices—the study of possible synergies of 2D-FPT with other categories of descriptors is beyond the purpose of this work.

2.5. Extrapolability and Structural Interpretation of 2D-FPT-Based Models. All the representative thrombin (Thr) inhibition models using default 2D-FPT descriptors were challenged to predict the affinity of two chemically

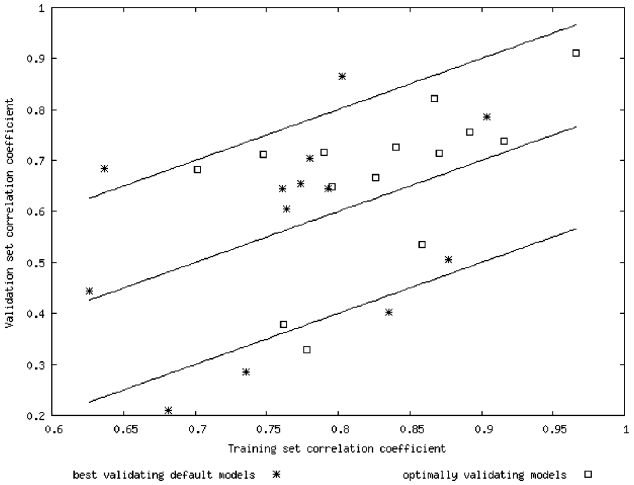

Figure 1. Comparative plot of training set (TS; R^2_T on X) vs validation set (VS; R^2_V on Y) correlation coefficients, for the globally optimal (row 1, Table 3) and default linear QSAR models (row 3, Table 3) of highest VS correlation coefficient, in context of grid lines $R^2_V = R^2_T$, $R^2_V = R^2_T - 0.2$, and $R^2_V = R^2_T - 0.4$, respectively.

different cocrystallized amidine/guanidine derivatives (PDB[30] codes 1BHX and 1D4P), plus a recently published compound,[28] of a radically different chemical class (2BXT). Since none of the Thr-trained equations passed the test—not even with respect to the amidine/guanidine derivatives—a novel series of models was refitted, after enrichment of the Thr training/validation series with presumed Thr inactives taken from 11 other compound sets (excluding FXa). The key pharmacophore triplets of these equations were traced back to their source atoms in the ligands, in order to check whether these include actual ligand-site anchoring points.

3. RESULTS AND DISCUSSIONS

3.1. 2D-FPT-Based Models Compare Favorably with Respect to Published QSARs of Higher Cost/Complexity. Prior to focusing on predictive success of validation sets—a necessary but not sufficient condition for any QSAR equation meant to serve for actual virtual screening of compound databases—Figure 1 provides, for each compound set, a concise outlook of the relationship between training and validation correlation coefficients, for the linear default and respectively optimal (top R^2_V) models. It is no surprise that no R^2_V -R^2_T correlation can be seen across multiple compound sets: while R^2_T values may to some extent relate to cross-validation scores (results not shown), correlations between training and validation scores are, even within a family of models based on a same compound set, rather rare (the "Kubinyi paradox"[31]). The reason for showing Figure 1 is to confirm that none of these models, selected due to their high R^2_V values, fail to apply to training set compounds. In principle, such models—artifacts 'explaining' the validation set by pure chance but unable to properly accommodate all the training examples—could be visited by the SQS procedure during its random walk in QSAR problem space. However, these are unlikely to enter the set of representative models regrouping only the most successful (training set) cross-validators—indeed, no situation with $R^2_V \gg R^2_T$ could be evidenced. Reversely, few models have $R^2_V \ll R^2_T$: the question whether these are 'overfitting' artifacts or whether some validation set compounds fall outside the applicability range granted by the training set will not directly addressed here. However, equivalently large R^2_V-R^2_T gaps are reported in the literature for the concerned compound sets: Cox2, GPB, and Ster (with the thrombin inhibitor set, Thr, only the default linear model displays large training-validation discrepancies).

Scrambling tests are routinely performed by the SQS approach: after termination due to failure to retrieve new fit models, the 10 best performing sets of operational parameters are used to pilot 10 independent attempts to build models on hand of Y-scrambled training set data (each of the 10 attempts relied on a different Y randomization). These attempts go beyond refitting of previously found equations and imply descriptor (re)selection and full-blown cross-validation. Typically, scrambling results met expectations: cross-validated Q^2 values were low (below 0.4) in most of the cases. For some compound set/descriptor version combinations, model fitting against unscrambled data failed to reach Q^2 values above 0.4—in these situations, there is significant overlap of the Q^2 ranges of scrambled and actual models. This is uninteresting—those cases would have been judged to represent QSAR buildup failures anyway, on the sheer basis of their low Q^2. Interestingly, some quite high scrambled Q^2 of up to 0.6 were obtained for the thrombin and steroids series, irrespectively of the employed descriptor versions. This is a critical warning signal about the extremely low intrinsic diversity of the sets: scrambling lead to swapping of activity values between molecules that are similar enough to 'stand' for others. Nevertheless, the top Q^2 scores with proper data exceeded 0.8 in all of these cases—therefore, the representative models discussed here, within a window of 0.2 Q^2 units, are all outside the Q^2 range covered by scrambling experiments. Under these circumstances, benchmarking may safely be based solely on validation criteria.

Table 3 shows that, with the notable exception of artemisinin analogues, 2D-FPT-based models were found (row 1) to equal or even significantly outperform the best validating published approaches (row 4; relative RMSPE shift in row 9—positive values standing in favor of 2D-FPT). [Albeit the correct classification rate of the FXa linear regression model is slightly lower than reported in literature, the former displays an excellent linear correlation score—which is a more constraining indicator of model quality than a classification rate. The reported GRIND-based discriminant model and the 2D-FPT linear equation are, as far as they can be compared, equipotent predictors.] This proves that 2D-FPT appropriately capture the structural information relative to reversible noncovalent binding to receptor sites and that the SQS methodology successfully mines for properly validating models. It is however inappropriate to claim that 2D-FPT are intrinsically more informative than CoMFA fields, although, for example, FPT-based results do outperform CoMFA even with the rigid and easy-to-align steroid (Ster), when the classical CoMFA drawbacks (uncertainties concerning the relevant conformations and alignment modes, etc.) are of little concern. Observed advantages may alternatively be explained by enhanced model sampling due to the parallelized, computer-intensive SQS approach, which might perhaps have found even better validating approaches if allowed to mine CoMFA field descriptors. This

Table 3. Benchmarking with Respect to Literature Results[a]

	ACE	AChE	AT1	AT2	Art	BZR	Cox2	DhfR	GPB	FXa	Ster	Ther	Thr
1	1.14	0.69	0.33	0.39	0.78	0.70	1.08	0.77	0.69	0.80	0.42	1.33	0.56
	0.713	*0.714*	*0.727*	*0.910*	*0.756*	*0.378*	*0.329*	*0.683*	*0.667*	*0.821*	*0.536*	*0.649*	*0.737*
	-	-	-	-	-	**81%**	**75%**	**69%**	-	**85%(a)**	-	-	-
2	D-R	D	C	O	D	D-R	O	D-S	O	D-R	D	C	O
	(L)	(N)	(N)	(N)	(N)	(L)	(N)	(N)	(L)	(L)	(N)	(L)	(N)
3	1.33	0.76	0.35	0.48	0.88	0.75	1.18	0.81	0.90	0.88	0.43	1.34	0.85
	0.605	*0.655*	*0.705*	*0.865*	*0.685*	*0.286*	*0.209*	*0.644*	*0.444*	*0.785*	*0.506*	*0.645*	*0.402*
	-	-	-	-	-	**75%**	**68%**	**92%**	-	**84%(a)**	-	-	-
4	1.48	0.95	0.42	0.51	0.70	0.87	1.17	0.84	0.79	-	0.69(c)	1.59	0.69
	-	-	-	-	-	**88%**	**70%**	**92%**	-	**88%(a)**	-	-	-
5	CoMSIA basic	CoMFA	CoMFA	CoMFA	NN(b)	2.5D	CoMSIA extra	HQSAR	CoMSIA extra	GRIND-PLS	CoMFA	CoMFA	CoMSIA basic
6	1.50	1.20	-	-	0.78(b)	0.87	1.25	0.99	1.20	-	-	2.24	0.96
						88%	**70%**	**81%**					
7	**10.1**	**16.8**	**16.7**	**5.9**	**−25.7**	**13.8**	**−0.9**	**3.6**	**−13.9**	**-**	**33.3**	**15.7**	**−23.2**
8	**11.3**	**36.7**	**-**	**-**	**−12.8**	**13.8**	**5.6**	**18.2**	**25.0**	**-**	**-**	**40.2**	**11.4**
9	**23.0**	**27.4**	**21.4**	**23.5**	**−11.4**	**17.5**	**7.7**	**8.3**	**12.7**	**-**	**39.1**	**16.4**	**18.8**

[a] **1** − Validation criteria for the globally optimal, best validating 2D-FPT models: RMSPE (plain text), validation set R^2_V (italics), and percentage of correctly classified inactives (bold) in an additional inactive validation set, except for (a), reporting the overall correct classification rate of both validation set actives and inactives. **2** − 2D-FPT setup and nonlinearity policy (linear, nonlinear) leading to results (1). **3** − Validation criteria, as in 1, of best validating linear model based on default 2D-FPT. **4** − validation criteria (RMSPE, correct classification rates) of most successful models reported in the literature (references in Table 1); RMSPE value (c) not reported as such, was calculated on hand of data reported in Table 2 of that publication.[24] **5** − Methodology leading to models (4). **6** − Validation criteria (as in 4) for literature models of comparable complexity to 2D-FPT equations. Except for artemisinin (b), reporting a linear model based on 2D and 3D descriptors, this row presents 2.5D descriptor-based models.[21] **7, 8, and 9** − relative prediction error decrease of 2D-FPT vs reported models: default 2D-FPT vs best reported, e.g., row 7 = ($RMSPE_4$-$RMSPE_3$)/$RMSPE_4$ (%), default 2D-FPT vs comparable reported (row 8: 3 vs 6) and best 2D-FPT vs best reported (row 9: 1 vs 4), respectively.

notwithstanding, 2D-FPT are clearly able to generate state-of-the-art QSAR models in conjunction with powerful model building procedures. Furthermore, previous results[25] actually showed that, in many cases, valid 2D-FPT models may well be obtained by less aggressive techniques, such as stepwise regression. Computationally effective topological and alignment-independent 2D-FPT have thus significant technical advantages over CoMFA.

Nonlinear approaches occupy the top position of the best validating model in eight out of 13 cases, an observation reinforcing the already reported[25] trend of improving model validation propensity upon allowing SQS to employ predefined nonlinear transformations.

Linear models using the default fingerprint version (row 3) never happened to represent the globally best validating approach. Their performances still equal or exceed the best literature values (row 4, relative shifts in row 7) in ten out of the 13 studied cases. In two situations, default linear models fail to meet literature standards, although better-than-literature 2D-FPT models could be found using other setups and/or nonlinearity policies. In the case of thrombin inhibitors Thr, the top linear model based on the O version performs only slightly better than the default (RMSPE of 0.82 instead of 0.85)—the dramatic drop to the global optimum at 0.56 is a specific consequence of nonlinearity. For glycogen phosphatase B inhibitors GPB, the globally optimal model is linear as well—see the next paragraph for a discussion of D- and O-version GPB equations.

Benchmarking of (row 3) default 2D-FPT equations against literature models of comparable complexity—i.e. linear, overlay-independent models not requesting any buildup of molecular geometries—found these latter (row 6, relative shifts in 8) to be outperformed in eight out of nine cases. Except for the artemisinin analogue series, where row 6 refers to a linear model based on 2D descriptors, other row 6 equations are based on '2.5D' indices. According to the original paper,[21] these models, using a mixture of standard 2D and 3D descriptors, outperform equations solely based on 2D indices. However, since the involved 3D descriptors are whole-molecule indices (such as molecular volume and surface values, replaceable by quick estimators based only on molecular connectivity), the 2.5D models were considered to be acceptable matches to 2D-FPT equations in terms of complexity.

The lesser performance of 2D-FPT with respect to the artemisinin series is actually not surprising. In this case, the monitored activity is an overall, systemic antimalarial potency score, normalized by molecular weight. These compounds are peroxides and act as heme alkylating agents,[32] not in reversibly binding to receptors. Fragment and/or topological descriptors are expected to (and actually do, according to the literature model[23]) better explain such a type of activity. 2D-FPT nevertheless come up with some reasonable models, which does not imply that some kind of pharmacophore recognition is required for the alkylating activity. More likely, the presence of certain pharmacophore triplets may correlate with specific substructures and therefore implicitly relate to reactivity (also see discussions below).

3.2. 2D-FPT Setup-Dependence of the Validation Performance. Within this and the following chapters, the validation set correlation score, systematically calculated according to eq 2, is used for comparison—either in terms of top R^2_V of the best validating models of the respective representative sets or in terms of average validation propensity $<R^2_V>$ over the representative sets of equations. In principle, due to the stochastic nature of the model builder, it is risky to extrapolate the intrinsic quality of descriptors from validation score differences of single models. However, the analysis of the 39 duplicated SQS simulations for all 13 data sets, performed with D, D-R, and D-S descriptors, respectively, showed that duplicate simulations generate significantly diverging representative sets[25] and different best validating models, which nevertheless have remarkably close R^2_V values. In 19 cases out of the 39 repeats (49%), the top R^2_V value was reproduced within 0.025, and in 25 cases

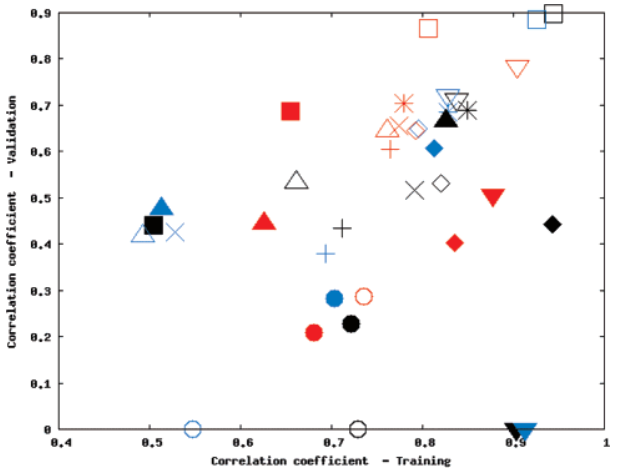

Figure 2. Training set (TS; R^2_T on X) vs validation set (VS; R^2_V on Y) correlation coefficients for the best validating linear models obtained for each compound set (see symbol coding in Table 2) and each descriptor setup version (red-D, black-O, blue-C).

(90%) the R^2_V shift did not exceed 0.05. The most irreproducible top R^2_V value, seen to shift by 0.2, concerned the steroid data set in conjunction with nonfuzzy descriptors (D-S). Since repeated simulations virtually never led to top R^2_V value differences above 0.1 units, this may be, in our opinion, taken as the significance threshold (shifts above 0.1, if observed, may be attributed to the differences in chemical information conveyed by each descriptor version).

The monitoring of the descriptor versions found at the basis of globally optimal equations (Table 3, row 1) shows a slight preference for the O setup (winner in 4 cases), followed by D (3), D-R (3), C (2), and finally D-S (1). Figure 2, representing the training and validation set R^2 values for every top validating linear model built at a given {compound set, descriptor version} combination, shows that the influence of 2D-FPT setup on resulting QSAR model performance is unfortunately not easy to foresee (compound sets are dot-shape coded, Table 2, while descriptor versions are color coded: D-red, O-black, C-blue).

3.2.1. Coarse FPT Are the Less Successful in QSAR. It appears that the coarse, less information-rich, descriptor version C is on the whole less well suited for QSAR modeling: in one case (Art ■) it actually failed to generate any useful models at all, while in two more cases (Bzr ○ and Ster ▼) all of the linear C models failed to validate. In all other cases, the blue mark corresponding to C-based models rarely tops the two others on the Y axis (R^2_V)—however, a remarkable exception was observed in the case of thrombin inhibitors (Thr ♦), where the C-based linear model only comes in second to the nonlinear O-based approach (not shown on the plot). No straightforward explanation of this unexpectedly good performance of the C version could be found. None of the C triplets entered either D or O models, and, furthermore, the C version model does not even include any specific triplet featuring the positive charge required for thrombin activity. This is, per se, not surprising, since all the Thr compounds, actives and inactives alike, include at least one protonable group. This training set does not emphasize the fact that the cationic center is important. It is thus likely that the C model exploits some local, family dependent chance correlation between C descriptors and activities. [In QSAR literature, a 'chance' correlation is said to apply within the training set but break down either at the cross-validation stage or, at latest, with respect to validation compounds. If, however, training and validation sets are subsets of the same structural family, 'chance' correlations that persist throughout training, cross-validation, *and* external validation may well exist—and explain the low success rate of QSAR models in actual virtual screening of diverse databases.] D and O models, however, feature at least one triplet involving the positive charge, i.e., implicitly suggesting that a cation flanked by specific groups at specific distances may play a role in binding. Such a conclusion may yet be an overinterpretation: it should not be forgotten that triplets are the only input options in this approach. Therefore, if the actually important element were a pharmacophore *pair*, not a triplet, selecting a triplet involving that pair plus the ubiquitous cation is merely the workaround found by the approach to compensate for a missing explicit pair of descriptors (also see the paragraph dedicated to structural interpretation of Thr models, further below).

3.2.2. O-Version Failures and Successes: Why FPT May, within a Structurally Homogeneous Family, Implicitly Behave like Fragment Descriptors. Like the C fingerprints, the O-based 2D-FPT also failed to generate any properly validating models for the Bzr ○ and Ster ▼ sets.

Steroid Models. For the steroid set, the D-version linear model utilizes three triplets, one being favorable for activity: (1) HA4-HA10-Hp12 − two acceptors at 12 bonds apart, e.g., located at both ends of the steroid scaffold and a hydrophobe at 4 bonds from one acceptor and at 10 from the other, e.g., part of the scaffold (in triplet nomenclature,[1] each corner is followed by the length of the opposed edge, in number of bonds). Two other triplets were seen to be most often populated in inactives and decrease predicted activity when present in validation set compounds: (2) Ar4-HA4-HA4 − an equilateral triangle of edge lengths 4 consisting of an aromatic and two acceptors. (3) Ar10-Ar10-HD4 − two aromatic atoms, 4 bonds apart, at the opposite of the steroid scaffold (at 10 bonds) from a hydrogen bond donor.

All these triplets are also part of the O-version basis set, but O-version population levels slightly differ due to different fuzziness and aromatic/hydrophobic equivalence parameters. The levels of Ar10-Ar10-HD4 are particularly low and only come from imperfect mapping of molecular triplets where the aromatic feature is down-weighted because it is actually represented by a hydrophobe. As 2D-FPT are integer value vectors, and given the overall poor match of actual molecular triplets, the actual population level of Ar10-Ar10-HD4 rounds up to either 0 or 1 (out of the 50 arbitrary units standing for a perfect match). Or, with the D setup, fuzziness and aromatic-hydrophobic interchangeability are defined such that the Ar10-Ar10-HD4 population level happens to be correlated with the presence of a hydroxyl group at position 3 of the A ring. All the 3-OH steroids have Ar10-Ar10-HD4 set to 1, and all but one of compounds with Ar10-Ar10-HD4 equaling 1 are 3-OH steroids. 3-OH steroids are inactive—their average activity (alcohols or phenols confounded) is 1.8 log units below the average over the rest of the molecules. When using the O or C setups, however, the privileged relationship between the 3-OH fragment and the particular 2D-FPT element breaks down, with immediate negative impact on the model statistics. The D model,

Table 4. Benchmarking of the Impact of Descriptor Fuzziness and pK_a-Dependence on the Quality of Resulting QSAR Models[a]

	best R^2_V			$<R^2_V>$ and (variance)			guaranteed shift δ due to	
set	D	D-S	D-R	D	D-S	D-R	fuzziness	pK_a- dependence
ACE	0.605	0.526	**0.713**	0.260 (0.110)	0.200 (0.119)	0.324 (0.131)	+0.049	−0.043
AChE	0.655	0.627	0.658	0.051 (0.104)	0.116 (0.160)	0.210 (0.178)	0.000	**−0.144**
AT1	0.705	0.671	0.718	0.554 (0.100)	0.478 (0.115)	0.489 (0.140)	+0.026	+0.024
AT2	0.867	0.873	0.884	0.744 (0.061)	0.760 (0.066)	0.754 (0.071)	−0.006	−0.002
Art	0.688	0.736	0.742	0.466 (0.107)	0.495 (0.156)	0.549 (0.083)	0.000	−0.060
BZR	0.286	0.214	**0.378**	0.009 (0.034)	0.004 (0.021)	0.023 (0.051)	+0.003	0.000
Cox2	0.209	0.247	0.171	0.012 (0.029)	0.029 (0.044)	0.006 (0.018)	−0.007	0.000
DhfR	0.644	0.670	0.590	0.172 (0.142)	0.364 (0.166)	0.173 (0.151)	**−0.139**	0.000
GPB	0.444	0.498	0.426	0.033 (0.068)	0.109 (0.098)	0.018 (0.065)	−0.029	0.000
FXa	0.785	0.819	0.841	0.639 (0.071)	0.706 (0.062)	0.689 (0.070)	−0.055	0.037
Ster	0.506	0.345	0.457	0.003 (0.037)	0.001 (0.016)	0.001 (0.022)	0.000	0.000
Ther	**0.645**	0.623	0.439	0.321 (0.154)	0.212 (0.143)	0.007 (0.039)	**+0.098**	**+0.307**
Thr	0.402	0.437	0.375	0.093 (0.108)	0.107 (0.142)	0.050 (0.070)	0.000	0.000

[a] In terms of both optimal and average validation propensities (R^2_V) of the linear models from the representative SQS sets.

although by any standards much better than reported CoMFA-based QSARs, is yet another example[11] of how QSAR may provide correct predictions based on wrong premises. Since the population level of Ar10-Ar10-HD4 is determined by the −OH group at carbon 3, there is no reason to imply that the aromatic corners of the triplet must be mechanistically involved in modulation of the affinity.

Benzodiazepine Receptor Inhibitor Models. Benzodiazepine receptor inhibitor models—including those from literature—all have low validation propensities. In this context, the deceiving behavior of O- and C-based models is not surprising. Since the D-based best validating linear model includes 14 different triplets, tracing the differences between D and O models back to the subjacent descriptors is a difficult task. The fact that several of the triplets entering the D model have edge lengths of two may be the first hint toward a possible explanation: with minimal edge lengths E_{min} of 4 and 5, respectively, neither the O nor C versions may account for such short-range pharmacophore elements.

Glycogen Phosphorylase B Models. GPB offers a counterexample where the O version is the most successful. In this case, all the triplets entering the O model also happen to be members of the D basis set. However, the O fingerprints are less fuzzy than their D counterparts, especially with respect to hydrophobic groups. Attempts to build relevant D models with the triplets entering the O-based top validating equation failed. Furthermore, the best validating linear D-S model, using unfuzzy triplets (see Table 4), performed slightly better than D but worse than the O-based equations. Apparently, the quality of the GPB QSAR models displays a peak at some optimal triplet fuzziness level.

3.2.3. Influence of pK_a-Dependence on QSAR Quality. Table 4 shows both the optimal R^2_V values and the average $<R^2_V>$ scores over the representative sets of linear models, from duplicate SQS runs using specified 2D-FPT versions. The reported guaranteed shifts are related to the respective (D-S vs D and D-R vs D) average score differences, conservatively corrected by the amount of average shift that might be attributable to $<R^2_V>$ score fluctuations.[25] Positive shift scores suggest the superiority of the default version with respect to rule-based and nonfuzzy approaches, respectively. Both top R^2_V values stand out by more than 0.1 units and guaranteed shifts exceeding 0.1 were highlighted.

Rule-based pharmacophore flagging lead to significantly better top models but not to significantly better average scores for the ACE and Bzr series. It also triggered a significant increase in the average validation propensity of AChE models, without however impacting on the quality of top equations. The only case where switching from D to D-R descriptors is seen to provoke a coherent and very large change of both optimal and average validation scores is the Ther compound set, with a net preference for pK_a-dependent D fingerprints. The following paragraphs suggest possible explanations for these observations:

Angiotensin Converting Enzyme Models. Within the ACE set, the top validating D-R equation includes five descriptors, compared to four entering the less well performing D top model. The essential difference is the participation of the Ar2-NC2-PC2 triplet in the former but not in the latter. This triplet is populated in α-amino acid moieties (with an actual hydrophobe replacing the aromatic): the D-R model learned that compounds including such moieties are, on the average, more active than others. This hypothesis finds itself confirmed by validation set compounds, of which all three (Figure 3a) that have populated Ar2-NC2-PC2 D-R triplets are nanomolar actives. The D-R Ar2-NC2-PC2 term contributes, for all three, an increment of +2.5 log units which is paramount for correct activity prediction. However, the D-R strategy ignores, by contrast to the pK_a-based approach, the existence of populated Ar2-NC2-PC2 in two additional, completely inactive, validation set compounds (Figure 3b). It rightly denies the cation status to the tertiary N atom, erroneously perceived as a quaternary pyridinium by the D flagging scheme, but wrongly ignores protonation of the tertiary amine in the second molecule. This latter is a technical problem that could be fixed by rewriting the default pharmacophore flagging rules precompiled by ChemAxon, which were used as such in this work (the aspect was already mentioned in the previous 2D-FPT paper[1]). Acknowledgment that the Ar2-NC2-PC2 triplet may stem from fragments other than α-amino acid moieties breaks down the correlation between Ar2-NC2-PC2 population levels and activity. Both D and D-R schemes each err once in the flagging of validation set compounds, but the D-R error leads to a happy coincidence, establishing a biased correlation between the pharmacophore triplet and a specific fragment. The number of models exploiting this artifact is however small compared to the set of relevant SQS equations—therefore, average validation propensities were not affected by switching from D to D-R.

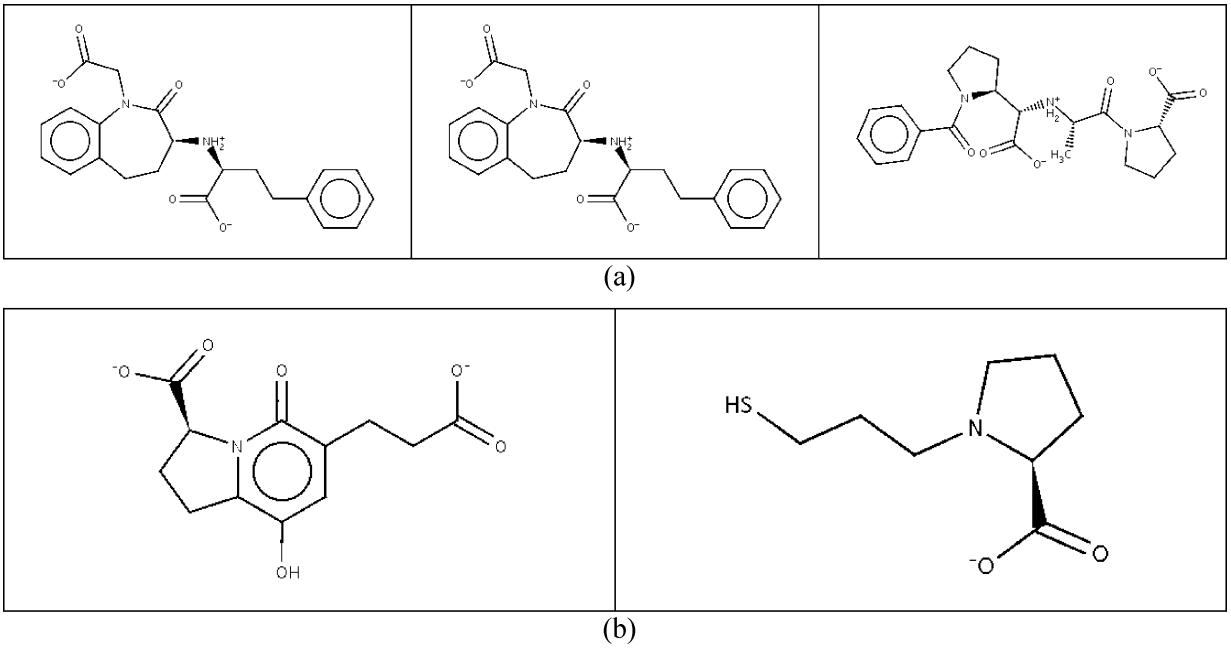

Figure 3. ACE compounds featuring the specific Ar2-NC2-PC2 triplet entering the top D-R model: (a) the three validation set compounds populating this triplet according to D-R are nanomolar binders and (b) the D (pK_a-sensitive) approach also finds the triplet in these two molecules, where it erroneously assumes a positive charge of the 'quaternary pyridinium' N and it correctly considers tertiary amines to be protonated at pH = 7.4. Both molecules (b) are inactive.

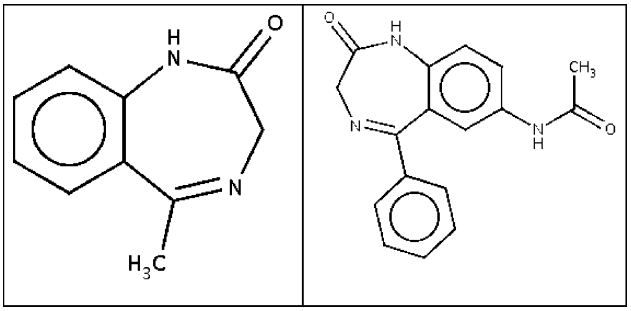

Figure 4. Typical Bzr set representatives featuring the imine moiety erroneously taken for an immonium cation by the D-R rule-based flagging strategy.

Benzodiazepine Receptor Inhibitor Models. In the case of Bzr inhibitors, the main difference between rule- and pK_a-based pharmacophore flagging concerns the imine nitrogen within the 7-membered ring, present in a majority of the compounds. The rule-based approach considers this N to be protonated because it possesses a free electron pair. This is actually not the case at pH = 7.4 (aliphatic imine pK_a values are about 4.0, with ChemAxon predicting values of 3.6 and 3.3 for the phenyl- and diphenylimine moieties in the Bzr compounds from Figure 4). In D fingerprints, this N is ignored, not being basic enough (pK_a cutoff of 5) to be flagged HA. This notwithstanding, imine fragments are nevertheless seen more often in actives than in inactives. Though it is impossible to state whether this relative enrichment is a set-specific accident or whether this fragment is mechanistically needed (electron density effects on conjugated phenyls, conformational constraints guaranteeing proper binding geometries) it makes sense to rely on the imine fragment count to explain activity trends within the set. As a consequence of the flagging error, in the D-R version imines are being assigned a special status (cations are rare features, so that they often stand out as the only 'cation' of the molecule). Therefore, the presence of the imine moiety is straightforwardly expressed by the population levels of specific PC-containing triplets. The better performance of the D-R approach in this case was again a lucky accident.

Acetylcholine Esterase Models. Unlike in the two above-mentioned situations witnessing accidental specific improvements of the top-validating D-R models, D-R based AChE models show an improvement of the average validation propensities, a trend not followed by top-validating models. In order to understand this phenomenon, the average prediction errors committed by each of the 1790 D and 2294 representative D-R models, respectively, were monitored for each of the 37 validation set compounds, in search for molecules that were systematically less well predicted by D approaches (Figure 5). It is important to note that in the AChe series the protonation states were explicitly provided in the input files—tertiary amines were protonated, forcing the D-R flagging scheme to recognize them as cations (if plain tertiary amines were input, D-R would have assigned[33] the hydrogen bond donor flag to the tertiary N, while the D flagging scheme is insensitive with respect to the actual protonation status of input compounds). The observed differences between D and D-R models is though not due to the treatment of tertiary amines by the latter. In-depth analysis pinpointed to another—related—flagging artifact of the D-R strategy: in fact, the above-mentioned peculiar flagging of trisubstituted N atoms equally (and erroneously) applies to N-disubsituted amides. Or, all four compounds in Figure 5 happen to belong to this category. The peculiar data set artifact allowing the chemically meaningless flagging of tertiary amides as hydrogen bond donors to translate into more accurate predictions remains obscure.

Thermolysin Models. The thermolysin compound set (Ther), the only one to show a consistent amelioration of both top and average validation propensities when using the

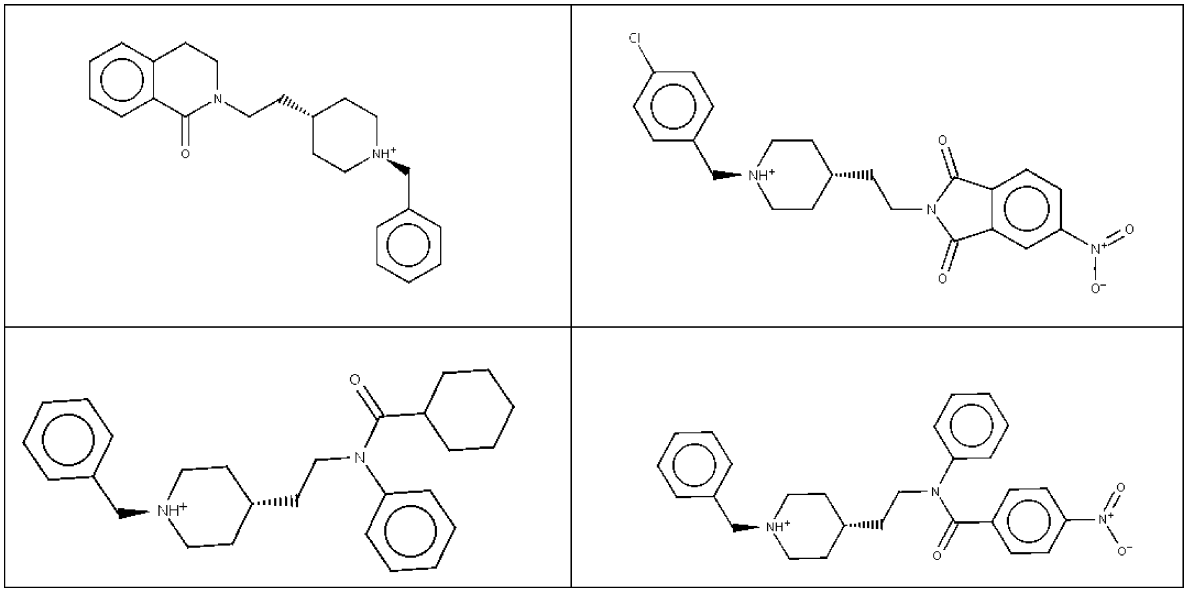

Figure 5. AChE validation set inhibitors for which the relevant D-R models provided, on the average, prediction errors smaller by one unit or more compared to the ones committed by D models.

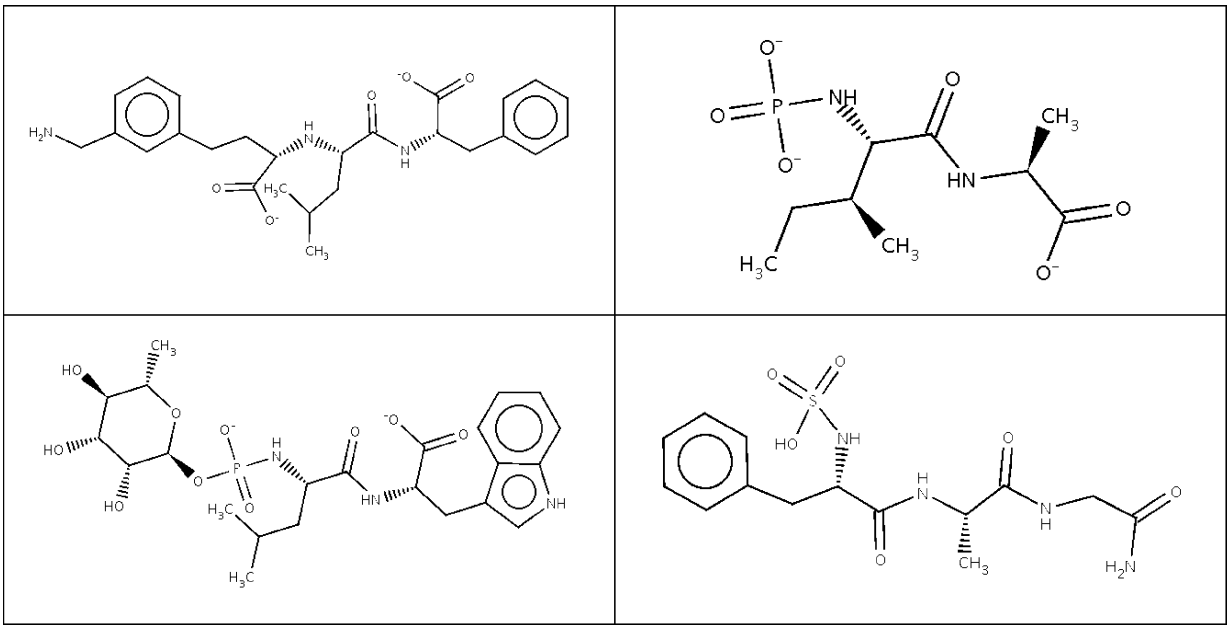

Figure 6. Thermolysin (Ther) validation set inhibitors having their activities properly predicted by the top D approach but highly overestimated if the same top D model is used with D-R descriptors.

pK_a-dependent flagging scheme, is also a series of outstanding structural diversity. Multiple, both acid and basic ionizable groups—some as atypical as thiophosphates, not perceived as anions by the D-R approach—are often seen in these compounds. Under these circumstances, the clear positive impact of a pK_a-dependent approach should not come as a surprise. The top validating D linear model is actually a quite simple equation, involving only four triplets, out of which three have positive coefficients (specifically populated in actives: Ar2-Hp4-Hp4, HA6-HD8-Hp8, and HD4-Hp6-NC4) and one with a negative coefficient (preferentially seen in inactives: Ar10-Ar10-HD6). In order to pinpoint key fingerprint differences upon switching to the D-R version, the validation set activities were also calculated according to the top D model but using D-R population triplet levels. For 10 out of 25 validation compounds, D and D-R population levels were identical, and so were predictions. In 8 cases, however, this led to a significant overestimation of activities (by 1 log unit or more). Figure 6 illustrates four of the concerned examples. For example, in the first represented compound, D-R population levels of both Ar2-Hp4-Hp4 and HD4-Hp6-NC4 were much higher than the corresponding D versions and thus triggered an activity overestimation of ∼5 log units. The explanation, however, is not in any way related to different protonation patterns (both D and D-R consider the two carboxylates as deprotonated and the primary and secondary amines as protonated) but to a peculiar difference in flagging strategy. While the D-R approach considers the anionic flag on the negatively charged oxygen, the D strategy assigns it to the carboxylate C, instead of the default hydrophobic flag (oxygens are flagged HDA). The D-R population level of HD4-Hp6-NC4

triplets increases because the contributing atom triplets both have the HD-NC and Hp-NC edges longer by one (the carboxylate C−O⁻) bond and are therefore a better match for the basis triplet. The Ar2-Hp4-Hp4 levels increase because the D-R version sets additional hydrophobes—the carboxylate C atoms. This is an example of 2D-FPT degeneracy impacting on QSAR propensity: if an atom triplet is responsible for activity, without being unique in the molecule (the others playing no role), then the key triplet will represent only a fraction of the total population level of the matching fingerprint element. This population level may, per se, not discriminate between deletion of the key triplet and deletion of an irrelevant contributor—other triplets, specifically designing the key atoms and their environment, must be taken into account. Setting hydrophobic flags on carboxylates caused 'drowning' of the relevant contributions to the Ar2-Hp4-Hp4 population level in noise from the additional, meaningless triplets. Of course, the fitting of specific D-R models, compensating for these flagging differences, lead to equations in which the molecules in Figure 6 are being better predicted than on hand of the D model using D-R fingerprints. This compensation is however incomplete—prediction by the D model with appropriate D fingerprints is still better. Also, the top validating D-R model requires a total of six different triplets, compared to only four for D. Although the D strategy is in this case the clearly better one, its advantages do not stem from capturing any subtle protonation effects but are more likely from the (chemically meaningful) deletion of the hydrophobic character of carboxylate C atoms.

All in all, this work did not produce any clear evidence that pK_a-dependent flagging may enhance descriptor performances in QSAR: if such evidence exists, it was unfortunately hidden by noise due to the peculiarities of the compound sets and by the other, unavoidable, flagging scheme differences. All the situations in which the rule-based approach appeared to perform better have been traced down to 'lucky' coincidences, where chemically meaningless flags happened to single out specific compound subfamilies, enriched in actives. On the opposite, in the single case where the D version brought clear improvements of both top and average validation propensities, benefits were due to pK_a-independent, albeit chemically meaningful, flagging differences. These findings apparently contradict the reported[1] importance of pK_a-sensitive flagging in evidencing NB violations ('activity cliffs'—structurally almost identical compound pairs with nonetheless differing activities). The problem is that differences between D and D-R flagging strategies are not strictly limited to proteolytic equilibria-related effects. In similarity scoring, however, systematic pK_a-unrelated flagging differences tend to cancel out: if A and A′ are, for example, two homologous carboxylic acids differing with respect to a single substituent, the dissimilarity score between A and A′ is largely independent of whether −COO⁻ carbons are both labeled as hydrophobes or both labeled as anions—they are just a common feature of both A and A′. If, however, the differing electronegativities of the varying substituents cause a shift of the −COOH ionization status in A vs A′, the difference is clearly reflected in the dissimilarity score. Things are different for QSAR: 'noise' from the allegedly hydrophobic carboxylate carbons happened to accumulate atop of an apparently relevant fingerprint element (Ar2-Hp4-Hp4), decreasing its propensity to enter QSARs and forcing the machine learning process to come up with alternative, less well performing models.

Pinpointing of specific pK_a-related effects in QSARs would have been possible if a top common model (or at least models including the same descriptors, with differing coefficients) would have been found for both D and D-R sets. Unfortunately, this was never the case. Trying to use D-R population levels in D models, or vice versa, always leads to significant prediction errors. Optimally validating models do not happen to differ solely because SQS fails to rediscover the same equation when run with the other descriptor set. Top validating models based on one fingerprint version were genuinely incompatible with other descriptors. Moreover, all attempts to fit D-R models with terms entering the top D equations, or vice versa, failed (results not shown). If the same top validating model would hold for both D and D-R series, with only predictions of proteolytic equilibrium-dependent compounds seen to vary in function of the flagging strategy, the direct impact of pK_a-dependence could have been monitored. In reality, switching from D to D-R prompts the SQS engine to come up with diverging sets of models, under the combined influence of both the pK_a-specific and nonspecific flagging differences that were highlighted above.

3.2.4. Influence of Fuzzy Mapping on QSAR Quality. The employment of fuzzy logic at the descriptor build-up stage has no significant impact on QSAR performance—at least not at the level of fuzziness proned by the D version—except for the DhfR inhibitor set. Here, fuzziness actually appears to be detrimental in terms of average validation propensity shifts, though it has no noteworthy impact on the top model quality. The D-S set, with a grid mesh $E_{step} = 2$, does not capture any information concerning atom triplets separated by an odd number of bonds. Fuzzy logic is mainly a tool to avoid triplets 'slipping' through the grid mesh defined by such a rarefied, smaller size triangle basis set and was shown to have a positive effect on the NB of 2D-FPT. However, it does not appear to be essential for QSAR model buildup. This makes sense if recalling that 2D-FPT fingerprints are highly redundant, in the sense that some triplet occurrences are necessarily correlated (if no positive charge is present, then all triplets featuring a PC will simultaneously have population levels of 0, etc.). Nevertheless, note that such interpopulation level correlations must not necessarily be of a linear nature: the more diverse and large the compound sets, the lesser the chance to find an even-edged triplet having its population level linearly correlated to one of the 'missed' odd-edge key triplets. If such correlations exist, the population level of the latter may thus implicitly account for one of the 'missed' triplets, throughout training and validation sets—very much in the same way in which a given triplet was shown to implicitly monitor the presence or absence of a single functional group. This is apparently the case within the DhfR inhibitor set. Unfortunately, the top validating models cannot reveal any more specific details of the problem, as they have similar validation propensities, and an in-depth analysis of all the relevant (and quite complex) D and D-S models, aimed at understanding the differing average behavior, is too cumbersome to undertake.

On the one hand, fuzziness plays an important role in mimicking the tolerance of certain receptors with respect to

varying spacer length between two key groups. There is however no straightforward way to detect such examples throughout the 13 data sets, although it is clear that compound sets with various small substituents around a large central scaffold (such as steroids) are not concerned. On the other hand, too much fuzziness will eventually lead to degenerated fingerprints: as less and less strict edge length matching criteria are imposed, more and more atom triplets—involved in binding or not—will get a chance to contribute an increment to the population level of the given basis triplet. Even with D-S fingerprints, there are chances to find more than one arrangement of three atoms having the required pharmacophore flags and topological distances to match the same basis triplet: if only one of these atom triplets is important for activity, its signal will be buried under the noise from the other fortuitous contributors. Fuzziness only worsens such pitfalls. The impact of fuzziness on QSAR performance is thus different from the impact on neighborhood behavior. In the latter case, considering a pair of close analogues A and A′ with a common scaffold, the fact that in the unfuzzy versions some atom triplets are ignored is not of paramount importance, as the ones slipping between the meshes of the grid will be roughly the same in A and A′. Also, degeneracy due to fuzziness lets contributions from equivalent triplets build up equivalent final population levels, i.e., it has no negative impact on similarity scoring. However, if A and A′ differ in terms of a centrally inserted $-CH_2-$ group between two moieties, then triplets specifically localized within each moiety will appear unchanged in the fingerprint, whereas the mapping of triplets featuring corners from both parts of the molecules may, in the absence of fuzzy logic, vary dramatically as even edge lengths become odd due to $-CH_2-$ insertion and vice versa. The dissimilarity of A and A′ is therefore at risk of being overestimated. In QSAR, however, triplet degeneracy is a serious problem, whereas the issue of varying long-range contributions with spacer length might be circumvented by letting the model simultaneously pick several correlated long-range triplets considering alternative corners adjacent to the actual porters of ligand-site interactions. Among these, some will be populated at odd and others at even spacer lengths—receptor tolerance with respect to varying spacer length may be mimicked without the explicit need for fuzzy logic.

3.3. Impact of Nonlinearity on QSAR Quality. As can be seen from Figure 7, the best validating nonlinear models outperform their linear counterparts in a majority of situations. The 59 cases represent compound set/fingerprint version (D, O, C, D-R, D-S) combinations for which both the best linear and the best nonlinear model scored $R^2_V > 0.1$. Out of these, in 36 situations the nonlinear approaches turned out to be more robust validators, sometimes (in 8 cases) by more than 0.1 R^2_V units. The most clear-cut improvements due to nonlinearity (>0.2 R^2_V units) are observed for thrombin models: with D descriptors, nonlinearity allows an improvement of R^2_V from 0.402 to 0.617, with D-R from 0.375 to 0.619, while with O descriptors a jump from 0.442 to 0.737 is observed. At the opposite, only three cases in which the introduction of nonlinearity triggers a decrease by 0.1 R^2_V units could be seen: the most clear-cut is observed for the Ster set and D-R fingerprints (from 0.457 to 0.319).

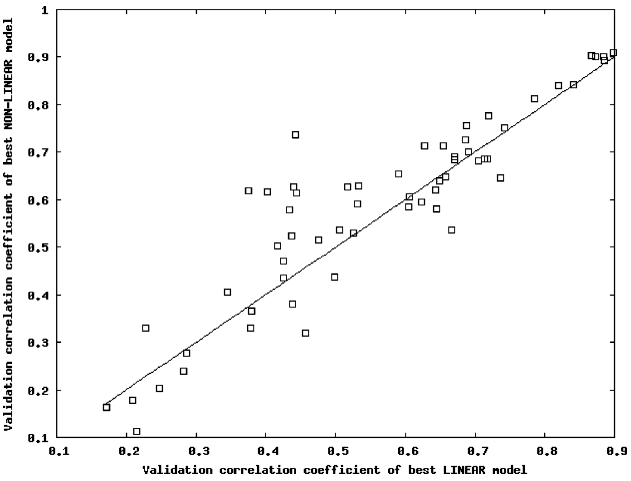

Figure 7. Comparative plots of R^2_V values scored, for each of the 13 compound sets, using each of the 5 descriptor versions, by the top validating nonlinear (on Y) and respectively linear (on X) equations (59 out of the $13 \times 5 = 65$ QSAR problems shown).

Nonlinear models are thus clearly better at extrapolating the knowledge extracted from the learning set to validation set molecules. These results reinforce the similar trend reported in earlier work.[25]

3.4. Beyond Validation: QSAR Extrapolability to Different Chemotypes. Successful validation is just a necessary but by no means sufficient guarantee of the actual usefulness of a model in virtual screening of random compound collections. Therefore, an in-depth assessment of QSAR models should, whenever possible, go beyond the simple comparison of R^2_V values. The first attempt to challenge the top validating nonlinear Thr QSAR model (D version) with predicting the activities of chemically different cocrystallized ligands (Figure 8) appeared quite promising at first sight: both compounds (b) and (c)—but not (d)—were predicted active. The latter, however, is known to adopt a different binding mode in the Thr active site—nothing in the training set could have hinted that such molecules may inhibit thrombin. Unfortunately, a closer look at the prediction showed that the high pK_i values predicted for (b) and (c) both stem from a single, very large contribution of the term 11.1×zexp3(HD4-Hp6-PC4). [zexp3(D) = exp[$-3(D-\langle D\rangle)^2/\sigma^2(D)$]—Please refer to Table 1 of the previous publication[25] for more details about the predefined nonlinear transformations in SQS.] Given the standard[25] average $\langle$HD4-Hp6-PC4$\rangle$ and variance σ(HD4-Hp6-PC4) population levels of 1.2 and 6.4, respectively, and knowing that HD4-Hp6-PC4 is not populated in either of the (b) and (c) molecules from Figure 8, the absence of such a triplet contributes 11.1×zexp3(0)=+10 to predicted pK_i values. This makes no sense—according to this model, any molecule without HD4-Hp6-PC4 triplets is a thrombin inhibitor (the considered 12-variable model does not contain any other negative potentially compensating contributions). Indeed, a quick verification confirmed that, according to this model—excellent training and validation statistics notwithstanding—all the compounds from the other sets used in this work should be nanomolar thrombin inhibitors. This is an artifact due to the low diversity of the training/validation set: the HD4-Hp6-PC4 triplet is populated in *all* training and *all* validation molecules, because it stems from the common amidine-phenylalanine moiety: the cation flag is set on the

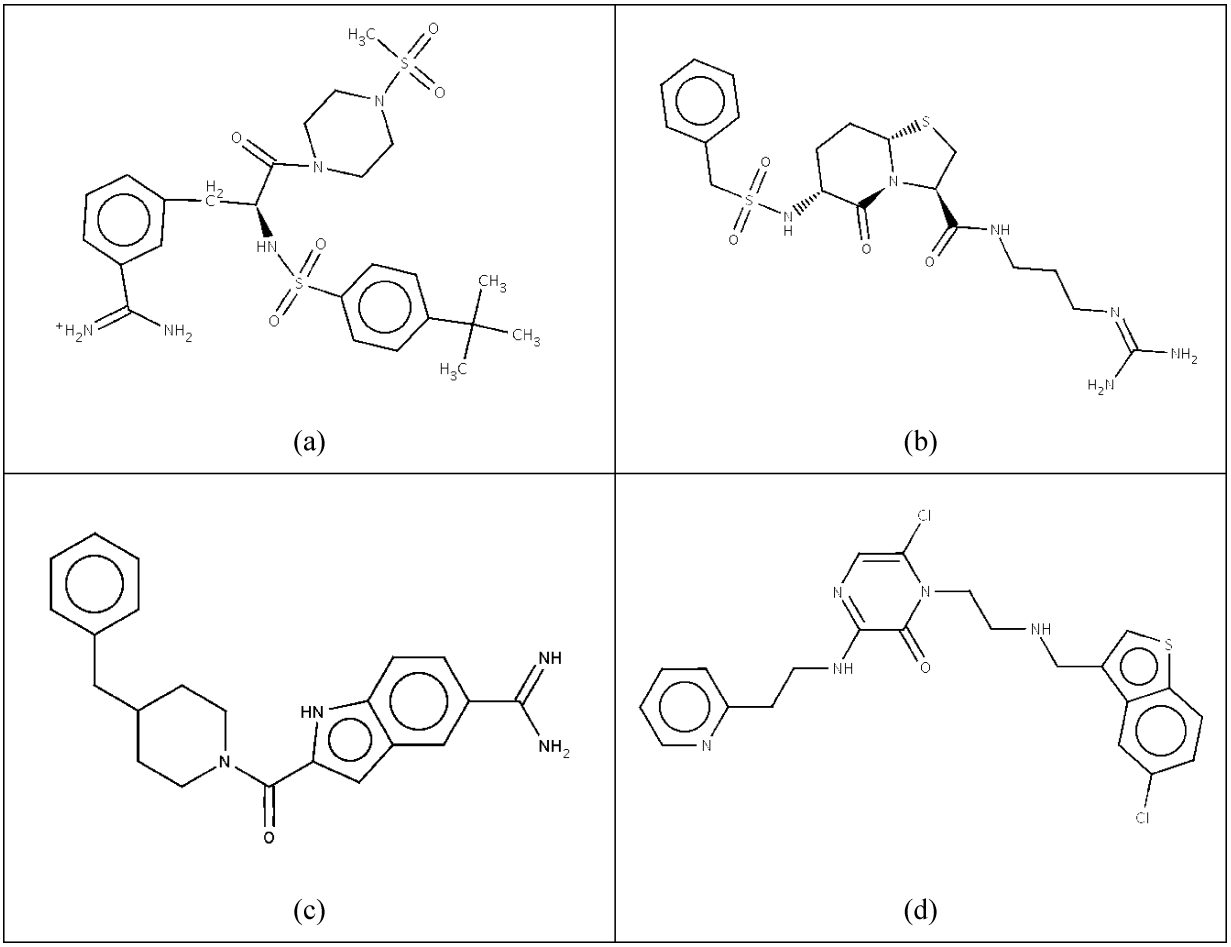

Figure 8. Part (a) is a typical thrombin inhibitor, featuring the amidine-phenylalanine scaffold characteristic of the Thr set, parts (b)[26] and (c)[27] are chemically different (cocrystallized) amidine/guanidine inhibitors, while part (d)[28] represents a radically new amidine-free class of ligands adopting a different binding mode.

amidine carbon, the donor is the phenylalanine >NH, at 6 bonds from the cation, while the phenyl ring carbon in *para* to the amidine, playing the role of the hydrophobe, is at 4 bonds from both PC and HD (other phenyl carbons also contribute, due to fuzzy mapping). If the phenylalanine carboxylate is coupled to a secondary amine, like in Figure 8(a), there are no other contributions to the population level of HD4-Hp6-PC4. The data set however contains a subset of primary amides: in this case, the HD flag of the CONH group is at 7 bonds from the PC and contributes to the HD4-Hp6-PC4 population. Primary amides have thus significantly higher HD4-Hp6-PC4 levels (i.e., lower zexp3 values), and, furthermore, they are on average significantly less active than secondary amides. Thus, zexp3(HD4-Hp6-PC4) entering the model with a large coefficient makes perfect sense in as far as the family of amidine-phenylalanines is concerned but is faulty outside this restricted applicability domain. Nonlinear models may indeed increase robustness of extrapolation from training to validation set but still do not offer guarantees of actual success in virtual screening. Improved validation set results might come at the price of a restricted applicability range or at least at the price of increased difficulty to properly define the applicability range in the presence of nonlinear terms.

None of properly validating ($R^2_V > 0.4$) representative nonlinear D models succeeded to specifically highlight (i.e., predict at submicromolar inhibition levels) (b) and (c) by contrast to randomly chosen inactives. Fortunately, 2D-FPT models are overlay-independent, which allows sets of arbitrarily high diversity to be used for training (training compounds need not have a common core, in order to be superimposable). Therefore, 125 compounds representing a randomly picked 10% of the other 11 data sets (FXa excluded), assumed to be inactive against thrombin (pK_i set to 4.0), were added to the initial Thr series. The resulting 'expanded' (ThrEx, 213 compounds) set was split into 169 training and 44 validation compounds and resubmitted to the SQS-driven nonlinear model buildup with D fingerprints. This time, the representative set of SQS equations featured two properly validating models, being both able to discriminate between Thr inhibitors and randomly picked compounds and to predict that (b) and (c) are submicromolar Thr inhibitors. Out of these two, one furthermore returned an excellent estimation of 50 nM for the affinity of (d) compared to the experimentally[28] reported 3 nM. This 12-variable nonlinear equation ($R^2_T = 0.864$, RMSPE = 0.73, $R^2_V = 0.762$) is given below, with *zQ(D:a:v)* denoting the predefined nonlinear functions[25] to be applied to descriptor *D* after its average/variance rescaling (*z*-transformation) with respect to the average value *a* and the variance value *v*, i.e. *zQ(D:a:v)=Q[(D-a)/v]*:

$$pK_i^{pred} = 0.07 \times HA8Hp6PC4 - 8.1 \times 10^{-4}\, Ar4Ar10HA10 - 0.57 \times HP10PC8PC10 - 2.1 \times 10^{-4}(HA12Hp12PC12)^2 - 0.16 \times (Hp6Hp6PC8)^2 + 0.3 \times \mathrm{zexp}(Ar10HP10PC6{:}2.3{:}13) - 0.45 \times \mathrm{zexp3}(Ar6Ar8Hp12{:}29.3{:}93.4) + 0.5 \times zsig3(Ar6HA2Hp6{:}119.1{:}156.8) - 0.5 \times \mathrm{zexp}(Ar8Ar10NC4{:}3.4{:}18.9) + 0.97 \times \mathrm{zexp3}(Ar8HA6Hp6{:}46.6{:}124.7) + 0.58 \times zsig(HA10HA12Hp4{:}18.8{:}77.9) + 3.46 \times \mathrm{zexp}(Ar6HA12NC10{:}0.2{:}2.0) \quad (3)$$

Equation 3 has the remarkable property to capture contributions not met in inactives but found in both active amidine-phenylalanine derivatives and (b), (c), or (d). It is however not the top validating nonlinear ThrEx model: this latter ($R^2_T = 0.901$, RMSPE = 0.32, $R^2_V = 0.956$) includes Thr-family specific contributions not shared by the structurally different inhibitors. As far as the machine learning process is left to focus only on the differences between actives and inactives within the Thr set, models exploiting all the idiosyncratic correlations due to the peculiar constitution of the data set perform well at training and cross-validation and will be selected. Some of these reveal themselves as meaningless when confronted to the diverse inactives of the extended set. Therefore, refitting with respect to the extended set leaves room for some less family specific, more general models to make it into the representative pool of equations as well.

It is also worth pointing out that out of 1113 distinct models—all of which boast outstanding training and validation criteria ($R^2_V > 0.7$)—only two stood up to the challenge of predicting compounds outside the training chemical family. In general, the QSAR problem is considered as solved if one well validating model has been found—what is the use of generating all these equally well performing 'redundant' models? The importance of aggressive QSAR problem space sampling resides in the fact that such 'redundant' models will cease to behave similarly when confronted to external molecules. The lower the informational content of the training set, the lower are the success expectations for any actual virtual screening based on thereon trained models, no matter how training is conducted. With stepwise/deterministic approaches, few equations—most likely all irrelevant—will be built. SQS may well enumerate relevant equations—but it will be impossible to guess which are the ones, unless an external test set can be used for further evaluations. The key advantage of SQS is that external sets may be too small to be useful at training (adding one or two external compounds to a homogeneous family does not help, with no cross-validation being possible) and yet allow for the discarding of most of the many thousands of sampled models, keeping only the ones that were not (yet) proved wrong. Classical QSAR buildup producing few equations is likely to end up with no models at all after confrontation with the external molecules.

3.5. Structural Interpretation of 2D-FPT Models—Do Topological Pharmacophores Make Sense? There is to our knowledge no direct experimental evidence of the binding mode of the amidine-phenylalanine derivatives of the Thr set. However, given the binding modes of related compounds

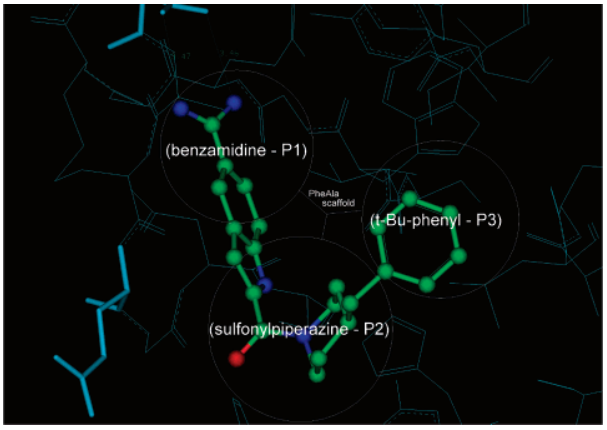

Figure 9. Thrombin active site with cocrystallized ligand—Figure 8(c)—and hypothesized binding mode of Thr set amidine-phenylalanine derivative from Figure 8(a).

and the overlay hypotheses standing at the basis of the original QSAR studies[34] concerning this family, it may be safely assumed that they would occupy all the three known binding pockets of thrombin. For example, compound (a) from Figure 8 would place the benzamidine moiety in P_1, the less hydrophobic sulfonylpiperazine substituent in P_2, and the t-Bu-phenyl group in P_3. Figure 9 illustrates this expected binding mode atop of the experimental bound geometry of compound (c). Clearly, compounds (a) and (c) are topologically different: in the former, the substituents filling the pockets feature a 'star' topology $P_1(-P_2)P_3$ centered on the phenylalanine α carbon, whereas in (c)—as well as in (b)—this arrangement is linear: $P_1-P_2-P_3$. Compounds like (b) or (c) have to adopt a U-shape geometry to close their P_1 and P_3 moieties up. This is a challenge to 2D-FPT-based models, since (a) and (b,c) do not share any topological triplets spanning all the three pocket-filling moieties: the P_1-P_3 topological distance in (b) or (c) is much larger than in (a). However, 3D-distance based common pharmacophore triangles might be found if the proper U-shaped fold is considered—should it be thus concluded that topological pharmacophore-based models prove unable to perform 'lead-hopping' from the star topology of the Thr series to the linear arrangement of the alternative ligands (b) and (c). Obviously not, since eq 3 applies to both of these topologies.

The high predicted pK_i values for the compounds in Figure 8 mainly stem from three main contributions. The highest one, an increment of +3.4 due to 3.46 × *z* exp(*Ar*6*HA*12*NC*10: 0.2:2.0) is constant for all four molecules, since none includes any negative charge. The term signals that compounds featuring such a triplet are not likely to be active—a 'lesson' learned from the additional inactives entering the ThrEx set. This makes sense insofar as thrombin clearly prefers cationic compounds. However, a negative charge will perhaps be detrimental even if it is not a part of this peculiar triplet chosen here.

Next, the Gaussian function of Ar8-HA6-Hp6 contributes with 0.6 to 0.9 pK_i units—the largest contribution seen in (a), where the triplet is represented once (fuzzy population level of 63), while the lowest occur if the triplet is not populated—in (c) and (d). Given the large variance of this triplet population level within the set of representative drugs used for 2D-FPT calibration,[1] this term may play an

important (activity-detrimental) role only in molecules containing several such triangles.

Insofar, the prediction that compounds (a)−(d) are active was only based on the fact that they are free of unwanted features, seemingly causing an affinity loss. Other contributions are, with one key exception, quite small (less than ± 0.5 pK_i units) and tend to cancel out. The remaining term is of paramount importance, based on a triplet actively favoring activity (HA8-Hp6-PC4). Figure 10 exemplifies the actual occurrences of this triplet in the molecules (recall that there are other atom triplets contributing, besides the highlighted ones—notably the ones including the symmetrically situated C atom in piperazine/cyclohexylamine rings). The triplet highlights two essential elements of the actual thrombin binding pharmacophore: the cation (amidine) interacting with Asp 189 from P_1 and the P_2 hydrophobic moiety 'sandwiched' between Trp 60 and Tyr 83. As the P_1 and P_2 binding moieties are topologically close in both (a) and (b/c), this particular triplet ensures model extrapolability from one topological family to the other.

Intriguingly, there is no role in binding directly attributable to the hydrogen accepting carbonyl of the triplet. However, this carbonyl is nevertheless 'important'—not structurally, but chemically, for synthesis reasons. As acylation is a preferred building block coupling reaction, it is not astonishing to see a conserved carbonyl throughout diverse series of compounds that were conceived as timers matching the three thrombin binding pockets. This is a nice example showing that QSARs will never represent absolute training-set independent laws, as training sets will always be biased, be it only for chemical feasibility reasons.

The P_3 pocket does not appear to play any important role: according to eq 3, compounds filling in P_1 and P_2 already score better than micromolar. This is arguably wrong, but, unfortunately, the data set presented to the machine learning tool cannot unambiguously tell whether hydrophobic groups in the P_3 pocket are absolutely necessary for activity or not. There are two compounds without a large hydrophobic group bound to the phenylalanine N, in which this group is actually not substituted at all and therefore cationic. The compounds are inactive, but this is too little evidence to make the model learn that hydrophobes in P_3 are important. Actually, the unsubstituted inactive compounds happen to be properly predicted, due to a penalty stemming from the square of the Hp6-Hp6-PC8 term. The extra positive charge in N-unsubstituted phenylalanines leads to increased population levels of this negatively weighted triplet, i.e., the model seems to suggest that inactivity is due the additional free charge. As far as the only examples missing a P_3 hydrophobe are also the only ones with a protonated phenylalanine N; there is no reason to prefer one explanation over the other.

Although the success of eq 3 appears to be partly due to the 'illusion' that the P_3 pocket may be ignored as something filled 'by default' with a hydrophobe in all the examples given, this does not mean that models accommodating various ligand topologies will be impossible to build once that the training set is furnished with enough examples to document the influence of pharmacophore pattern variation in the P_3 region. Such an equation may be based on several triplets, each regrouping elements from (P_1 and P_2), (P_2 and P_3), and (P_2 and P_3), respectively, binding moieties: there is no need to enter a triplet having each corner from a different

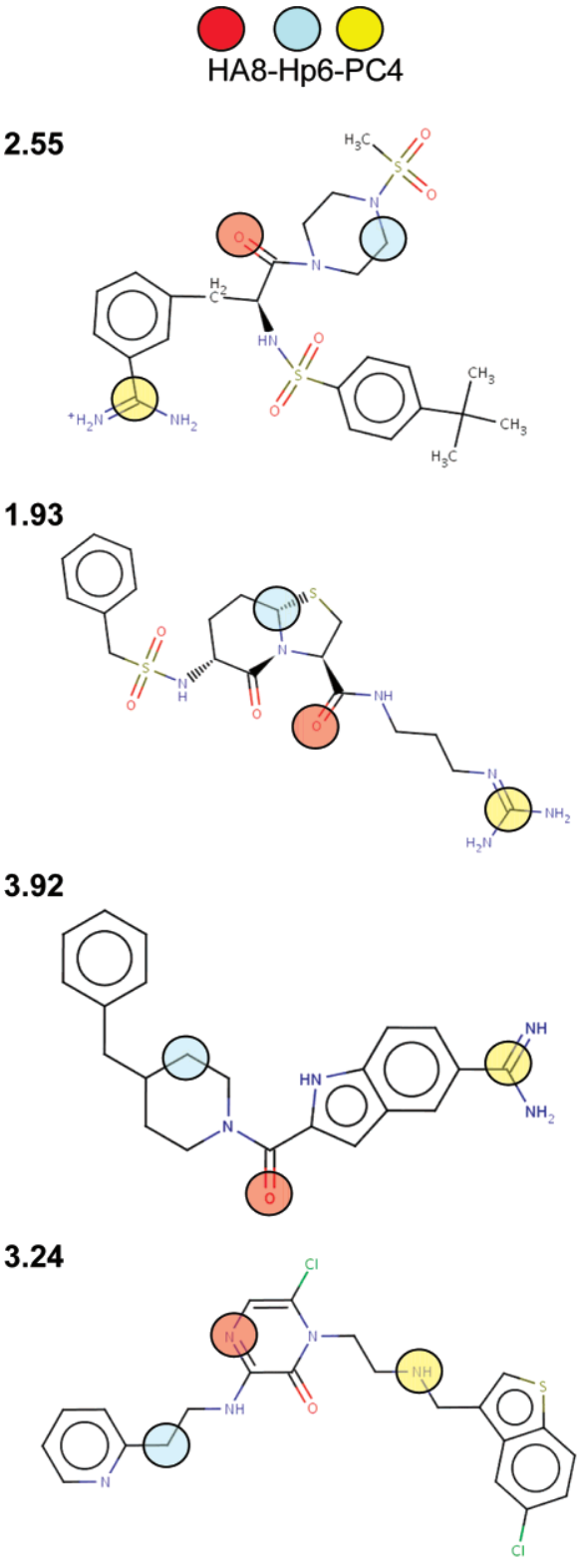

Figure 10. Color-coded display of the occurrence of the key Thr affinity modulating triplet HA8-Hp6-PC4 in the four chemically different inhibitors.

moiety, since such triplets will not be shared thorough topologically different series.

The successful prediction of a typical compound from Figure 8(d) is due to the herein present triplet HA8-Hp6-PC4. However, this very same topological pharmacophore

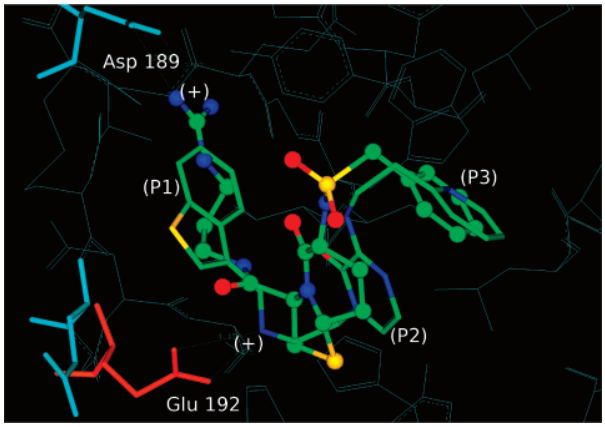

Figure 11. Superimposed thrombin active sites with aligned cocrystallized ligands—Figure 8(b),(d). Glu 192 forming the atypical salt bridge with the cation of compound (d) is seen (in red, vs default light blue) to shift its side chain in order to interact.

seen to previously match elements binding to the P_1 and P_2 binding pickets covers in compound (d) elements seen to go into P_3 (the hydrophobic spot of the pyridine linker) and P_2, respectively. This is a purely accidental example of 'inverse' degeneracy, where the same topological triplet may be accommodated in two different ways in an active site. [This 'inverse degeneracy' is antonymic to the previously illustrated 'classical' degeneracy of 2D-FPT, where different atom triplets may indistinctively contribute to the population level of the same basis triplet.] The cation now forms a salt bridge with Glu192, which reorients its side chain in order to enter this interaction. While the HA corner of the relevant triplet has no direct binding role in compounds (a)−(c), the pyrazone oxygen, which is an alternative contributor to HA8-Hp6-PC4 (not highlighted in Figure 10 for the sake of simplicity), is actually involved in the interaction with the main chain >NH of Gly 216. All this is however anecdotic: the QSAR model did not foresee that the thrombin active site supports this alternative binding mode. Out of the two equations that correctly extrapolated the activity of (b) and (c), on the one hand, of the family of (a), only eq 3 included triplets also shared by (d). Due to the peculiarities of the thrombin active site, the two distinct binding modes might be explained by the same model. From a practical point of view, using eq 3 in a virtual screening would have triggered a major breakthrough in thrombin inhibitor research, (serendipitously) leading to a completely new family. Unfortunately, there are no deterministic recipes to find such models, if ever they happen to exist.

4. CONCLUSIONS

As far as the benchmarking exercise goes, 2D-FPT-based QSARs fare extremely well, outperforming not only 2D and 3D-index-based models but also the elaborate, overlay-based CoMFA approaches. The biological property less well handled by pharmacophore triplet models is, unsurprisingly, the heme alkylating activity of artemisinin analogues—the only studied property not reflecting a reversible noncovalent target inhibition process, conceptually associated with 'binding pharmacophores'. 2D-FPT are thus information-rich and relevant descriptors of site-ligand recognition processes. The study of optimal 2D-FPT fuzziness highlighted the problem of 2D-FPT degeneracy, which may be of serious concern in descriptor selection-based QSARs (much more so than in similarity scoring), although pharmacophore triplets suffer much less from this problem than pairwise descriptors.

Nevertheless, the 'topological pharmacophores' defined by triplets entering 2D-FPT models are not necessarily representatives of ligand-site anchoring points. This work highlighted the very limited scope of the training and validation sets typically used for QSAR buildup and benchmarking, showing many situations where the successful QSAR fitting and validation relied on family specific idiosyncrasies. Another symptom of training set limitations is the generation of models predicting high activity values by default and relying on penalizing terms to reduce the score for the known inactives containing 'unwanted' features, or these models predicting high activities for any molecules too small to contain any triplets, be it wanted or unwanted, are thus senseless.

Although pK_a-dependent pharmacophore flagging was proven to be more rigorous than the rule-based one, leading to a much better understanding of the molecular similarity principle,[1] in QSAR studies, set-specific artifacts gained the upper hand over pK_a-related effects: the best performing flagging scheme was often the one best exploiting some set-specific coincidence.

The broad range of encountered set-specific artifacts (and which surely appeared under different forms with the various descriptors used in the cited literature studies) is a serious incentive to reconsider the actual sense of QSAR buildup, validation, and benchmarking on such limited series. In light of the many examples of chemically flawed equations brilliantly passing 'external' validation tests—against new members of the training family, more precisely—the present work suggests that (a) any training set should be completed with a set of diverse (presumed) inactives before QSAR buildup. This is easily feasible with 2D-FPT and other overlay-independent descriptors but problematic with CoMFA and related tools. (b) An additional challenge against topologically different actives should be regularly included in benchmarking. General equations based on chemically meaningful terms may be enumerated upon extensive sampling of the QSAR problem space, among many other successfully validating family specific models. They are likely to perform reasonably well, without being the best in terms of training/validation scores (therefore, deterministic QSAR build-up procedures may not find them). The challenge to predict topologically different actives is needed to highlight them among the many apparently redundant alternative models.

Concerning the interpretability of 2D-FPT models, it must be pointed out that these were excellent tools to highlight training set deficiencies: the chemically interpretable terms responsible for observed artifacts allow a straightforward comprehension of the problem. Whether or not selected triplets match actual binding pharmacophores is mainly a question of training set diversity. 2D-FPT may lead to valuable QSAR models, provided the training set diversity is sufficient to force the learning of key features, not of secondary pharmacophore signatures that serendipitously reflect subsets locally enriched in actives. If this is the case, the applicability range of such models may extend over several chemotypes—and may even go beyond expectations

if the targeted active site offers alternative models to accommodate a topological triplet.

The setup files.xml controlling 2D-FPT buildup are available upon request from the author.

Supporting Information Available: Thirteen considered data sets plus the compiled extended thrombin inhibitor set ThrEx, for each set, a two-column (SMILES, activity score) <set>.smi.txt file, a list of the molecules entering the validation set <set>.vset.txt, and the activity-descriptor matrices <set>.<descriptor-version>.txt are available, for each descriptor version D, D-R, D-S, O, and C (all files−Unix ASCII). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457−2477.

(2) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces − A Benchmark for Neighborhood Behavior Assessment of Different In Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 691−698.

(3) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Towards generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and related methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.

(4) Milicevic, A.; Nikolic, S.; Trinajstic, N. Toxicity of aliphatic ethers: A comparative study. *Mol. Diversity* **2006**, *10*, 95−99.

(5) Adam, M. Integrating research and development: the emergence of rational drug design in the pharmaceutical industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(6) Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A. M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. Pharmacophore-based design of HIV-1 integrase strand-transfer inhibitors. *J. Med. Chem*. **2005**, *48*, 7084−7088.

(7) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem*. **2005**, *48*, 6997−7004.

(8) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold hopping with molecular field points: identification of a cholecystokinin-2 (CCK2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based CCK2 antagonists. *J. Med. Chem.* **2005**, *48*, 6790−6802.

(9) Cramer, R. D., III; Patterson, D. E.; Bunce, J. E. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(10) Horvath, D. ComPharm − Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design. In *QSPR /QSAR Studies by Molecular Descriptors*; Diudea, M., Eds.; Nova Science Publishers, Inc.: New York, New York State, 2001; pp 395−439.

(11) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Chemoinformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; pp 117−137.

(12) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687−698.

(13) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687−2694.

(14) Fechner, U.; Paetz, J.; Schneider, G. Comparison of Three Holographic Fingerprint Descriptors and their Binary Counterparts. *QSAR Comb. Sci.* **2005**, *24*, 961−967.

(15) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of d(1) dopaminergic antagonists for database mining. *J. Med. Chem*. **2005**, *48*, 7322−7332.

(16) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J. Med. Chem.* **2005**, *48*, 6563−6574.

(17) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1214−1223.

(18) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1998**, *38*, 144−150.

(19) Sciabola, S.; Morao, I.; de Groot, M. J. Pharmacophoric Fingerprint Method (TOPP) for 3D-QSAR Modeling: Application to CYP2D6 Metabolic Stability. *J. Chem. Inf. Model.* **2007**, *47*, 76−84.

(20) Güner, O. F. *Pharmacophore Perception, Use and Development in Drug Design, IUL Biotechnologies Series;* Güner, O. F., Eds.; International University Line: La Jolla, CA. 2000.

(21) Sutherland, J. J.; OBrien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(22) Skold, C.; Karlen, A. Development of CoMFA models of affinity and selectivity to angiotensin II type-1 and type-2 receptors. *J. Mol. Graphics Modell.* **2007**, *26*, 145−153.

(23) Guha, R.; Jurs, P. C. Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.

(24) Coats, E. A. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discovery Des.* **1998**, *12*−*14*, 199−213.

(25) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation - How much effort may the mining for successful QSAR models take? *J. Chem. Inf. Model.* **2007**, *47*, 927−939.

(26) Wagner, J.; Kallen, J.; Ehrhardt, C.; Evenou, J. P.; Wagner, D. Rational design, synthesis and X-ray structure of two selective noncovalent thrombin inhibitors. *J. Med. Chem.* **1998**, *41*, 3664−3674.

(27) Chirgadze, N. Y.; Sall, D. J.; Klimkowski, V. J.; Clawson, D. K.; Briggs, S. L.; Hermann, R.; Smith, G. F.; GiffordMoore, D. S.; Wery, J. P. The crystal structure of human alpha-thrombin complexed with LY178550, a nonpeptidyl, active site-directed inhibitor. *Prot. Sci.* **1997**, *6*, 1412−1417.

(28) Bulat, S.; Bosio, S.; Papadopoulos, M. A.; Cerezo-Galvez, S.; Grabowski, E.; Rosenbaum, C.; Matassa, V. G.; Ott, I.; Metz, G.; Schamberger, J.; Sekul, R.; Feurer, A. Design and Discovery of Novel, Potent Pyrazinone-Based Thrombin Inhibitors with a Solubilizing Amino P1−P2-Linker. *Lett. Drug Des. Discovery* **2006**, *3*, 289−292.

(29) Horvath, D. et. al. − unpublished work.

(30) RCSB Protein Data Bank. http://www.rcsb.org/pdb/ (accessed Oct 08, 2007).

(31) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(32) Cazelles J.; Robert A.; Meunier B. Alkylation of heme by artemisinin, an antimalarial drug. *C. R. Acad. Sci., Ser. IIc: Chim*. **2001**, *4*, 85−89.

(33) Chemaxon − Pmapper user guide. http://www.chemaxon.com/jchem/doc/user/PMapper.html#config (accessed Oct 10, 2007). Also check the pharma-frag.xml configuration file in the JChem distribution.

(34) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.

CI7003237

Neuvième partie

L'utilisation de Cartes Auto-Organisatrices pour accélerer les recherches par similarité

Les cartes auto-organisatrices (SOMs) développées par Teuvo Kohonen ([7]) sont populaires en chémoinformatique car :

- Elles sont des représentations bidimensionnelles faciles à comprendre d'un espace d'entrée multidimensionnel,
- Elles sont capables de préserver les propriétés topologiques de l'espace d'entrée
- Elles permettent de représenter les données d'entrée sous forme de regroupements (*clusters*) : des objets projetés sur un même neurone (ou nœud) ou sur des neurones adjacents sont considérés comme étant similaires.
- Elles permettent de faire des prédictions : un objet nouveau projeté sur un neurone aura une forte probabilité de posséder des caractéristiques communes avec les autres membres du neurone.

Grâce à leurs propriétés particulièrement utiles, elles sont souvent utilisées pour projeter et décrire l'espace chimique d'un ensemble de composés. Cependant, nous avons choisi dans nos travaux de les utiliser dans un but différent : les prendre comme base afin d'accélérer les recherches par similarité dans une base de données.

Afin se sélectionner les cartes les plus aptes à aider à l'accélération des recherches, un test de performance a été mis en place. Il consiste à récupérer, à partir d'une base de données de composés (**DB**, constituée d'environ 55,000 molécules), les *Hits* virtuels (c'est à dire, les plus proches voisins se trouvant en dessous d'un seuil de dissimilarité spécifié) pour chacun des membres d'un "ensemble de requêtes" (***QS*** – *Query Set*) de 2000 composés. De plus, de bonnes cartes ont été mises à l'épreuve dans des tests grandeur nature, en utilisant comme requête un ensemble d'environ 12.000 composés (**ExtQ**) à comparer à une base de données industrielle de 160.000 composés (**ExtDB**).

Les 2D-FPTs ont été utilisés comme descripteurs sur l'intégralité des composés de cette étude. Etant de haute dimension (chaque molécule est associée à un vecteur de descripteurs à 4418 dimensions) et basés sur des nombres réels (au lieu d'opérations sur les bits), les procédures classiques d'accélération utilisées sur des empreintes binaires ne peuvent être appliquées ([83, 8, 36]). Cependant, les cartes auto-organisatrices permettent d'appliquer des techniques de division des données afin de sélectionner lesquelles seront à utiliser dans le criblage.

Plusieurs centaines de cartes ont été générées afin de les soumettre aux tests d'accélération. Elles ont été entrainées sur deux ensembles de tailles différentes (environ 11.000 et 53.000 molécules respectivement), en variant les paramètres d'entrainement ainsi que les paramètres de taille, de topologie et de fonctions de voisinage. Après la phase d'entrainement, les deux ensembles de test **QS** et **DB** sont projetés sur chaque SOM. Les composés du **QS** ne sont comparés qu'avec les composés de **DB** résidant dans le même neurone ou dans les neurones voisins. La taille de la zone des neurones voisins autour du neurone requête varie en fonction des résultats du test (tant que 90% des *Hits* ne sont pas retrouvés, le programme continue à chercher parmi les neurones un cran plus loin). Plus cette zone est petite, moins il ya de composés de la *DB* à comparer à chaque requête : le temps de criblage virtuel est donc réduit. Il faut noter cependant que les composés de **DB** qui se trouvent en dehors de la zone sont considérés comme "dissimilaires", ce qui a un impact sur le taux de récupération des ensembles initiaux de *Hits* virtuels. Un compromis est donc à trouver entre accélération et perte de *Hits* virtuels.

Un critère conciliant ces tendances opposées a donc été défini, afin de caractériser l'accélération par la SOM contre l'efficacité du taux de récupération. Ce critère simple permet la comparaison des performances relative de toutes les cartes créées. Nous avons grâce à ce critère étudié l'impact des choix de construction des cartes (taille de l'ensemble d'entrainement, taille et géométrie des cartes, critère de convergence imposé, choix des fonctions de voisinage) sur leur efficacité. Le but étant de donner un ensemble de recommandations pratiques pour entrainer efficacement des cartes ayant une efficacité optimale d'amélioration du criblage virtuel, nous avons de plus comparé notre critère d'accélération aux critères de qualité classiques des SOMs (erreur de quantization, visualisation, homogénéité).

Il est démontré dans ces travaux qu'augmenter la taille du set d'entrainement au delà d'une certaine limite se fait au détriment de la qualité de la carte : trop de composés d'entrainement entrainent des problèmes de convergence, ce qui pourrait supplanter les bénéfices supposés de l'ajout de nouvelles informations. De plus, l'impact de l'entrainement est analysé en profondeur et il est montré qu'il est aussi important de bien entrainer une carte que de bien choisir l'ensemble sur lequel cet entrainement est fait. La meilleure carte ressortant des comparaisons est décryptée et présentée dans ces travaux. Enfin, les tests grandeur nature sur 4 cartes sélectionnées pour leurs bons résultats démontrent que notre critère permet effectivement de décrire le comportement des cartes sur de nouveaux jeux de données.

Dixième partie

Using Self - Organizing Maps to Accelerate Similarity Search (*submitted on 31th january 2012*)

Using Self-Organizing Maps to Accelerate Similarity Search

Bonachera Fanny[a,b], Marcou Gilles[a], Kireeva Natalia[a], Varnek Alexandre[a], Horvath Dragos[a,*]

[a] *Laboratoire d'Infochimie UMR 7177, Université de Strasbourg, 1, rue B. Pascal, 67000, Strasbourg, France*
[b] *Unité de glycobiologie structurale et fonctionnelle UMR 8576, Université Lille I, Bâtiment C9, 59655, Villeneuve d'Ascq Cedex, France*

Abstract

While Kohonen Self-Organizing Maps (SOM) have often been used in Chemoinformatics to map and describe chemical space, this paper focuses on their use to accelerate similarity searches based on information-rich, high-dimensional real-value descriptors in a database of small molecules. Similarity (both Euclidean and Tanimoto) is calculated in terms of vectors of fuzzy tricentric pharmacophore (FPT) descriptors. These are real-value, high-dimensional descriptors for which classical, binary fingerprint-based similarity speed-up procedures do not apply. Similarity search speed-up was achieved by repositioning candidate compounds on SOM, then focusing the search for analogues on the neurons in the direct neighbourhood on the one in which the query compounds reside. Smaller neighbourhood means shorter virtual screening time, but lower analogues retrieval rates. An enhancement criterion, conciliating the opposite trends is defined. It directly depends on map definition and build-up protocol (training set size, map size & geometry, imposed convergence criteria, choice of neighbourhood functions). The main goal of this paper was to discover and validate SOMs of optimal quality with respect to this criterion. It was shown that increasing the size of the training set beyond a certain limit becomes detrimental to map quality: too many training compounds raise convergence problems. Also, using an excessively large number of training iterations may lead to over-fitting. Gradual training with *en-route* checking of VS enhancement propensity is the best strategy to follow. Eventually, maps were successfully challenged to accelerate the large-scale virtual screening of 12,000 queries against a database of 160,000 compounds and also shown to provide a meaningful mapping of activity-annotated compounds in chemical space.

Keywords: Chemical space navigation, Virtual screening, Similarity searching, Fuzzy pharmacophores, Kohonen maps

*Corresponding author

1. Introduction

Similarity-based virtual screening is an integral part of modern *in silico* drug discovery. [1]

This approach is based on the paradigm that structurally close compounds may also have similar activity against a given target. [2] A candidate molecule from some large structural database will be considered as potentially active if a similarity search shows that it is related (in terms of its molecular descriptors, and using an appropriate similarity measure [3]) with one or more known actives. The results of these searches are ranked lists of all screened compounds, along with similarity scores. The highest ranked compounds of these lists are assumed to be the closest to the query in terms of activity. [4]

However, with the ever-growing size of databases, similarity searches become more and more time-consuming. While this is not a practical concern when binary fingerprints are used for similarity searches - for which very fast search methods have been designed - the situation is different if one wishes to employ information-rich, high-dimensional real number vectors, such as fuzzy pharmacophore descriptors. While chemically relevant, and intrinsically very powerful in similarity searching (particularly well suited for "lead hopping" [5]), they require floating-point operations instead of fast bit-wise matching and, furthermore, are not eligible for search acceleration procedures developed for binary fingerprints [6, 7, 8].

Similarity search speed-up techniques rely on some kind of "divide and conquer" approach aimed to split the original search space into sub-domains, and then decide beforehand, in function of the given query, which sub-domain(s) may be ignored during the search. Real-value descriptors do not naturally provide a "granular" search space, by contrast to binary fingerprints. However,Self-Organizing maps (SOMs) may be used with such descriptors in order to obtain the needed "tessellation" of the chemical space. Furthermore, SOM-driven tessellation is non-linear with respect to the chemical space (SOMs being a particular type of neural networks [9, 10]), and easy to visualize (output being a 2D grid of "nodes" or "neurons"). Each node is represented by a weight vector (the code vector) of the same dimension as the molecular descriptors. The fact that the Self-Organizing Maps are able to preserve the topological properties of the input space have made them popular in Chemoinformatics [11]. The two-dimensional generated maps are :

- Easy to understand two-dimensional representations of the high-dimensional input.
- Clustered representations of the input : objects that are mapped in the same node or in adjacent nodes can be considered as similar.

The applications of Kohonen Self-Organizing Maps in Chemoinformatics research are varied. It has been demonstrated that Kohonen maps can effectively be used to produce topology-preserving maps of small molecules, providing ways to compare compounds and assess similarity [12, 13]. They have also proven

to be very effective in classifying and clustering compounds as well as describing the chemical space of a database [14], detecting novelties [15, 16], studying structure-activity relationships [17], mapping pharmacophores [18], selecting virtual screening candidates [19, 20], predicting or mapping properties [21], or comparing chemical libraries [22].

The use of Kohonen Self-Organizing Maps to accelerate search has already been published in the image retrieval domain [23, 24]. The association of classification of images on a SOM with a k-Nearest Neighbors similarity scoring function proved to be a quick and effective way of retrieving images.

Here, SOMs will be used to enhance similarity-based virtual screening (VS). While typically tree-based acceleration methods [6, 7] are used to speed up search in large databases, these have several drawbacks compared to the use of self-organizing maps:

- SOMs as virtual screening enhancers are compatible with various similarity metrics, whereas trees are constructed with respect to one particular metric. Although SOMs operate on the basis of euclidean metric, they are nevertheless able to successfully accelerate Tanimoto similarity-based searches, as will be shown in this work. Actually, any similarity metric can be used "on the fly", in as far as its neighbourhood behaviour [25] is correct.

- Furthermore, there is no need to rebuild the SOMs if the database is extended. This is systematically required with search trees. New compounds need just be mapped into corresponding nodes of the existing map. Building a tree requires the entire database, whereas building a relevant and robust SOM only needs a representative fraction of the database. Actually, as will be shown in this work, oversized SOM training sets may have detrimental influence on SOM performance.

- A tree model can sometimes miss Hits (because some branches are ignored during a search), whereas in the herein presented approach the hit retrieval rate is tunable, and may be balanced off against the time gain by specifying the size of the searched map neighbourhood around the residence node of the query compound.

The VS benchmark consists in retrieving, out of a compound database (DB, $\sim$ 55,000 molecules), the virtual hits (nearest neighbours below a specified dissimilarity threshold) for each of the members of a "query set" (QS) of 2000 compounds. Similarity (both Euclidean and Tanimoto) is calculated in terms of vectors of fuzzy tricentric pharmacophore (FPT) descriptors. The CPU times required to match each QS member against every DB compound are determined and used as reference values. To speed up this virtual screening protocol, a set of SOMs have been trained on two training datasets of different sizes (of about 11,000 and 53,000 molecules, respectively). After the training phase, both QS and DB were then mapped thereon, and QS members were only matched against DB compounds residing on the same or on neighbouring neurons of the SOM

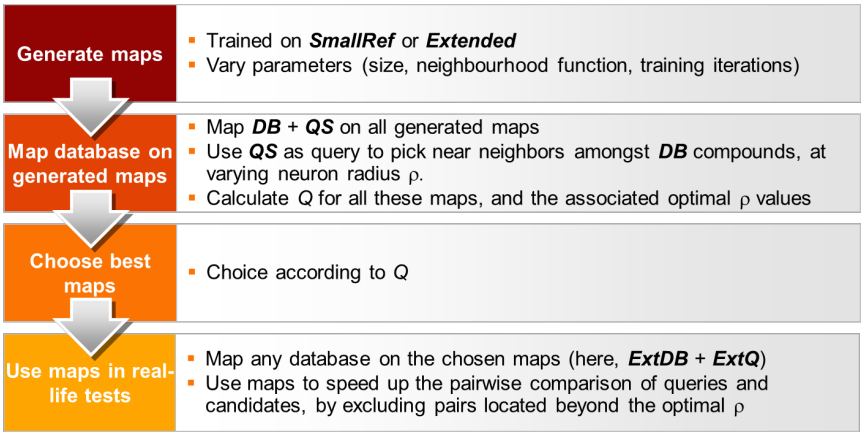

Figure 1: Workflow of building, assessment, selection and testing of SOMs with good similarity-based virtual screening enhancement propensities

(at increasing size of the squared neighbouring neuron area around the query neuron). The smaller this area, the less DB compounds must be matched against each query: virtual screening time is therefore decreasing. However, since DB compounds outside this area are by default considered "dissimilar", the retrieval rate of the initially established sets of virtual hits will also decrease. A best compromise criterion, conciliating thee opposite trends, is defined in order to characterize the SOM speed-up *vs.* retrieval rate efficiency, and thus compare the relative performances of the various SOMs. The workflow is shown in Figure 1.

The final goal of this work is to come up with a set of practical recommendations about how to train maps with optimal virtual screening enhancement proficiency, by studying the impact of map build-up choices (training set size, map size & geometry, imposed convergence criteria, choice of neighbourhood functions) on their proficiency. The following text is structured as follows: in Methods, after presentation of the employed data sets and molecular descriptors, the SOM technology is briefly introduced. The systematic scan of combinations of considered SOM parameters is described. The definition of the VS enhancement criterion recommended as the objective score of SOM quality is given. Eventually, a real-life VS experiment validating the usefulness of SOMs as VS accelerators is presented. The Results section devotes a lengthy discussion to the problem of proper map training: reaching convergence while avoiding overfitting. The impact of training set choice on map quality is the next important chapter, followed by a presentation of the to-date best ranked map in terms of the herein proposed VS enhancement criterion. Eventually, the behaviour of top-ranking maps in the real-life VS experiment is illustrated, before drawing the final Conclusions.

2. Methods

2.1. Data Sets

There are six, partly overlapping data sets to which this work refers. These are as diverse as possible compound collections, randomly extracted from a large panel of sources, in order to minimize set-related artefacts. Care has been taken to include both chemically "dense" series of analogues, "sparse" sets of drugs and reference compounds (in which only the diverse marketed structures stand alone, not surrounded by analogues - unless several related structures have been marketed as drugs) and finally, commercial organic compound databases of more or less drug-like compounds, part of chemical series or singletons. Also, since SOMs are unsupervised learning methods, no particular attention was paid to ensure that SOM training sets are distinct from the VS molecules.

- The DataBase **DB** represents a pool of 55613 molecules including random subsets of 11 different analogue series used to model structure-property relationships in literature [26], marketed drugs and biological reference compounds, 1870 ligands from the Pubchem database tested on the hERG channel [27], and a majority of randomly picked ZINC compounds.

- The Query Set **QS**, is composed of 2000 molecules, regroups the remainders of the 11 above-cited series, further marketed drugs and biological reference compounds and commercially available molecules (picked randomly from the ZINC database, [28]). There is no overlap between DB and QS. These molecules serve as starting points (queries) for similarity-based virtual screening against DB, in order to assess the ability of our SOM-diven screening tool to find, for all of the 2000 queries, a maximum of their nearest neighbours amongst DB compounds, with a minimum of effort. Thus, QS and DB are the data sets used to assess SOM quality, *not* to train the SOMs. The SOM training sets, which only partially overlap with QS, are the following:

- The Large SOM training set **Extended** of 53206 molecules is basically a subset of previously available molecules (DB+QS), excluding the analogue series members, the Pubchem compounds and some 900 ZINC molecules. More precisely, out of the 2000 QS molecules, 149 are members of both *SmallRef* and *Extended*, while the others are never used to train maps.

- The Small SOM training set **SmallRef** of 11168 molecules features all drugs and biological reference compounds seen in *Extended*, but significantly less ZINC molecules.

- Eventually, the External Database **ExtDB** of roughly 160,000 molecules from the corporate collection of one of our industrial partners: this was used, in the final stage, to verify the performance of the map-enhanced virtual screening tool, deployed on multiple processors, under real-life conditions. This set has been screened against 12,491 query compounds - taken

basically from *SmallRef*, and completed with randomly picked commercial compounds: let us refer to this query set as **ExtQ**

- A subset of ligands from the Database of Useful Decoys (**DUD**) [29], regrouping the binders to the ten most ligand-rich targets displayed in Table 3.3, and used to assess the proficiency of maps in telling various ligand classes apart.

2.2. Descriptors and Metric: The Chemical Space of the Similarity-based Virtual Screening Approach

Fuzzy Pharmacophore Triplets (FPT)[30, 26] represent fuzzy counts of monitored triplets of potential pharmacophore points (PPP), at given topological inter-feature distances, i.e. "edge lengths" of the considered triangles. As discussed in the original publication [30], six different PPP types (hydrophobic, aromatic, hydrogen bond donor and acceptor, positive and negative charge) were assigned to the atoms within every microspecies present (at significant population levels) in the proteolytic equilibrium of the molecule at a given pH of 7.4. Molecular fingerprints are averages of microspecies fingerprints, accounting for their relative population levels. In this work, only FPT1 descriptors, corresponding to the default setup of fuzzy triplets in the original work [30] were considered. In the *SmallRef* set, 4418 different pharmacophore triplets were found to be populated (out of the 4494 defined for the FPT1 setup). Therefore, this subset of 4418 triplets has been systematically used to construct the 4418-dimensional descriptor vectors positioning the molecules in the Chemical Space (CS). The averages $< D_i >_{SmallRef}$ and standard deviations $\Sigma(D_i)_{SmallRef}$ of the 4418 vector elements were taken over the *SmallRef* compounds and employed to normalize the descriptors prior to both virtual screening and SOM buildup/mapping. Therefore, the finally employed descriptor vectors are $d_i = [D_i - < D_i >_{SmallRef}]/\Sigma(D_i)_{SmallRef}$, which, for *SmallRef* molecules, corresponds to classical "Z-transformed" [31] vectors. Please note that, with other sets, *SmallRef* averages and deviations are employed rather than the actual values over the respective sets.

For each query compound of QS, both its Euclidean and its Tanimoto distance to every DB molecule (the latter being defined as 1-Tanimoto index) was computed, on hand of the d_i values. Generically speaking, let the inter-molecular dissimilarity score (whether Euclidean or Tanimoto) between two molecules m and M be denoted $\Sigma(M, m)$. An empirical distance threshold of 0.25 was picked to delimit "virtual hits" in terms of Tanimoto distances (meaning a maximum of 25% of tolerated dissimilarity). In Euclidean space, the equivalent cutoff value (roughly corresponding to the same number of "virtual hits") was found to be of 9.0. Therefore, for each query compound, two virtual hit lists VHL - based on Tanimoto VLH_T and Euclidean VLH_E distances, respectively - were established. They include the first 300 (or fewer) nearest neighbours from DB, found to be within the respective cutoff radii with respect to the query. Virtual screening has been carried out by a FORTRAN executable on a x86_64 Intel

CPU, and the CPU times t_E^{ref} and t_T^{ref} for the complete virtual screening according to Euclidean, respectively Tanimoto distance matrices were measured using the unix *time* command. These times include, next to the actual time span of the computation of the $QS \times DB$ distance matrices, the overheads for data input and virtual hit list output.

2.3. Build-up of the Self-Organizing Maps (SOM)

2.3.1. The SOM_PAK software

SOM_PAK, first published in 1992 [32] is a program package written in ANSI C that contains all necessary tools to build and exploit Self-Organizing Maps.

The package provides tools to initialize, train maps, evaluate the quantization error (see below "evaluating maps quality") and visualize maps. To initialize a map, the user needs to specify the following input parameters:

- The name of the file containing the molecular descriptors of the training compounds.
- The map topology: with SOM_PAK, rectangular or hexagonal arrangements of the neurons in a 2D lattice are supported. Only the "rectangular" option has been used here. By contrast to the multidimensional chemical space (4418-dimensional, according to the number of monitored fuzzy triplets), also referred to as "input space" in the following, the rectangular 2D lattice of neurons is the "output space" of the map. In output space, molecules are located onto their "winning neuron", the one of minimal Euclidean distance to that molecule, in input space. Distances in output space are defined between two neurons, as the Euclidean distance between the associated lattice nodes. Implicitly, in output space the dissimilarity between two molecules is given by the distance between their winning neurons
- The wanted dimensions of the resulting map, in terms of two integers defining $length \times width$ of the rectangular neuron lattice. For each of the $length \times width$ neurons, a characteristic "code vector", positioning them in the chemical space of input molecules, need to be fitted. The number of fittable parameters in a SOM therefore equals to $length \times width \times$ *dimension of the chemical space*
- The neighbourhood function, which can be "Gaussian" or "bubble". When updating the code vectors of each neuron, the shift of the code vector of a neuron induced by a given molecule must be obviously a decreasing function of the distance between this neuron and the "winning neuron" associated to that molecule. This function may be either a Gaussian or a bubble function (*i.e.* return one if the distance is below a given threshold or "learning" radius, and zero otherwise). Likewise, the width of the Gaussian bell is also tunable (the "learning radius" being here associated to the half-width of the Gaussian bell). SOM_PAK will decrease the initial, user-input learning radius linearly with the number of iterations, in

order to enhance the convergence. To the same purpose, the absolute value of molecule-induced shifts is also being dampened as fitting proceeds, by multiplying them to a learning parameter taking an user-defined value α at the beginning of the fit procedure, and linearly decreasing with respect to the iteration number.

Map initialization. The reference vectors (code vectors) were initialized using the *randinit* program. All reference vectors components are first set to random values evenly distributed in the area of corresponding training data input vectors components.

Map training. After initialization, we have used the *vsom* program to train the reference vectors. This program uses rough random map parameterized at the initialization step. During this training phase, *vsom* finds the best matching node (or neuron) for each training input data vector and uses the neighbourhood function to update the nodes neighbours.

The important parameters are *rlen* (the number of training steps), *alpha* (initial learning rate parameter, which decreases linearly to zero during training). and the *radius* (initial radius of the training area, which decreases linearly to one during training) parameter.

Further training processes, taking as input already trained maps instead of randomly initialized ones, may be used to continue the refinement.

2.3.2. Building the SOMs

Various maps have been built using the two training sets (*SmallRef* and *Extended*). We chose to vary X and Y dimensions from 8x6 to 30x30, which corresponds a range of 48 to 900 neurons. Practically, the 36 explored map geometries were: 6x12, 6x14, 8x10, 8x6, 10x10, 10x14, 10x20, 12x8, 14x6, 18x18, 18x20, 20x20, 22x24, 22x26, 22x28, 22x30, 24x22, 24x24, 24x26, 24x28, 24x30, 26x22, 26x24, 26x26, 26x28, 26x30, 28x22, 28x24, 28x26, 28x28, 28x30, 30x22, 30x24, 30x26, 30x28, 30x30. Each such map has been built, based on every training set, in both "Gaussian" and "bubble" version, leading to a total of $`2 \times 2 \times 36 = 144$ processed SOMs.

Map fitting is driven by the objective to minimize the quantization error *QE*, the average Euclidean distance between each molecular descriptor vector and the closest code vector (the one of the neuron to which it was assigned). In terms of fitting strategies, the calibration of the SOMs based on *Extended* was expected to be more difficult to achieve, therefore different fitting strategies were used:

Three successive runs were employed for map calibration.

- Brute training at *rlen* of 1000, *alpha* of 0.05 and a *radius* equalling the *lenght* of the neuron lattice.
- Refinement at *rlen* of 10000, *alpha* of 0.02 and a *radius* of $max(6, length/2)$.

- HyperRefinement at variable *rlen* (typically 50000, which will also be reffered as "Standard HyperRefinement" - other values being indicated in the text), *alpha* of 0.01 and a *radius* of $max(3, length/4)$.

A study of the convergence of the maps. was conducted in order to make sure that the above-mentioned conditions are sufficient, and that the map quality criteria are not biased due to their failure to converge. To this purpose, based on *Extended*, using Gaussian neighbourhood and for two different sizes (a modest 10×10 and a larger 22×28), series of successive maps were saved at every step of a succession of training runs. Training always included a Brute and a Refinement phase, followed by HyperRefinement stages of different lengths. Map quality criteria were then monitored throughout these series, in order to check how many iterations were required before they reached a stable value. In parallel, a similar study was conducted for the bubble 22×28 configuration, based on *SmallRef*, in order to verify the set size impact on the convergence rate.

2.4. Maps visualization with SomView

The software used for map visualization and to make all SOM figures of this paper is an in-house developed software, SomView. It allows us to open .fvs files (lists of coordinates corresponding to the best-matching nodes in the map for each training data sample, created with the *visual* program), together with the associated .sdf files of mapped molecules, and to easily visualize rectangular Self-Organizing Maps.

The maps are depicted as rectangular grids in which each node is a pie colored by labels or by label majority, depending on the chosen option. If an .sdf file is provided, the contents of each node can be viewed in 2-dimensional structures. Other nodes visualization options are available : node color and size by topographic index (see below - "Evaluation of Self-Organizing Maps quality"), node size by quantization error.

2.5. Map-enhanced Similarity Searching

2.5.1. Assigning compounds onto map neurons

Before performing similarity search tests, both QS and DB sets are mapped on the current Kohonen map κ, *i.e.* each of the compounds is being assigned to the closest map neuron. This is performed with the *visualize* tool. Formally, each compound is thus being assigned a 2D position on the map grid, coded by two integers: the neuron position (N_x,N_y). Unlike in complete similarity-based screening, a given QS molecule will now be compared to a DB compound only if the host neurons of the two structures are acceptably close to each other. The highest acceptable distance between two neurons for which direct similarity scoring needs to be performed is called the *neuron radius* ρ. Let $(N_x, N_y)|_M$ and $(N_x, N_y)|_m$ be the residence neurons of molecules M and m respectively. These two molecules are subjected to an actual similarity calculation of their descriptor vectors only if $|N_{x,M} - N_{x,m}| \leq R$ and $|N_{y,M} - N_{y,m}| \leq R$. For a

query compound M located on the query neuron in Figure 2, only DB molecules located on neurons within the box of size $2R+1$ around the query neuron would qualify for explicit $\Sigma(m, M)$ calculations. Otherwise, the inter-molecular dissimilarity score $\Sigma(M, m)$ will be by default assumed infinitely large. In other words, unless they reside on neurons that are no more than ρ units apart, on either dimension of the map grid, M and m do no longer count as neighbours. This working hypothesis allows a quick, and -if the map is meaningful- effective discarding of (m, M) pairs for which the actual costly calculation of $\Sigma(M, m)$ in the 4000-plus dimensional descriptor space may be spared. Figure 3 illustrates how the inter-molecular dissimilarity matrix calculation is impacted by the choice of ρ. When rebuilding the virtual hit lists VHL using map κ, at a given ρ value, the lower ρ - and the lower that quality of κ, the higher the odds that neighbour pairs originally seen in the VHL issued from the exhaustive pairwise comparison (see subsection 2.2) will now be missed. Reversely, the lower ρ and the better the quality of κ, the lesser the effort to calculate the $QS \times DB$ dissimilarity matrix.

2.5.2. Virtual Screening Enhancement Factor of a Map.

Let the *retrieval rate* RR_Σ represent the average, over all query compounds, of the fraction of retrieved nearest neighbours (using κ, at ρ), with respect to the total number of neighbours in the exhaustive VHL_Σ at metric $\Sigma = \{E(uclidean), T(animoto)\}$. Note: for the roughly 25% of singletons having no nearest neighbours at all, RR will always count as 1.0: map-driven virtual screening at any ρ values cannot be held responsible for failing to detect near neighbours when there are none. Let f_Σ represent the time it took to perform the $QS \times DB$ dissimilarity matrix calculation, using κ at ρ, compared to the reference time t_Σ^{ref}. The "$RR-f$" plot of $RR_\Sigma^{\kappa@\rho}$ *vs.* $f_\Sigma^{\kappa@\rho}$ at increasing ρ values describes a curve originating at $\rho = 0$ ($\Sigma(M, m)$ calculations restricted only to molecules m residing on the same neuron as the query M - the point at lowest time fraction and at lowest retrieval rate). Increasing ρ will eventually lead to picking all the possible (m, M) pairs for explicit dissimilarity calculation, thus $RR = 1$ and $f = 1$ if the box of size ρ covers the entire map. In between theses extremes, the Virtual Screening Enhancement factor with respect to dissimilarity metric Σ, for the map κ needs to be optimized by scanning for the best time enhancement $1 - f_\Sigma^{\kappa@\rho}$ *vs.* retrieval rate compromise, over increasing radii:

$$Q_\Sigma^\kappa = \max_\rho [RR_\Sigma^{\kappa@\rho} \times (1 - f_\Sigma^{\kappa@\rho})^2] \quad (1)$$

In equation 1, time enhancement is squared, in order to emphasize that time effectiveness is, in this case, a more stringent demand than a perfect retrieval of all the possible near neighbours. Eventually, since a map is expected to be an efficient VS enhancer irrespective of the actually employed dissimilarity metric, the average criterion Q^κ is taken as the arithmetic mean of Q_T^κ and Q_E^κ as defined in equation 1.

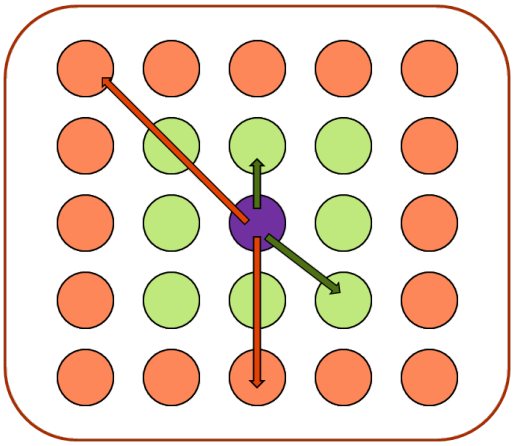

Figure 2: Radius around the query node (in purple) : green neurons correspond to $\rho = 1$, orange neurons correspond to $\rho = 2$.

2.5.3. Real-Life Testing: Large Library Comparison with the Map-Enhanced Virtual Screening Tool

In order to allow rapid matching of large sets of commercial compounds against a typical corporate collection - either with the goal to discover similar, potential hits close to already existing corporate compounds, or to seek for original pharmacophore patterns, not yet represented in the corporate collection - a SOM-driven Virtual Screening tool has been designed to automatically break up the problem into tractable sub-problems, and then deploy each sub-problem on an independent CPU. First, the guiding map (some optimal configuration discovered during benchmarking) must be installed, step at which the compounds of the corporate collection *ExtDB* are assigned to their winning neurons of the new map, and the FPT1 descriptors of compounds residing on a same neuron are grouped together in a common, neuron-specific file. For example, reference compounds assigned to neuron (2,1) will have their FPT1 values - associated to the actual molecular IDs - temporarily saved in a file called **REF2_1.FPT1**. This operation must be done only once, for a given map: then, accordingly reshuffled *ExtDB* fingerprints are ready to be confronted to an arbitrary number of external queries *ExtQ*. External queries are first encoded under the form of FPT1 descriptors as well, then they are assigned to their winning neurons too. Likewise, the FPT1 vector of co-resident query compounds on neuron (x,y) are compacted into common files, say **Qx_y.FPT1**. Eventually, given the neuron radius ρ recommended for use with the installed map, it is straightforward to list, for each query file **Qx_y.FPT1**, the set of reference files **REFX_Y.FPT1** which require to be confronted to the given set of co-resident queries. For a given (x, y) and the corresponding sets (X, Y) with $|x - X| \leq \rho$ and $|y - Y| \leq \rho$, the query file and its associated reference FPT1 files are dispatched for exact Tanimoto-based compound dissimilarity scoring (of every member in **Qx_y.FPT1**, against

m	M	Σ(m,M)	NeurXY	$\Sigma^{\rho=1}$(m,M)	$\Sigma^{\rho=2}$(m,M)
1	2	0.123	(3,8)-(4,8)	0.123	0.123
1	3	0.751	(3,8)-(7,1)	∞	∞
...	...	...	(...)-(...)	...	...
i	j	0.057	(4,3)-(6,1)	∞	0.057
i	j+1	0.438	(4,3)-(4,5)	∞	0.438
...	...	...	(...)-(...)	...	...
N-1	N	0.632	(1,3)-(1,3)	0.632	0.632

Figure 3: Illustration showing how the choice of the neuron radius ρ impacts on the calculation of the inter-molecular dissimilarity score matrix $\Sigma(M, m)$. Given the actual values in column 3, and the neuron co-ordinates of compounds in the pairs (column 4), it can be seen that at small ρ values, most of the pairwise comparisons are no longer carried out: at $\rho = 1$, dissimilarity is supposed to be infinite unless compounds are not residing within the same or within adjacent neurons. This is justified in most cases - coloured in green, where the effort of the calculation of high Σ values is spared. Sometimes, this may cause pairs of similar compounds to be not recognized as such: situation depicted in red. Further increasing of the radius may eventually bring such pairs back into the subset of potentially close neighbours. Orange boxes represent situations of dissimilar molecules, nevertheless located on neurons within the range of ρ.

each molecule in **REFX_Y.FPT1**) on some free node on a cluster or - in the present test - on an available CPU out of the four of the used x86_64 RedHat workstation. The master script then pauses until a new computational resource is available to dispatch new Q-Ref file bundles, until all the query compounds have been confronted - each with the relevant subset of *ExtDB*, the one residing on neurons close to its own. Hits outside the employed neuron radius ρ will never be retrieved and are not known to this date (the baseline, complete computation of the similarity matrix of $12,491 \times 1.6\ 10^5 = 2$ billion floating-point Tanimoto score calculations in 4418 dimensions has not been attempted). The number of reported *ExtQ-ExtDB* hits (at Tanimoto ≥ 0.75), and the physical time taken to complete the calculations (deployment waiting times included) have been monitored with respect to several maps, found to be optimal or near-optimal during the benchmarking work. Some of these maps were also used at variable ρ values, in order to check in how far the choice of ρ in order to optimize Q as outlined in equation 1 actually selects a radius values which performs well in this real-life test.

3. Results and discussion

3.1. Monitoring Map Convergence

Map training is an iterative process through time. It is quite time consuming and requires a lot of computational effort. While training, the software learns the code vectors of the neurons from sample vectors of the input data. The only stopping criterion is the number of steps, decided by the user at the beginning of the training process. The question addressed here is how the tuning process of the code vectors affects the VS Enhancement propensities of maps. These are different form the actual training objective (quantization error) - yet, it is clear that a poorly trained map (*i.e.* with almost random code vectors) may not significantly enhance the VS process. With a random map, the retrieval rate of similar compound pairs should linearly scale with the number of compound pairs subjected to explicit Σ value calculations, since there is no meaningful grouping of related compounds on a same or on neighbouring neurons.

Map convergence depends both on the training set size and the map size. Figure 4 shows the $RR - f$ plots (in terms of Euclidean-driven VS, $\Sigma = E$) for a succession of maps trained on the *SmallRef* set, of dimensions 22x28, rectangular topology and bubble neighbourhood function. The first map in the series was obtained after 1000 steps of training from a randomly initialized set of code vectors, whereas successive maps were each obtained by further training of their predecessor, by the indicated number of steps: the second map evolved from the first after further 10,000 training iterations (thus accumulates a total of 11,000 training steps,*etc.*). This 11K step map is the starting point of further hyperrefinement runs of various length (50, 80, 100 and 200 thousand steps). The map labelled '50 K' in Figure 4 accumulates thus a total of 61,000 training steps, *etc.* In terms of the map training strategies outlined in section 2.3.2, this map is the product of three successive steps: Brute training, Refinement and HyperRefinement. Other maps correspond to even longer HyperRefinement runs. This refitting scheme starting from a common 'ancestor' ensures that the relative behaviour of successive maps is purely due to the fitting process, and not due to the stochastic choice of initial code vectors.

This Figure shows that, with *SmallRef*, the convergence is quickly obtained. The Brute training step is insufficient, leading to relatively large and not very homogeneous nodes. However, the following Refinement run significantly improves the $RR - f$ characteristic of the map, pushing it into the near-optimality area. Relevant nodes hosting query compounds shrank (at $\rho = 0$ only 64% of expected hits were found to reside on the same neuron of the query). Yet, after extending the scope of the search to $\rho = 1$ (query neuron and its neighbours), more than 80% of hits are covered, and more effectively than at $\rho = 0$ of the Brute map. A slight improvement is still witnessed after HyperRefinement. However, further pushing of the fitting process at 80 or 100 thousand additional steps does no longer bring any improvement - on the contrary, high retrieval rates seem to become relatively more expensive in terms of computational effort (larger neuron radii needed). Apparently, fitting starts by assigning the few gross families of compounds to some neurons of the map, while all the others

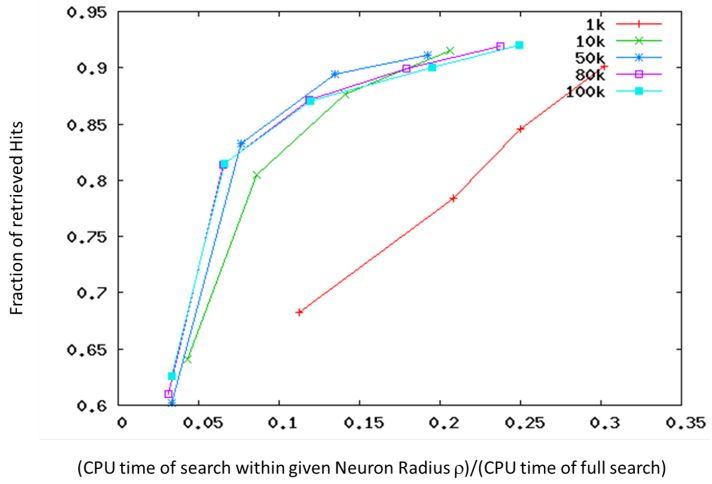

Figure 4: Retrieval Rate - Time fraction ($RR - f$) curves after different training steps, of a 22x28 map, Bubble neighbourhood with the *SmallRef* dataset

are empty (it is highly unlikely to generate, at the random initialization step, a 4418-dimensional code vector which by chance strongly correlates to an existing pharmacophore pattern in a molecule. Only a dwindling minority of possible code vectors encode chemically meaningful code vectors, anyway). Further refinement aims at splitting the gross families on the few populated neurons into more specifically defined subfamilies, to be hosted on the so-far empty neurons in the vicinity of the original attractor.

Over-fitting artefacts are even stronger when the larger training set *Extended* is employed. Figure 5, first of all, shows that achieving convergence is globally more difficult with larger sets: Brute+Refinement steps alone are not yet able to push the $RR - f$ curve into the optimality zone. HyperRefinement at 50K aditional steps is necessary to achieve this. More aggressive refinement, notably at 200K additional steps, clearly illustrates an unwanted 'upwards' bend of the $RR - f$ curves, signalling a loss of the ability to effectively retrieve the nearest neighbours.

The distribution of the molecule populations in the neurons of the corresponding maps in given in Figure 6, and supports the same explanation suggested during the discussion of *SmallRef*-trained maps: it can be seen that more aggressive map training results in a steady homogenization of population sizes allocated to every neuron. At a certain point, however, this homogenization appears to be artefactual, and no longer match the 'natural' compound family sizes found in the data set. Too much refinement is not necessarily beneficial. First, different gross families may be spread out over zones of different sizes, which makes the choice of the optimal ρ value a frustrating exercise: low ρ is effectively dealing with families that were not dispatched over large areas in

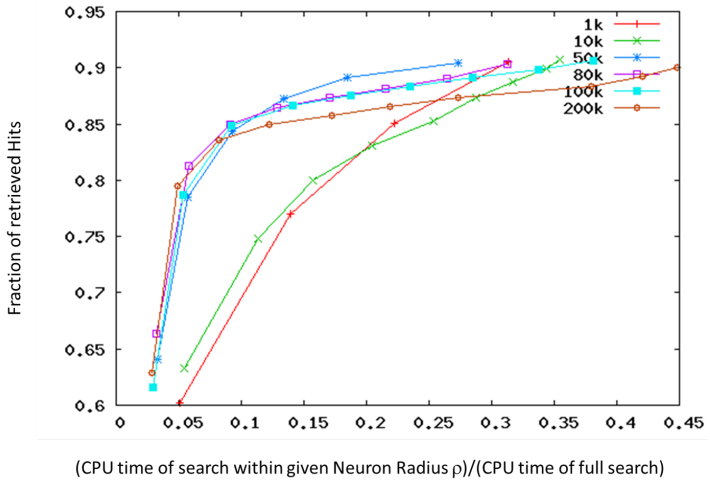

Figure 5: $RR-f$ curves after different training steps, of a 22x28 map, Bubble neighbourhood with the *Extended* dataset

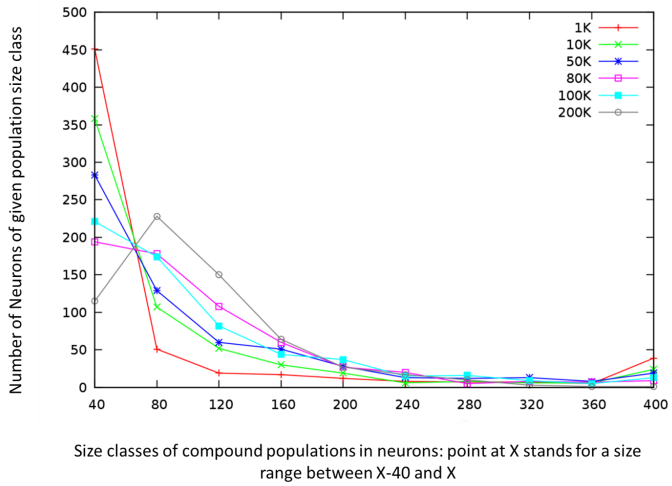

Figure 6: Count of neurons (on Y), out of the total of 22×28, hosting a number of compounds within the given X range: between 0 and 40 (first point), 40 to 80, 80 to 120.. and more than 400 (rightmost point)

spite of continued fitting. High ρ values are a must in order to ensure decent retrieval rates within families that have been dispatched over many neighbouring neurons, but are time-wasting choices with respect to the queries in 'localized' subfamilies. At a certain point in the fitting process, the selective spreading of specific families over much larger zones, while localized families remain confined

to a restrained set of neighbouring neurons, may therefore cause an overall loss of efficacy of the map as a generic Virtual Screening enhancer - as observed here. Interestingly, albeit SOMs are unsupervised learning methods, and albeit the fact that the minimized criterion is the quantization error, not the quality of the $RR - f$ characteristic, the typical signature of over-fitting artefacts can nevertheless be evidenced in terms of $RR - f$ quality.

In parallel the Quantization error monotonously decreases upon more aggressive training: from 11.61 for the 'Brute' map (1k), to 11.50 for 'Brute+Ref' (10k), to 10.13 at HyperRefinement (50k), 9.93 (80k), 9.62 (100k), 9.51(200k). As can be seen the most significant quantization error decrease (during the additional 50k steps of HyperRefinement) is also matched by the most significant improvement of the $RR-f$ curve. However - and expectedly, since quantization error is the objective function minimized at SOM training - quantization error cannot signal over-fitting.

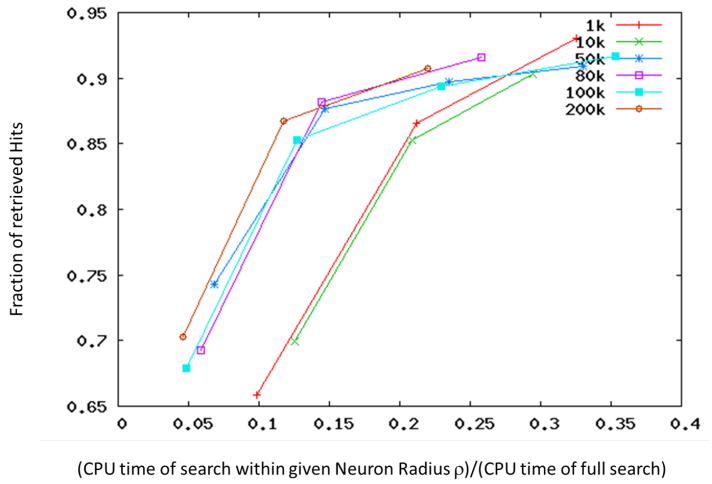

Figure 7: $RR - f$ curves, after different training steps, of a 10x10 map, Gaussian neighbourhood with the *Extended* dataset

However, convergence of maps seems to be heavily map size dependent. Unlike the previously shown results, fitting (Figure 7) of the much smaller 10x10 map (Gaussian neighbourhood function) does not seem to follow the same pattern. Brute and Reference training also appear insufficient in this configuration (the additional 10k Ref steps actually seem to slightly decrease map quality). Further training, however, fails to reveal clear signs of reaching an over-fitted configuration and, by contrast to the previous pattern of the evolution of the neuron population levels (Figure 6), aggressive fitting of the 10x10 map actually seems to favour apparition of massive nodes - see Figure 8 -, while sparsely populated nodes are rare (they do exist, though, and particularly in the Brute 1k map). *Per se*, this behaviour is not surprising: the more degrees of freedom

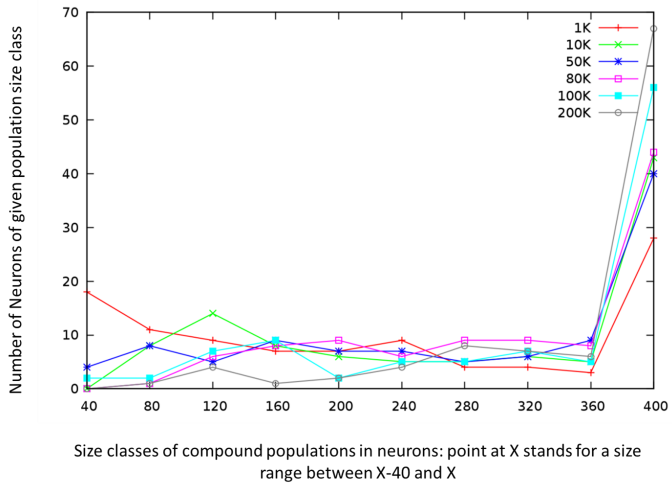

Figure 8: Count of neurons (on Y), out of the total of 10×10, hosting a number of compounds within the given X range: between 0 and 40 (first point), 40 to 80, 80 to 120.. and more than 400 (rightmost point), in the map referred to in Figure 7

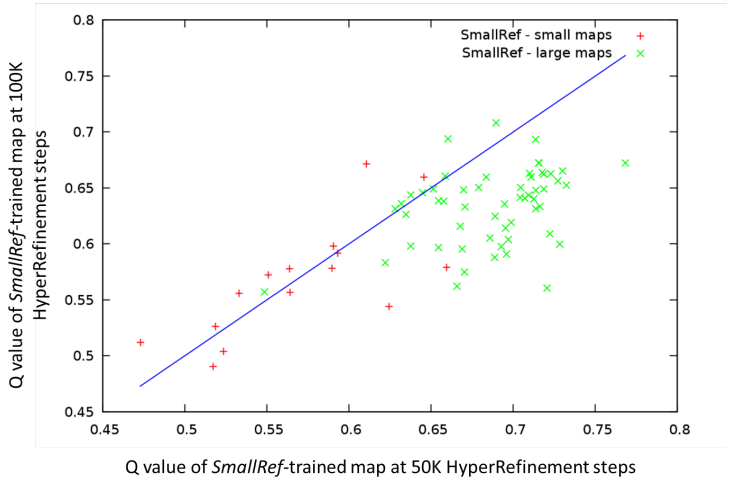

Figure 9: On X, the plot monitors Q factor of *SmallRef*-trained maps after a 50K Hyper-Refinement run against, on Y, the corresponding value for a map obtained after a 100K HyperRefinement task. Green dots correspond to "large" maps - of more of 200 neurons, red dots stand for small maps (from 48 to 200 neurons). The straight line is the diagonal: points below diagonal are over-fitting-prone maps

in a model (here: the more neurons in a map), the more likely it is prone to over-fitting artefacts. Practically, though, it is not easy to predict (without

running this very time-consuming scan at successive training levels) whether a peculiar map, at peculiar geometry, is being over-fitted or not, after the HyperRefinement step - or whether, on the contrary, it might still benefit from more training. However, the general trend emerging from Figure 9 clearly evidences over-fitting of twice HyperRefined maps with respect to simply HyperRefined ones, the effect being overall visible for the SmallRef set, and quite specific for large maps, in agreement with the previous discussion.

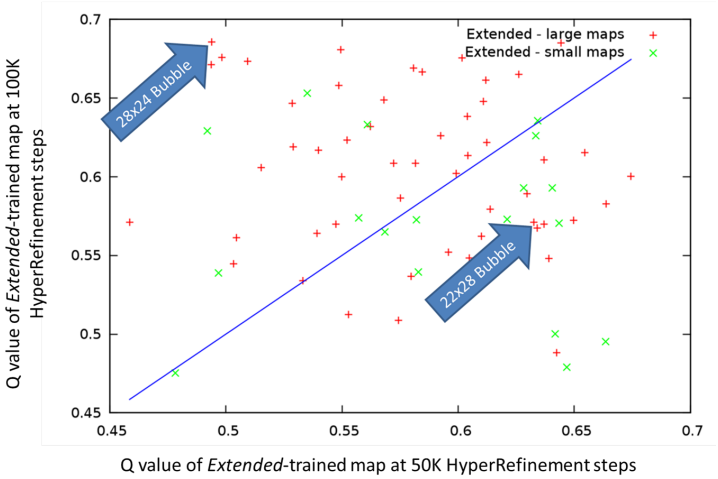

Figure 10: On X, the plot monitors Q factor of *Extended*-trained maps after a 50K HyperRefinement run against, on Y, the corresponding value for a map obtained after a 100K HyperRefinement task. Green dots correspond to "large" maps - of more of 200 neurons, red dots stand for small maps (from 48 to 200 neurons). Arrows pinpoint the behaviour of two very similar maps which nevertheless differ significantly: the upper significantly improves at 100k, while the lower suffers from over-fitting after 50k HyperRefinement steps

When using the Extended training set, however, the optimal training effort appears to display (Figure 10) an almost chaotic dependence on the precise map geometry: sometimes HyperRefinement at 100k steps triggers significant over-fitting artefacts over 50k steps only - as had been the case with the 22×28 map studied in detail previously, see Figure 5. However, the opposite scenario seems to occur equally often: 50k HyperRefinement steps may as well be insufficient. Note that the most striking example thereof occurs with the 28×24 Bubble map - a a very similar setup to the above-discussed 22×28 case study.

In the following, for each map configuration, fitting will be performed either 50k or 100k HyperRefinement steps, and the best map in terms of Q criterion will be taken to 'represent' the specific setup.

3.2. Impact of the training set size. Top performance maps.

The study of map convergence has already shown that a training set size increase does not only render convergence more difficult to achieve, but also

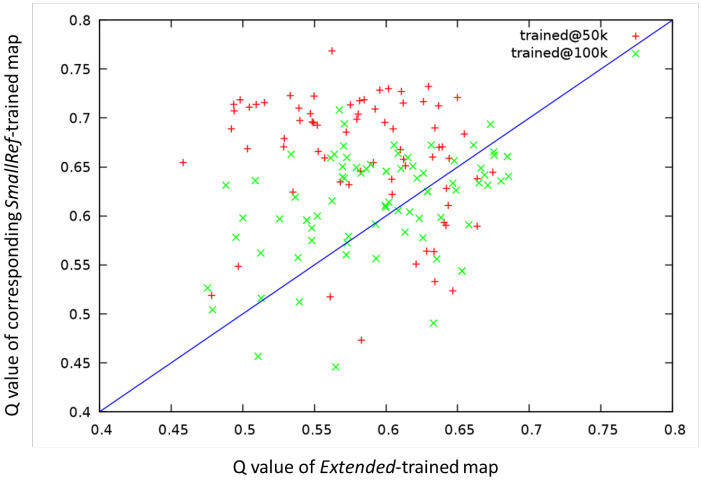

Figure 11: On X, the plot monitors Q factor of *Extended*-trained maps against, on Y, the corresponding value for the equivalent map built on hand of *SmallRef*, all other parameters being equal. Green dots correspond to aggressively trained maps at 100k HyperRefinement steps, by contrast to 50k step-maps rendered as red points.

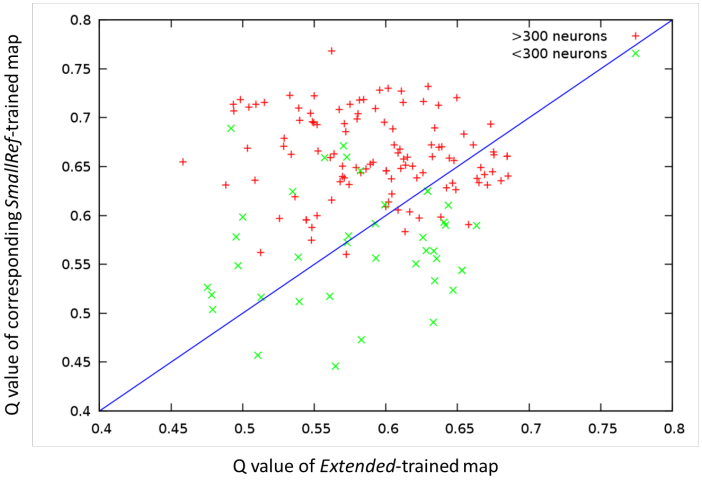

Figure 12: Same plot as in Figure 11, but colour coded by map size: in red, large maps with 300 neurons or more, in green, small maps

erratically affects the position of the borderline between optimal training and over-fitting. However, how does the set size affect absolute map quality? Let Q_s represent the Q score for a map trained on the *SmallRef* set - plotted on Y in Figure 11, and Q_e (on X) its counterpart of the *Extended*-trained map. The

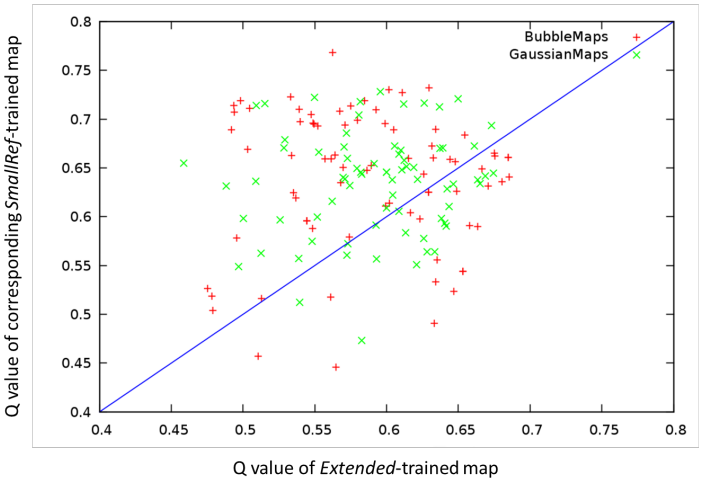

Figure 13: Same plot as in Figure 11, but colour coded by map neighbourhood function

strong prevalence of above-diagonal points in Figure 11 clearly shows that, all other things being equal, maps trained on *SmallRef* tend to be more potent VS enhancers than those built on the basis of *Extended*. All in all, the dominant maps at $Q_s > 0.7$ are all based on *SmallRef* and built with 50k HyperRefinement steps. By contrast, no setup whatsoever, all sizes, neighbouring functions and training strategies confounded, led to an *Extended*-trained map of $Q_e > 0.7$. Also, Figure 12 proves that the dominant maps are all large (with more than 300 neurons). By contrast, neither one of employed neighbouring function (Gaussian *vs.* Bubble) displays any dominance in terms of associated high quality maps (Figure 13).

It may seem puzzling that increasing the training set size should result in a global decrease of map performances, albeit the *Extended* training set actually makes up for a large majority of the DB used in the similarity screening simulation producing the Q criteria. Yet, Kohonen nets are unsupervised learning algorithms: more input information does not necessarily lead to maps of improved Q scores. Furthermore, the non-linear training procedure is prone, like any complex objective function minimizations, to risks of being trapped in local minima, *etc.* In light of the insights provided by the detailed study of convergence, it can thus be said that any potential benefits stemming from the supplementary information provided by *Extended* are being cancelled by the increased difficulty of map training in presence of more input molecules. This notwithstanding, please note that the so-called *SmallRef* set is already a consistent collection of eleven thousand diverse and relevant molecules, thus some two order of magnitude larger than a typical structure-activity learning set. The flattening out and eventual decrease of the map performance with training set size allegedly happens somewhere within the range between 10 and 50 thou-

sands compounds. Albeit this work does not furnish a formal proof, the further reducing of the training set below *SmallRef* is not likely to witness a further grow of map quality. Our argument in this respect is the coherent behaviour of *SmallRef*-trained maps in the convergence study: at 50k HyperRefinement steps, these seem indeed to reach an optimum - thus fully exploit all the chemical information in the training set. *SmallRef* does hence not appear to be as large as to jeopardize map convergence, by contrast to *Extended*, for which the study showed a chaotic behaviour in the convergence study, and failed to highlight an optimal number of training iterations.

3.3. An overview of the best map

Amongst studied maps shown in Figure 12, the best quality criterion $Q = 0.77$ corresponds to a SOM trained on the *SmallRef* dataset, which contains $18 \times 20 = 360$ neurons and has been trained in three steps (Brute + Refinement + HyperRefinement of 50k iterations), with a rectangular topology and a bubble neighbourhood function.

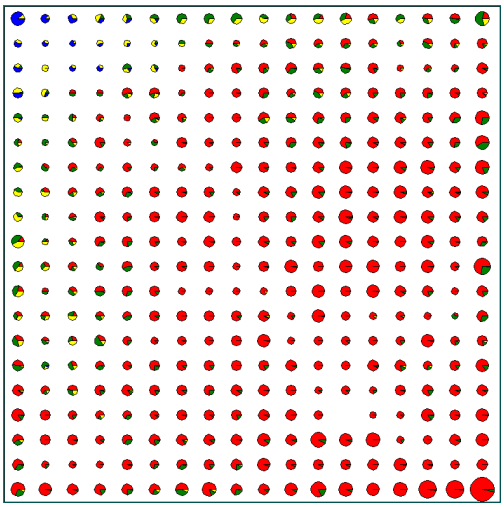

Figure 14: Repartition of compounds on the 18×20 rectangle bubble map. Neuron circle sizes represent the number of resident compounds. Neurons are coloured proportionally to the fraction of compounds at given number of Lipinksi rule violations. Red=0 violations, green=1, yellow=2, blue=3.Only one compound was found to violate all 4 rules and is therefore not visible.

Figure 14 visualizes the mapping of *SmallRef* on this SOM, where molecules have been colour-coded by the number of Lipinski [33] rule violations, in order to convey a general idea of the chemical space mapping quality. For a better view of , Figure 15 represents the same map, with neurons coloured in terms of the most often encountered number of rules violated by their resident compounds. The map coherently accounts for drug-likeness [34], with a clear drug-likeness gradient on the north-west (non-druglike) to south-east (drug-like), albeit it was

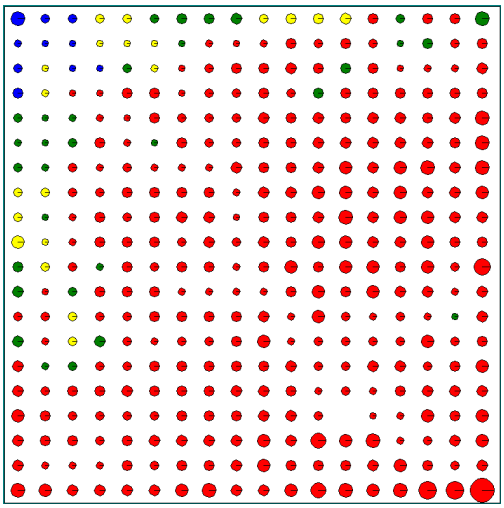

Figure 15: Repartition of compounds on the 18 × 20 rectangle bubble map. Neurons are coloured according to the majority class they display. As before, classes correspond to the number of Lipinksi rules violations. Red=0 violations, green=1, yellow=2, blue=3.

not trained by any means to account for drug-likeness (SOMs are unsuperwised learning methods, anyway). Furthermore, the map is relatively homogeneously populated by the *SmallRef* compounds, with only one empty neuron and a largest neuron (on the far bottom right) with 300 compounds (the smallest molecules in the set, of 100<MW<200), where 291 thereof break no Lipinski rules.

For a better view of the quality of this map, a set of 2170 active compounds from the DUD database ([29]) have been mapped on it and coloured according to their associated targets (see table 3.3). Figure 16 shows their repartition on the map. This DUD subset has been selected to regroup targets with a maximum of associated ligands. Unfortunately, nothing is known about potentially promiscuous ligands, which may bind to other targets out of the ten selected, in addition to the 'officially' assigned one. Since six targets out of then are kinases, for which the discovery of selective ligands is notoriously difficult (ATP-mimicking ligands will hit many kinases, for they all have a more or less well conserved ATP binding pocket), pharmacophorically similar compounds will reside on a same neuron, even if they are formally assigned to distinct categories with respect to the targets they bind. Ache, Fxa and Src ligands are well separated from the others. Src is a tyrosine kinase, represented by a set of specific inhibitors. Intriguingly, Cox2 ligands share a node ((18,6), the biggest with 263 compounds - more than 10% of the total) with p38 kinase ligands. Albeit cyclooxygenase 2 and the p38 kinase are functionally different, Cox2 also happens to recognize ligands with large aromatic moieties and hydrogen bond acceptors or anions. Indeed, the two ligand series are structurally very close,

especially as far as pharmacophore patterns can tell. Discovery of the right-hand p38 ligand by similarity-driven virtual screening, with the left-hand Cox2 compound of Figure 17 as a query, is an example of potentially promising 'lead hopping' - changing of scaffold while preserving the actual pharmacophore pattern. Whether the respective Cox2 and p38 ligands are actually binders to both targets is unknown to us at this point. All ligands are indeed in the "Lipinski-compliant" area of the map, with the bigger ones closer to the top left areas of the map.

A closer look at one heterogeneous neuron (11,6) shows that some of the depicted targets share structurally close ligands. This node regroups ligands binding respectively to dhfr, egfr, fxa, pdgfrb, vegfr2. (see Figure 18). Despite their different primary targets, it is visible that they share similar pharmacophoric patterns (according to our descriptors). Knowing that Dihydrofolate reductase readily binds heterocyclic bases and favours negatively charged compounds, like the kinases, this is actually not surprising at all - the most atypical resident of the node is the fxa inhibitor, which nevertheless features extended aromatic systems 'ornated' with several hydrogen bond acceptors (here, carbonyl groups instead of pyridine nitrogens) and donors (amide NH groups, instead of phenylamines). Again, the affinity of these ligands for the alternative targets is not known.

Factor X inhibitors are spread all over the map (but usually in nodes with a good purity). However, this is not a weak point of the mapping *per se*, but a general limitation of global similarity-based virtual screening: overall pharmacophore pattern similarity is by no means a *necessary* condition for two compounds to bind a same target (see detailed discussion of the Neighbourhood Behaviour problem [25]). Indeed, the necessary conditions for two compounds to be recognized by a same target is they possess key anchoring groups actually interacting with the receptor - *i.e.* a *binding pharmacophore*. But the overall pharmacophore pattern encoded by the fuzzy triplet fingerprint is a function of all putatively interacting groups - thus, imposing a strict similarity of the entire pattern is far too stringent. First, molecular moieties that never interact with the site may arbitrarily vary. Second, a protein site possess many putative interaction centres, and not all of these are systematically used to bind all the ligands. Ligands of comparable activities may exploit some different anchoring points - only few anchoring points are recurrently used by all the known actives, and these form the consensus binding pharmacophore. But two molecules may share a same consensus pharmacophore, yet be widely dissimilar - with respect to the other, much more numerous, irrelevant or occasionally relevant groups - which are accounted for in the overall pharmacophore pattern fingerprint, by definition. This is the case here (Figure 19): on one hand, the benzamidine group in the left-most fxa ligand is a hallmark of fxa activity, and is positively involved in the interaction with the active site. It also has a prominent role in shaping the overall pharmacophore pattern, as it is a carrier of a positive charge and of several H bond donors. However, it is not *compulsory*, as proven by the right-most fxa binder. The two extreme fxa ligands are beyond doubt *dissimilar* in terms of overall pharmacophore patterns: one is cationic and rich

Nb	Target	Nb of actives	Colour
1	ache (Acetylcholinesterase)	106	Red
2	cox2 (Cytochrome c oxidase subunit II)	409	Yellow
3	dhfr (Dihydrofolate reductase)	408	Light blue
4	egfr (Epidermal growth factor receptor kinase)	427	White
5	fgfr1 (Fibroblast growth factor receptor 1 kinase)	97	Grey
6	fxa (Factor X)	146	Pink
7	pdgfrb (Beta-type platelet-derived growth factor receptor kinase)	110	Gray blue
8	p38 (p38 mitogen-activated protein kinase)	342	Dark blue
9	src (Proto-oncogene tyrosine-protein kinase)	49	Dark red
10	vegfr2 (Vascular endothelial growth factor receptor kinase 2)	76	Green

Table 1: Label, targets and number of DUD compounds found for each target. The colors represent classes on the map.

in H-bond donors, the second is neutral (actually, as FPT descriptors are pH-dependent, at 7.4 the latter fingerprint captures some low contributions from the anionic species obtained by deprotonation of the sulfonamide at a predicted pH of about 8.0). No surprise, thus, to see these ligands on remote neurons, beyond the neuron radius of this map. Starting the search with the left-hand ligand as a query cannot find the right-hand molecule - and rightly so, not because of the map and its neuron radius, but because the pharmacophore dissimilarity score of these compounds is high: they would have not selected each other in all-pair-based similarity scoring, either. The middle ligand, however, is reasonably similar to the right-hand molecules (and their residence neurons are adjacent) - yet, it is still far from the left-hand benzamidine, although the o-aminopyridine moiety (shown as charged) is a bioisostere of benzamidine. However, due to pH-senstivity of the fingerprints, in actual FPT calculations the charged form as shown in the Figure contributes only roughly 50 %, whereas the neutral one (with the pyridine N as an acceptors) is equivalently important: the pK_a of this o-aminopyridine is estimated at roughly 7 by the ChemAxon plugin [35] called by the FPT generator. True, a medicinal chemist may decry the failure to pick the middle ligand starting from the benzamidine as a query, but this is not due to employing the map as VS enhancer, but again due to the intrinsically high dissimilarity of pharmacophore patterns. Rule-based pharmacophore flagging (stating that both benzamidine and o-aminopyridine are charged, thus equivalent) would have perhaps been more satisfactory for the end user - in this situation. In many others, apparently insignificant structural changes triggering important pK_a shifts will immediately result in puzzling *activiy cliffs* [36] unless pH-dependence of the descriptors is not accounted for [30].

3.4. Real-life virtual screening enhancement tests and benchmark

A preliminary real-life test of VS enhancement has been performed on a subset of 4 maps selected for their good *a priori* performances - at the point of this undertaking, which preceded the discovery of the best map discussed

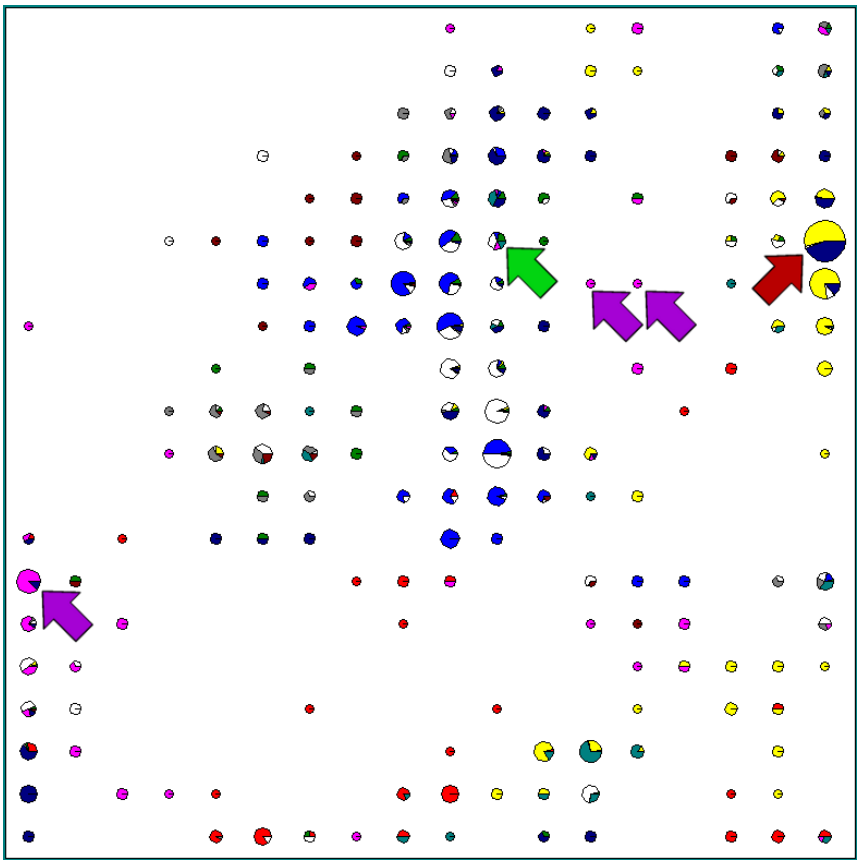

Figure 16: Mapping of the 2170 DUD compounds on the 18 × 20 rectangle bubble map. Neurons are coloured according to the class they display. See table 3.3 for colors. The red arrow points to neuron (18,6), the green arrow points to neuron (11,6) and the violet arrows point to neurons (1,14), (13,7) and (14,7).

above. The goal of this simulation was to double-check in how far the actual VS enhancement propensities of these maps match their relative ranking by the Q criterion - *i.e.* whether Q-based design of maps will lead to effective VS enhancement tools.

The four selected maps, which all displayed Q values between 0.5 and 0.7 were :

- top1 : 22 × 28 rectangle bubble trained on the SmallRef set
- top2 : 20 × 40 rectangle gaussian trained on the Extended set
- top3 : 28 × 30 rectangle bubble trained on the SmallRef set
- small : 6 × 6 rectangle bubble trained on the Extended set

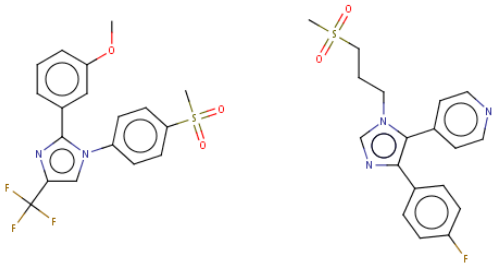

Figure 17: Left : a Cox2 inhibitor, and right a P38 inhibitor of the DUD data set. Albeit binding to different targets, they clearly display a highly similar overall pharmacophore pattern and reside both in neuron (18,6) - red arrow in Figure 16

Each of the four maps was used to confront the *ExtQ* set to the *ExtDB* compounds. In addition, for the map top2, the virtual screening was conducted with various neuron zone sizes, below and above the optimally established $\rho = 3$. For each run, we monitored the effective physical time the virtual screening took to complete, as well as the number of detected pairs of Tanimoto dissimilarity below 0.25 (i.e. the number of retrieved hits).

3.4.1. Hit Rate Analysis

The total number of (*ExtQ*, *ExtDB*) compound pairs within 0.25 of Tanimoto dissimilarity is not known, since the cumbersome systematic comparison of $\sim 2 \times 10^9$ fingerprint pairs has not been attempted. Note - the retrieved hits are unequally distributed among the external compounds: if the "average" external compound turns out to have slightly above 2 neighbors among the ExtDB molecules, this is because some 7000 of the 12k do not have any hit at all, while 1500 feature 10 hits or more (only the 10 top hits are returned and counted, anyway). This is not surprising, given the ubiquity of series in corporate databases: compounds that are related to an in-house motif are more likely to return many hits, for there are multiple "incarnations" of that motif. The tool therefore does detect diversity holes. (see figure 20).

3.4.2. Map-dependent Hit Rate: how many of the pairs of neighbours escape detection when using maps for acceleration, and what time gain does one get in compensation?

The optimality criterion Q used to choose these maps, as a "best" compromise between speed-up and hit loss (while taking care that hit loss does not exceed 15%) may not be a direct quality indicator for a real-scale experiment because (a) the sizes and the nature of the involved sets are not the same - the peculiarities of compound distribution in a corporate database were not accounted for, and (b) the real-scale experiment has been parallelized on multiple (4) CPUs, thus biasing time gain measures over the original map optimality.

The table 3.4.2 displays, for each map - at nominal or variable neuron radius

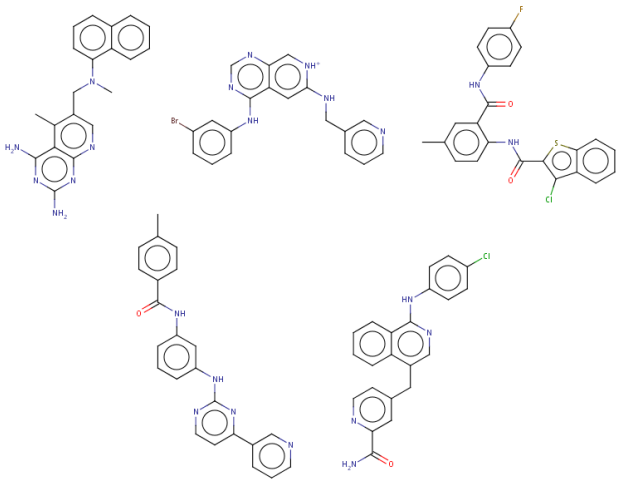

Figure 18: Example of five compounds found in the heterogeneous node (11,6) - green arrow in Figure 16. From left to right : dhfr inhibitor, egfr inhibitor, fxa inhibitor, pdgfrb inhibitor, vegfr2 inhibitor.

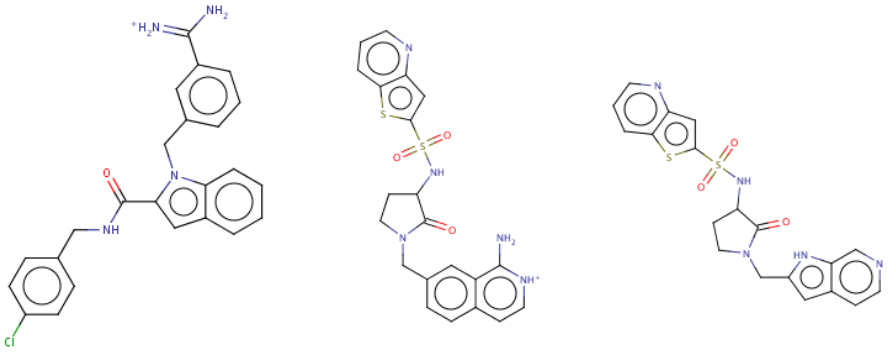

Figure 19: Comparison of three fxa ligands residing in three different neurons (from left to right : neuron(1,14), neuron(13,7), neuron(14,7) - violet arrows in Figure 16).

ρ, the number of retrieved hits and the time it took to complete the screening.

It is clear from above that the initial "small" 6x6 bubble map was the best in detecting similar pairs (the baseline number of which is not known, albeit the scanning of ρ with the map top2 suggests that convergence at full map coverage should stabilize this number somewhere not far beyond 30000). This is achieved in a relative good time of less than 3 hours, whereas bigger - thus finer - maps such as top2 witness an aggressive increase in computer effort at increased ρ. Conclusion: similar compounds which, for whatever reasons, are

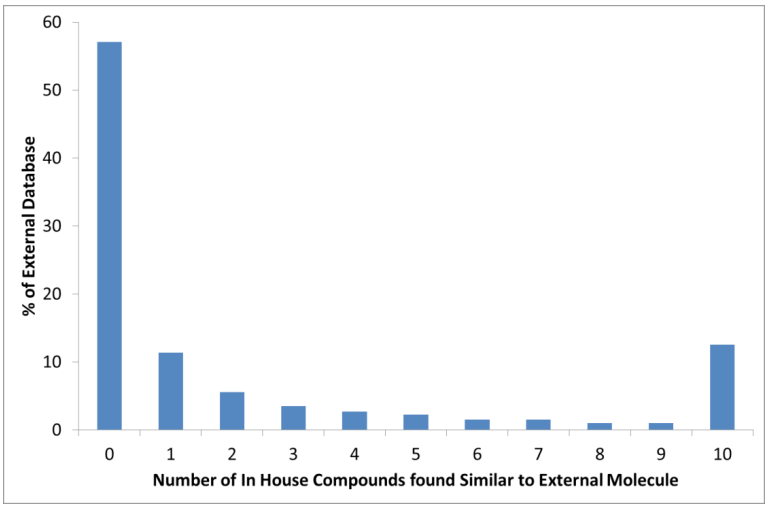

Figure 20: Percentage of ExtQ (on Y) found to have N (on X) near neighbours among in-House compounds (ExtDB). The almost 60% at "0" are brand-new pharmacophore patterns, not present in the corporate database ExtDB. The others may still be potentially interesting, if they are original from a scaffold-centric view. Distribution obtained with map "top2", where the average number of neighbours/external compound is of 2.17

not located on close-by neurons, are at risk to be dispatched anywhere in the map - retrieving them by increasing R may be very costly. In terms of the speed-up, solutions top1 and, in a lesser extent, top3 are very satisfying. All in all, the good behaviour of the maps as evidenced at their primary benchmarking stage (2000 external compounds (QS) x 53000 playing the role of 'in-house molecules' (DB)) was confirmed in this real-scale experiment.

As hinted by the primary benchmarking, the optimal neuron zone size ρ for top2 is indeed 3, and the top1 and top3 maps have clear speed-up advantages over the "small" solution. The second-best ranked map top2 was slightly deceiving - albeit its best performances happened at the nominal neuron radius value as assigned by the Q-driven set-up process, it was outperformed by both top3 and small maps in the real-life VS simulation. Obviously, small Q variances as the ones between the 'top' maps are not relevant, neither is it reasonable to expect that Q may represent some universal quality criterion. While high Q may not guarantee excellent map performances under any arbitrary VS conditions, low Q proves a map to be a bad performer.

4. Conclusions

Using Kohonen Self-Organizing Maps is an effective way to accelerate similarity searches in a database of small compounds. The acceleration tests performed on 57613 compounds, using the first 2000 as query, have proven that mapping the molecules on a SOM can considerably accelerate similarity searches

map	Neuron radius ρ	# Detected similar pairs	Time (mins)	Q
small	1	29718	158	0.51
top1	1	27107	65	0.71
top2	1*	24588	74	-
top2	3	27096	139	0.70
top2	5*	27707	260	-
top2	10*	28571	597	-
top3	2	27837	96	0.69

Table 2: Results of real-scale inter-set similarity assessment. Neuron radii values labelled by * were checked out for testing purposes, and do not represent the nominal values associated to that map. The last column reports the Q factor of the map, based on the benchmarking study.

without significant losses of virtual hits, as proven by high quality scores Q, specifically developed to the purpose of synthetically capturing the compromise between speed-up and hit loss. The best maps may retrieve about 90% of the relevant neighbours of the query in about 10% of the total time required to scan the entire database.

During the training phase, attention should be focused on the convergence of the maps. Failure to converge results in suboptimal performances, but - somehow surprisingly for an unsupervised learning method - excessive map training was shown to lead to over-fitting artefacts. There seem to be no unequivocal rules on how to establish the optimal number of training iterations - convergence behaviour being notably determined by both map geometry and training set size. A gradual training strategy, with intermediate checks of VS enhancement scores, is advised in order to discover an optimal map.

Furthermore, care is advised while choosing the training set (which must be representative of the targeted chemical space: here - drug-like molecules). Two training sets have been compared, the *SmallRef* (11168 compounds) and the *Extended* (53206) set. The smaller training set is sufficient to create maps that depict correctly the chemical space of our database. Surprisingly, the employment of *Extended* training set was not only unhelpful to further increase map performance, but often detrimental (its specific impact varied in function of map geometry). Too many training molecules tend to render convergence more difficult to achieve, and thus cancel out any potential benefits from the additional information they provide. Note, however, that the 'small' set is already a significant and diverse compound collection.

The winner map of the Q-based benchmarking study has been visually rendered, with respect to its ability to monitor drug-likeness. A clear-cut separation of drug-like and non-druglike compounds can be observed along its diagonal, showing that this is a potentially meaningful chart of medicinal chemical space. Furthermore, mapping of diverse series of binders to 10 different targets (both inter-related kinases and widely differing enzymes) lead to a coherent and sensible picture, highlighting no inherent weakness of the map as such, but rather

the well-known pitfalls and limitation of global similarity-based search of active analogues. Given that map training was never oriented towards discrimination of activity classes, and that the DUD reference binders were never employed at any stage of training or map selection, this is a remarkable result underlining the excellent propensity of the Q-score to recognize meaningful chemical space maps.

Eventually, real-life performance tests have shown that the maps may successfully stand the challenge of scaling down the effort of a 2×10^9-fold matching of 4000-dimensional fingerprints to hardly more than one hour on a 4-CPU workstation, without dramatic losses of virtual hits.

[1] P. Lyne, DDT 7 (2002) 1047–1055.

[2] M. Johnson, G. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley Interscience, 1990.

[3] P. Willett, Annu. Rev. Inform. Sci. 43 (2009) 1–117.

[4] P. Willett, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.

[5] E. Lipp, Genet. Eng. Biotechn. N. 28 (2008) 22.

[6] A. Smellie, J. Chem. Inf. Model. 49 (2009) 257–262.

[7] J. Bentley, Communications of the ACM 18 (1975) 509–517.

[8] A. Gionis, P. Indyk, M. R., VLDB '99 Proceedings of the 25th International Conference on Very Large Data Bases (1999).

[9] T. Kohonen, Biol. Cybernetics 43 (1982).

[10] T. Kohonen, Proc. IEEE 78 (1990).

[11] G. Schneider, P. Wrede, Prog. Biophys. Mol. Biol. 70 (1998) 175–222.

[12] J. Polanski, J. Gasteiger, Acta Pol. Pharm. 56 (1999) 112–122.

[13] S. Anzali, W. Mederski, M. Osswald, D. Dorsch, Bioorg. Med. Chem. Lett. 8 (1997) 11–16.

[14] S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, S. J., A. Wagener, Perspect. Drug Discov. 9-11 (1998) 273–299.

[15] D. Hristozov, T. Oprea, J. Gasteiger, J. Comput. Aided Mol. Des. 21 (2007) 617–640.

[16] D. Hristozov, T. Oprea, J. Gasteiger, J. Chem. Inf. Model. 47 (2007) 2044–2062.

[17] T. Aoyama, Y. Suzuki, H. Ichikawa, J. Med. Chem. 33 (1990) 905–908.

[18] J. Polanski, Adv. Drug Deliv. Rev. 55 (2003) 1149–1162.

[19] J. Polanski, K. Jarzembek, J. Gasteiger, Comb. Chem. High Throughput Screen. 3 (2000) 481–495.

[20] P. Selzer, P. Ertl, J. Chem. Inf. Model. 46 (2006) 2319–2323.

[21] J. Gasteiger, X. Li, Angew. Chem. Int. Ed. Engl. 33 (1994) 643–646.

[22] J. Sadowski, M. Wagener, J. Gasteiger, Angew. Chem. Int. Ed. Engl. 34 (1995) 2674–2677.

[23] D.-J. Im, M. Lee, Y. Lee, T. Kim, S. Lee, J. Lee, K. Lee, K. Cho, Lect. Notes Comput. Sc. 3481 (2005) 334–342.

[24] K. Oh, A. Zaher, P. Kim, Lect. Notes Comput. Sc. 2383 (2002) 131–151.

[25] D. Horvath, C. Jeandenans, J. Comput. Inf. Comp. Sci. 43 (2003) 680–690.

[26] F. Bonachera, D. Horvath, J. Chem. Inf. Model. 48 (2008) 409–425.

[27] E. Bolton, Y. Wang, P. Thiessen, S. Bryant, Annual Reports in Computational Chemistry 4 (2008).

[28] J. Irwin, B. Shoichet, J. Chem. Inf. Model. 45 (2005) 177–182.

[29] N. Huang, B. Shoichet, J. Irwin, J. Med. Chem. 49 (2006) 6789–6801.

[30] F. Bonachera, B. Parent, F. Barbosa, N. Froloff, D. Horvath, J. Chem. Inf. Model. 46 (2006) 2457–2477.

[31] J. Kornhuber, L. Terfloth, S. Bleich, J. Wiltfang, R. Rupprecht, Eur. J. Med. Chem. 44 (2009) 2667–2672.

[32] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, Report A31 (1996).

[33] C. Lipinski, F. Lombardo, B. Dominy, P. Feeney, Adv. Drug Delivery Rev. 23 (1997) 3–25.

[34] J. R. Proudfoot, Bioorganic amp; Medicinal Chemistry Letters 12 (2002) 1647–1650.

[35] ChemAxon, pka calculator plugin, http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html, 2007.

[36] G. Maggiora, J. Chem. Inf. Model. 46 (2006) 1535.

Onzième partie

Conclusion

Les 2D-FPTs apportent, par rapport à d'autres descripteurs pharmacophoriques flous 2D ou 3D, des améliorations majeures ayant du sens chimiquement parlant. De ce fait, les calculs de similarité basés sur les 2D-FPTs sont souvent plus efficaces.

La projection floue des triplets moléculaires sur l'ensemble des triplets de base permet de prendre en compte la tolérance naturelle des récepteurs par rapport aux ligands. En effet, un site de liaison tolère l'insertion ou la délétion de liaisons dans les analogues de son ligand. Ceci est traduit par la variabilité tolérée des distances topologiques entre les groupes pharmacophoriques lors de la projection des triplets. De plus, d'un point de vue pratique, la dimensionalité des descripteurs peut etre significativement réduite grâce à cette technique, en choisissant les triplets de base de manière appropriée. Il est ainsi possible de restreindre le jeu de référence à quelques milliers de descripteurs par rapport à plus de 50.000 descripteurs dans des empreintes binaires.

La prise en compte du pK_a lors de la pondération des pharmacophores permet de corriger certaines incohérences trouvées lors d'un marquage basé sur des règles et de rendre les descripteurs chimiquement plus pertinents. Bien entendu, une prédiction correcte des états de protonation est pré-requise pour assurer le succès de cette méthode (d'où l'utilisation dans nos travaux du calculateur de pK_a de Chemaxon). Des changements locaux de substituants n'entrainent pas de changements dans le modèle pharmacophorique dans le cas d'un marquage basé sur des règles. Cependant, ces changements peuvent faire varier les valeurs de pK_a au delà du seuil de pH et entrainer ainsi des changements conséquents dans l'activité du composé à l'équilibre. Cet état de fait est retranscrit dans les 2D-FPTs par des changements de population des triplets. Ainsi, dans l'espace des 2D-FPTs, nous avons pu gérer l'une des limitations de la similarité basée sur des pharmacophores : le problème des aspérités du paysage. Il a été démontré dans nos travaux que le fait que les espaces de descripteurs classiques ne prennent pas en compte les changements d'équilibre protéolytique peut expliquer certains artefacts : des "pics d'activité" observés dans le paysage des relations structure-activité.

Le score de similarité développé dans nos travaux a été comparé aux critères "classiques" de similarité ou de dissimilarité (Euclidienne ou Dice). Ces scores sont génériques et peuvent être appliqués dans des espaces indépendamment de la nature réelle des descripteurs moléculaires utilisés. Les schémas de score classiques n'étant pas capables de traduire les contributions des triplets en fonction de leur type (partagés, nuls ou exclusifs), il était donc important de mettre en place un score nouveau qui soit basé sur l'interprétation de l'information apportée par nos descripteurs. L'idée simple et logique à la base de notre nouveau score était de séparer la prise en compte des contributions des triplets partagés, exclusifs ou encore nuls entre deux composés. Un triplet absent dans les deux composés sera donc considéré plus faiblement qu'un triplet présent dans les deux - ce qui entrainera dans ce cas la similarité.

Dans cette première étude, le comportement au voisinage (*NB - neighbourhood behaviour*, [48]) des FPTs (tout comme d'autres descripteurs basés sur des pharmacophores) a été de bonne qualité, à la fois sur des ensembles de composés divers (*BioPrint*) mais aussi sur des ensembles contenant plusieurs séries d'analogues. Dans cette dernière situation, il est généralement plus facile de voir un bon comportement au voisinage des descripteurs car une discrimination simple entre les chémotypes principaux trouvés sur la base de la série d'analogues suffit. Il faut cependant préciser que les conclusions apportées par ce type d'études peuvent être biaisées par différentes sources : La taille relative des composés, la complexité chimique ainsi que d'autres particularités de la série d'analogues considérée.

Il était plus complexe de rechercher en profondeur la similarité pharmacophorique dans des séries contenant peu de représentants pour chaque châssis moléculaire, mais la méthodologie des 2D-FPTs a permis d'accomplir ces recherches par similarité avec succès.

Ce bon comportement doit être cependant nuancé au vu d'une étude récente ([46]). Dans ces travaux, le comportement au voisinage local des FPTs ainsi qu'un certain nombre d'autres descripteurs a été étudié sur 2500 composés provenant d'une base de données combinatoire de ligands de protéases (Tryptase, Chymotrypsine, Facteur Xa, Trypsine, UPA). Les FPTs ont été étudiés avec 4 configurations (FPT1 - la configuration par défaut avec prise en compte du pK_a au ph=7.4, FPT-noPK - ne se basant pas sur la prédiction du pK_a mais uniquement sur des règles "classiques" de marquage, FPT-ph1 - avec prise en compte du pK_a à un pH de 1, et FPT-ph5 - avec prise en compte du pK_a à un pH de 5.). Chacun de ces descripteurs ont été utilisés avec ou sans normalisation, et en conjonction avec 3 métriques de distances (**EUCLID** - Distance Euclidienne, **DICE** - Basée sur le coefficient de distance Dice, et **FDIFF** - La fraction des différences qui compte la fraction de caractéristiques différemment peuplées dans une paire de molécules (m,M).).

Cet ensemble de $4 \times 2 \times 3 = 24$ configuration des FPTs a été utilisé afin d'explorer leur qualité, c'est à dire, sélectionner (via SQS) des ensembles de descripteurs qui corrèlent de manière optimale avec l'affinité expérimentale de la Tryptase, lors d'une étude QSAR. Pour cela, les ensembles de descripteurs sont utilisés sur les ligands de la Tryptase mais aussi sur ceux des 4 autres protéases et chacun des résultats est analysé.

La conclusion de cette étude a indiqué que les FPTs sont retrouvés 4 fois sur 5 parmi les meilleurs descripteurs. Cependant, les résultats observés lors de notre étude n'ont pas pu être reproduits (c'est à dire, l'amélioration drastique apportée par la prise en compte du pK_a). Ici, les FPT1 et les FPT-nopK sont apparus comme étant des descripteurs aussi valides les uns que les autres. Leurs résultats sont équivalents, mais il est important de noter que la majorité des composés utilisés possèdent un seul groupement ionisable (ou plusieurs, mais éloignés). Les cas où les effets du pK_a sont visibles sont assez rares dans cette base de données.

Les 2 autres configurations (FPT-ph1 et FPT-ph5) sont par contre clairement dépassées par les configurations FPT1 et FPT-nopK.

En résumé, la prise en compte du pK_a dans le calcul des FPTs n'est pas forcément toujours un gage d'efficacité et est dépendante des données d'entrée. En effet, quel que soit l'espace de descripteurs utilisé, la réussite des expériences de criblage virtuel dépend fortement de la requête utilisée.

Après leur mise en place, les FPTs ont été utilisés dans des études de *QSAR*, en comparaison avec d'autres descripteurs (ISIDA – des descripteurs fragmentaux, CAX – des empreintes pharmacophoriques à deux points développées par Chemaxon). Le but était de mettre en place une procédure de test afin de comparer deux méthodes de sélection de descripteurs pour les études *QSAR*. Les études ont été réalisées sur 3 jeux de données de petite taille (CU (*Cyclic Urea derivatives*) - 118 composés, HEPT *(1-[2-hydroxy-(ethoxy)methyl]-6-(phenylthio)thymines) derivatives* - 93 composés, TIBO (*tetrahydroimidazobenzodiazepinones dérivatives*) - 84 composés).

Le *SQS* (Stochastic QSAR Sampler) est un échantillonneur QSAR basé sur un algorithme génétique original, piloté par un méta-algorithme génétique.

Les résultats obtenus sont les suivants :

Des simulations de construction de modèles *QSAR* ont été faites sur 75 combinaisons de 3 jeux de données $\times$ 5 manière de diviser ces données en ensemble d'entrainement et ensemble de test $\times$ 5 choix de descripteurs, en se basant sur des hypothèses de travail variées (recherches de modèles linéaires et non-linéaires). Nous avons pu de cette manière découvrir avec succès de multiples modèles de validation efficaces. Voici le résumé des observation que nous avons pu faire.

Reproductibilité Lors de la répétition d'un processus d'échantillonnage avec les mêmes paramètres que le processus précédent, la probabilité de retrouver un chromosome modèle déjà conservé auparavant est assez faible (de 0.03% à 6%). Le volume d'espace à échantillonner étant particulièrement grand, il est donc très peu probable de redécouvrir les mêmes modèles lors d'une répétition. De plus, le *SQS* ne pourra jamais lister l'ensemble des modèles *QSAR* possibles car il ne visite pas l'intégralité du volume de l'espace des phases du problème. Le volume d'échantillonnage le plus grand est obtenu avec les FPTs, qui présentent les taux de redécouverte les plus bas. L'ensemble des modèles produits (avec les 3 types de descripteurs) possède de bonnes performances de validation.

Cependant, la reproductibilité de la qualité des modèles basés sur les FPTs semble être dépendante des données d'entrée. Les modèles non-linéaires découverts pour CU montrent une excellente reproductibilité de la qualité, malgré les bas taux de redécouverte. Par contraste, l'échantillonnage des modèles TIBO présente des problèmes de reproductibilité significatifs. L'analyse plus en profondeur de ce problème a montré que la propension à la validation est dépendante de la manière de diviser les données. Ceci signale probablement la présente d'un triplet spécifique important dans certains composés.

Dépendance au volume d'espace des phases : non-linéarité et choix des descripteurs Il n'a pas été observé de perte de propension à la validation des modèles lors de l'augmentation du volume de l'espace des phases : ni l'introduction de transformations non-linéaires préétablies, ni le changement entre des ensembles de descripteurs petits ou grands n'ont déclenché d'artefacts d'*overfitting*. L'introduction de la non-linéarité a eu des effets positifs, les modèles non-linéaires entrainant de meilleures performances que les modèles linéaires.

Les modèles non-linéaires basés sur des FPTs ont là aussi des scores différents en fonction du jeu de données. Il est montré que la propension à la validation est liée à l'information chimique encodée par les descripteurs. En effet, les descripteurs ISIDA ont de meilleures performances sur les ensembles HEPT et TIBO, alors que les FPTs sont les plus efficaces pour construire des modèles de CU. Ceci s'explique par la nature des composés des jeux de données : la famille des urées cyclique est construite autour d'un large châssis moléculaire commun, ce qui la rend peu diverse. Les FPTs, capables d'encoder l'information spécifique des pharmacophores autour du châssis, sont ainsi plus indiqués pour décrire cet ensemble de données.

Construction d'un modèles stochastique ou pas à pas Le taux de succès de validation a servi de point de comparaison entre le *SQS* et le *SR*, sur les 3 jeux de données avec 2 types de descripteurs (ISIDA et FPTs). Pour les jeu de données HEPT, les modèles *SQS* sont aussi efficaces (voir plus, dans le cas des FPTs) à la validation que les modèles *SR*. La situation est différente sur l'ensemble TIBO : le SQS génère un ensemble de modèles très bien validés, malheureusement dépassé par des modèles non-validés. Les descripteurs FPTs sont les seuls qui entrainent, pour ce jeu de donnés, une famille d'équations non-validantes avec le *SR*.

Ainsi, alors que la *SR* (*Stepwise Regression*) peut échouer à produire des modèles de validation là où le *SQS* y parvient (ceci ayant été observé une fois sur six - HEPT/FPTs), cette dernière approche peut occasionnellement "cacher" les nombreuses équations valables au sein d'un ensemble encore plus large de non-validantes (HEPT/FPTs à nouveau). Au vu de ces résultats, il semble que les avantages du *SQS* sont peu nombreux par rapport aux hauts coûts en temps de calcul qu'il nécessite.

Stratégie consensus L'utilisation d'équations consensus est l'alternative au choix aléatoire de modèles uniques. Les modèles *SQS* consensus ont montré un comportement extrêmement robuste, en ayant virtuellement de meilleures caractéristiques que 70-90% des modèles individuels. Les performances du *SQS* et du *SR* sont une fois de plus très similaires malgré la procédure de construction de modèles consensus du *SR*. Les modèles consensus du *SQS* pourraient cependant posséder des avantages décisifs si l'on souhaite les utiliser pour le criblage virtuel de grandes bases de données.

Quelques remarques Si le but ultime d'une étude *QSAR* est de découvrir au moins une équation qui soit validée, alors le *SQS* ne semble pas présenter suffisamment d'avantage par rapport à l'effort nécessaire en temps le calcul pour le déployer. Il est vrai que le fait de trouver plus de modèles n'implique pas forcément d'en trouver des meilleurs, dans le sens des tests de validation standards intra-familles. Cependant, deux modèles bien validés peuvent néanmoins renvoyer des prédictions divergentes lorsqu'on les applique à des composés externes. En effet, les tests de validation, nécessaires mais non suffisants ne sont pas capables de dire quelle est l'équation physiquement interprétable parmi les nombreuses formes "apparemment" équivalentes (l'équivalence est ici désignée par rapport aux ensembles d'entraînement et de validation, pas forcément dans l'espace des composés *druglike*). C'est pourquoi peu d'études *QSAR* permettent d'aider à découvrir réellement des actifs par criblage virtuel. Une projection extensive de l'espace des problèmes du *QSAR* pourrait permettre de réduire les chances de ratés du criblage virtuel. L'échantillonnage de modèles par *SQS* pourrait ainsi être utilisé afin d'avoir une idée du degré de dégénérescence de l'information chimique dans le set d'entraînement et pourrait permettre de savoir quels composés nouveau il serait nécessaire d'ajouter à cet ensemble afin de lever certaines de ces dégénérescences.

Une grande étude de performance a été réalisée à la suite de ces observations. L'utilisation des FPTs dans des études de *QSAR* a donc été étudiée plus en profondeur, en comparaison avec des descripteurs déjà existants sur un ensemble de 11 jeux de données provenant de la littérature. De plus, nous avons comparé différentes versions de triplets entre elles : la version par défaut et l'optimale décrites toutes deux dans le premier chapitre, et une version *Coarse* plus grossière. Une version des FPTs ne prenant pas en compte le pK_a mais des règles d'attribution des pharmacophores a aussi été prise en compte. Deux manière d'assigner les triplets molécules aux triplets de base ont été étudiées, la manière floue habituelle et une assignation strict. Le but était de mettre en lumière quel type de FPTs peut être utilisé de manière optimale dans les études de *QSAR*.

Performance Les études *QSAR* basés sur les 2D-FPTs se sont extrêmement bien déroulées tout au long de cet exercice de performance. Les modèles basés sur les FPTs ont égalisé, voire dépassé significativement les modèles publiés basés sur l'index 2D et 3D, mais aussi les approches élaborées du *CoMFA*, basées sur la superposition.

Une exception notable a cependant été détectée : l'activité d'alkalysation de l'hème des analogues de l'artémisinine est, de manière non surprenante, la propriété biologique la moins bien traitée par le modèle pharmacophorique des triplets. Cette propriété est la seule étudiée qui ne reflète pas de procédé d'inhibition non-covalent de la cible, conceptuellement associé aux "pharmacophores de liaison".Les 2D-FPTs sont des descripteurs plus appropriés et riches en information pour décrire le processus de reconnaissance entre site et ligand.

Il ne faut cependant pas affirmer que les FPTs sont intrinsèquement plus informatifs que les champs CoMFA. En effet, les avantages observés pour les FPTs pourraient être expliqués par l'approche utilisée pour choisir les modèles : le SQS, lancé en parallèle et permettant du calcul intensif, aurait pu donner d'aussi bons voire de meilleurs modèles avec les descripteurs CoMFA. Il n'en reste pas moins que les FPTs sont capables de générer de très bons modèles *QSAR* en conjonction avec une méthode puissante de construction de ces modèles. Nous avons pu noter grâce à cette étude que les approches non-linéaires donnent de bons résultats en termes de validation, ce qui confirme les résultats du chapitre IV.

En résumé, les résultats obtenus grâce aux FPTs (version par défaut) sont très bons, sauf dans les cas où la nature même des composés n'est pas compatible avec ce type de descripteurs.

Le flou optimal L'emploi de la logique floue n'a pas d'impact significatif sur la performance du *QSAR*. Le flou, auparavant montré comme ayant des effets positifs sur le comportement au voisinage des FPTs, ne semble pas être essentiel pour la construction de modèles *QSAR*. D'un coté, le flou permet de simuler la tolérance des récepteurs vis à vis de leurs ligands. Cependant, le flou peut aussi entrainer une dégénérescence des descripteurs : plus le flou est grand, plus des triplets (qu'ils sont réellement impliqués dans la liaison ou non) ont de chance de contribuer à incrémenter la population d'un triplet de base donné. Ainsi, il y a un risque les signaux importants soient noyés sous le bruit des autres contributions. L'impact du flou est donc différent dans des études *QSAR* que dans des études de comportement au voisinage. Ceci pourrait poser des problèmes conséquents dans des études de *QSAR* basées sur la sélection des descripteurs (bien plus que lors de l'utilisation de scores de similarité).

Impact de la non-linéarité Nos travaux ont montré que, dans la majorité des situations, les modèles non-linéaires donnent de meilleurs résultats que leurs contreparties linéaires. Ils semblent ainsi être plus efficaces pour extrapoler aux données de validation l'information obtenue par les FPTs sur les jeux de données d'entrainement.

La représentativité Une bonne validation n'est pas une garantie suffisante de l'utilité réelle d'un modèle pour le criblage virtuel de composés choisis au hasard. Il est important de vérifier que les modèles soient représentatifs du mécanisme de liaison entre les ligands et la cible. Nous avons vu dans cette étude que les "pharmacophores topologiques" définis par les triplets entrant dans les modèles 2D-FPTs ne sont pas nécessairement caractéristiques des points d'ancrage entre le site et le ligand. En effet, nos travaux ont mis en exergue la portée très limitée des ensembles d'entraînement et de validation habituellement utilisés pour la mise en place et pour les tests de performance des études *QSAR*. A cause de cette limite, certains modèles se révèlent dépourvus de sens lorsqu'ils sont confrontés au divers composés hors du jeu de données utilisé. Il ressort que dans de nombreuses situations, le *fitting* et la validation du modèle *QSAR* sont basés sur des particularités spécifiques à chaque famille. Un des avantages du *SQS* est mis en valeur ici : sa capacité à créer de nombreux modèles lui permet de faire face à la confrontation aux molécules externes, ce qui n'est pas le cas d'autres méthodes qui produisent peu de modèles validés.

Un autre symptôme dû aux limitations des ensembles d'entraînement est la génération de modèles prédisant de hautes valeurs d'activité par défaut et se basant sur les termes pénalisants pour réduire le score des inactifs connus contenant des caractéristiques "non désirées". De même, des modèles prédisant de fortes valeurs d'activité pour n'importe quelle molécule trop petite pour contenir des triplets, qu'ils soient voulus ou non, n'ont pas de sens.

De plus, nous avons pu observer que les artefacts spécifiques aux jeux de données prennent le dessus sur les effets liés au pK_a. Bien qu'il ait été prouvé que le marquage pharmacophorique dépendant du pK_a soit parfois plus rigoureux que le marquage basé sur des règles, le schéma de marquage le plus efficace était souvent celui qui exploitait les coïncidences spécifiques à l'ensemble.

Comment faire face aux artefacts ? Nous avons rencontré dans cette étude un vaste nombre d'artefacts spécifiques aux jeux de données. Ceux-ci sont certainement apparus sous différentes formes avec les descripteurs utilisés dans les études citées dans la littérature. A la lumière des nombreux exemples d'équations chimiquement défectueuses qui pourtant passent brillamment les tests de validation "externes" - plus précisément contre de nouveaux membres de la famille d'entraînement – nos travaux suggèrent que :

- N'importe quel set d'entraînement devrait être complété par un ensemble divers d'inactifs (présumés) avant la mise en place de l'étude *QSAR*. Cela est facilement faisable avec les descripteurs 2D-FPTs ainsi qu'avec les autres descripteurs indépendants de la superposition, mais problématique avec *CoMFA* et les outils analogues.
- Lors des tests de performance, il faudrait ajouter un test supplémentaire, incluant des actifs topologiquement différents (c'est à dire, ayant des structures bidimensionnelles ne dérivant pas de la même structure que les actifs déjà présents dans l'ensemble d'entrainement). Le challenge de la prédiction d'actifs topologiquement différents est nécessaire, pour mettre l'accent sur eux parmi les nombreux modèles alternatifs apparemment redondants. Ainsi, es équations générales basées sur des termes chimiquement significatifs devraient être énumérées lors de l'échantillonnage intensif de l'espace des problèmes de l'étude *QSAR*. Elles devraient montrer de bonnes performances, sans être les meilleures en termes de scores d'entraînement / de validation (par conséquent, des procédures déterministes de mise en place de *QSAR* pourraient ne pas les trouver - d'où l'intérêt de l'utilisation du *SQS*).

En conclusion Les FPTs ont prouvé une nouvelle fois leur utilité. En effet, ils peuvent donner des modèles *QSAR* valables, pour peu que la diversité du set d'entraînement est assez grande pour permettre l'apprentissage des caractéristiques clés, et non pas de signatures pharmacophoriques secondaires qui reflètent des sous-sets localement enrichis en actifs. Que les triplets sélectionnés correspondent ou pas aux véritables pharmacophores de liaison est une question de diversité du jeu de données. Lorsque le jeu de données n'est pas assez représentatif, les FPTs ont été d'excellents moyens de mettre en valeur les déficiences des sets d'entraînement : les termes interprétables chimiquement et responsables des artefacts observés permettent une compréhension simple du problème. Si l'ensemble d'entrainement est suffisamment divers, le rayon d'applicabilité des modèles basés sur les FPTs peut s'étendre à plusieurs chémotypes – et pourrait même aller au delà des espérances si le site actif ciblé offre des modèles alternatifs pour accommoder un nouveau

triplet topologique.

Les excellents résultats obtenus dans les collaborations industrielles ont clairement démontré qu'à partir d'un jeu d'apprentissage de milliers de composés divers, les modèles $QSAR$ basés sur les 2D-FPTs ont en effet un vrai pouvoir prédictif validé expérimentalement.

Pour finir, nous avons utilisé les FPTs afin de construire des cartes auto-organisatrices (SOMs). Le but de cette étude est d'aider à accélerer les recherches par similarité dans des bases de données. En effet, le nombre de plus en plus considérable de composés regroupés dans les bases de données entraine des recherches qui peuvent se révéler très longues pour chaque requête. Il existe des manières "classiques" d'accélération des recherches dans des bases de données, mais elles sont basées sur des manipulations de bits. Or, Les triplets pharmacophoriques ne sont pas éligibles pour des méthodes de ce genre. Cependant, les cartes auto-organisatrices peuvent être utilisées sur ce type de descripteurs et permettent de regrouper l'espace des données d'entrée de telle manière à ce que les composés similaires se retrouvent dans des zones proches (dans le même neurone ou dans des neurones voisins).

De nombreuses cartes ont été construites, sur deux ensembles de données (*SmallRef* et *Extended*, comportant 11168 et 53206 composés respectivement). Une fois les cartes générées, un jeu de données de test de 57613 composés (dont les 2000 premiers composés sont les requêtes) a été projeté sur chacune des cartes, afin de comparer leur propension a accélérer les recherches par similarité. Nous avons basé notre idée d'accélération sur un principe simple : un composé requête ne sera comparé qu'à ses plus proches voisins, c'est à dire les composés de la base de donées qui se trouvent soient dans le même neurone que lui, soit dans les neurones voisins. Le critère de qualité Q a représenté notre manière principale de faire un tri entre les cartes. Ce critère permet d'établir un compromis entre le temps utilisé pour rechercher les 2000 composés requêtes parmi les 55613 composés restants et le nombre de *Hits* virtuel retrouvés. Nous avons utilisé deux types de scores (Euclidien et Tanimoto) pour calculer la similarité entre nos requêtes et la base données. Le critère de qualité prend en compte ces deux types de valeurs afin de donner un aperçu global de la qualité de la recherche effectuée.

Il apparait que cette méthode d'accélération peut s'avérer très efficace, si tant est que les cartes ont été correctement choisies. En effet, certaines cartes permettent d'accélérer considérablement les recherches dans la base de données tout en limitant les pertes de Hits virtuels. Ceci est représenté par de hauts scores de qualité Q. En effet, les meilleures cartes que nous avons pu générer sont capables de retrouver 90% des *Hits* virtuels en seulement 10% du temps qui aurait été nécessaire pour calculer les similarités sur l'ensemble de la base de données.

Nous avons pu voir dans cette étude qu'une attention toute particulière doit être portée à la phase d'entrainement des cartes. En effet, les résultats d'accélération sont considérablement dépendants de la qualité des cartes, et donc de la manière dont elles ont été entrainées. Nous avons ainsi comparé plusieurs types d'entrainements pour chacun des paramètres des cartes : taille, fonction de voisinage, taille du jeu de données d'entrainement, nombre d'itérations en phase d'entrainement. De manière assez surprenante pour le type de méthode utilisée (les cartes auto-organisatrices sont basées sur un algorithme d'apprentissage non supervisé), nous avons détecté des artefacts d'*overfitting* lorsque les cartes sont sur-entrainées, et ce, sur les deux types de jeux de données d'entrainement.

Les cartes présentant ce type d'artefacts donnent lieu à des performances plus basses, probablement dues à un réarrangement des familles de molécules dans les neurones voisins de leur neurone "idéal". Bien que nous ayons pu déterminer un certain nombre de "bonnes" cartes (présentant de manière claire un bon entrainement et pas d'artefacts) sur les deux ensembles d'entrainement, il reste cependant impossible de trouver des règles sans équivoque pour la création de cartes. En effet, la qualité reste dépendante non seulement de l'entrainement mais aussi de la taille des cartes ainsi que du jeu de données utilisé. La stratégie la plus efficace reste de vérifier graduellement la qualité de chaque carte au cours de son entrainement, et de choisir le nombre d'itérations optimal correspondant à chaque ensemble de données pour chaque paramétrisation de cartes.

Le jeu de données utilisé pour créer les cartes a aussi son importance. En effet, il est important de choisir cet ensemble de manière à ce qu'il soit représentatif de l'espace chimique visé. Dans notre cas, nous avons souhaité orienter nos recherches par similarité dans un ensemble de molécules *druglike*. La comparaison entre les cartes générées sur l'ensemble *SmallRef* et l'ensemble *Extended* nous a permis de conclure que le jeu de données le plus petit (comprenant tout de même 11168 molécules) semble parfaitement suffisant pour décrire l'espace chimique de notre base de données.

L'ensemble *Extended*, avec ses 53206 molécules n'a pas semblé apporter d'améliorations à la performance des cartes et a même dans certains cas fait baisser leur qualité. Ainsi, de manière surprenante, l'ajout de nouvelles informations apportées par cet ensemble de données n'a pas été bénéfique pour la qualité des cartes. Ceci peut être dû au fait que les cartes ont plus de difficulté à converger avec un grand jeu de données d'entrainement. Il reste important de noter que notre "petit" ensemble d'entrainement était déjà d'une taille assez conséquente et aussi suffisamment divers et représentatif de notre plus grande base de données.

L'ensemble de nos test basés sur le critère de qualité Q ont permis d'identifier la meilleure carte ($Q = 0.77$). Cette carte, entrainée sur le petit jeu de données *SmallRef*, possède une topologie rectangulaire et une fonction de voisinage Bubble. Elle contient $18 \times 20 = 360$ neurones et a été entrainée en 3 étapes : *Brute* (1000 itérations) puis *Refinement* (10000 itérations) et enfin *HyperRefinement* (50000 itérations). Pour une meilleure visualisation de la répartition des composés sur la carte, nous avons attribué à chaque composé de l'ensemble *SmallRef* une valeur comprise entre 0 et 5, et indicant le nombre de fois que le composé viole les règles de Lipinski. Nous avons pu observer une coupure nette sur la diagonale de la carte entre les composés *druglike* (0 ou 1 violation) et les composés *non-druglike* (tous les autres). Cette séparation indique que notre meilleure carte est capable de représenter avec succès l'espace chimique de nos composés d'entrainement d'un point de vue "médicinal".

Pour conclure, nous avons effectué des tests grandeur nature afin de vérifier si notre critère de qualité Q a un sens en dehors des tests sur notre propre base de données. Pour ce faire, nous avons projeté deux ensembles de données (*ExtQ* - ensemble de 12491 molécules requêtes et *ExtDB* - ensemble d'environ 160000 composés provenant de notre partenaire industriel) sur 4 cartes sélectionnées pour leur diversité de taille et de jeu de données d'entrainement. Les 4 cartes ont été utilisées dans différentes situations (avec un rayon de recherche différent) afin de comparer leur propension à accélerer les recherches, mais aussi afin de voir si le fait d'agrandir la zone de recherche autour des requêtes entraine une augmentation significative du nombre de *Hits* virtuels trouvés. Nous n'avions pas d'idée du nombre total de *Hits* virtuels qu'il est possible de retrouver (la taille totale des ensembles à comparer étant trop grande), mais nous pouvions cependant comparer l'efficacité des cartes entre elles. Il s'avère que l'utilisation des cartes auto-organisatrices a permis de renvoyer des résultats de criblage virtuel (basé sur la comparaison d'empreintes de dimension 4418) en un peu plus d'une heure, sans rencontrer de pertes significative de *Hits* virtuels. Ce résultat indique non seulement que notre méthode d'accélération est efficace, mais aussi que le le bon comportement des cartes dans le processus de test sur une plus petite base de données est un bon indicateur de son comportement en conditions réelles.

Références

[1] Moe (molecular operating environment). Chemical Computing Group, Inc., 2005.

[2] J. Bajorath. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, 41(2) :233–245, 2001.

[3] A. Balaban. Higly discriminating distance based topological index. *Chem. Phys. Lett.*, 89 :399–404, 1982.

[4] J. M. Barnard and G. M. Downs. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.*, 37 :141–142, 1997.

[5] M. Barreca, S. Ferro, A. Rao, L. De Luca, M. Zappala, A. Monforte, Z. Debyser, M. Witvrouw, and A. Chimirri. Pharmacophore-based design of hiv-1 integrase strand-transfer inhibitors. *J. Med. Chem.*, 48 :7084–7088., 2005.

[6] J. Batista, J. Godden, and J. Bajorath. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, 46 :1937–1944, 2006.

[7] H. U. Bauer and K. R. Pawelzik. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Trans. Neur. Netw.*, 3 :570–579, 1992.

[8] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9) :509–517, 1975.

[9] T. Blundell. Structure-based drug design. *Nature*, 384 :23–36, 1996.

[10] E. Bolton, Y. Wang, P. Thiessen, and S. Bryant. Pubchem : Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, 4, 2008.

[11] F. Brown. Chemoinformatics : what it is and how does it impact drug discovery ? *Annu. Rep. Med. Chem.*, 33 :375–384, 1998.

[12] F. R. Burden. Molecular identifiation number for substructure searches. *J. Chem. Inf. Comput. Sci.*, 29 :225–227, 1989.

[13] A. Cammarata. An apparent correlation between the in vitro activity of chloramphenicol analogs and electronic polarizability. *J. Med. Chem.*, 10(4) :525–552, 1967.

[14] E. Carhart, D. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies : Definition and applications. *J. Chem. Inf. Comput. Sci.*, 25 :64–73, 1985.

[15] ChemAxon. pka calculator plugin. http ://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html, 2007.

[16] ChemAxon. Screen user guide. http ://www.chemaxon.com/jchem/ index.html ?content=doc/user/Screen.html, 2007 2007.

[17] A. Cheng, D. Diller, S. Dixon, W. Egan, G. Lauri, and K. Merz. Computation of the physio-chemical properties and data mining of large molecular collections. *J. Comput. Chem.*, 23(1) :172–183, 2002.

[18] A. Chiriac, D. Ciubotariu, S. Funar-Timofei, L. Kurunczi, M. Mracec, M. Mracec, Z. Szabadai, E. Seclaman, and Z. Simon. Qsar and 3d-qsar in timisoara, 1972-2005. *Rev. Roum. Chim.*, 51 :71–99, 2006.

[19] C. Cortes and V. Vapnik. Support-vector networks . *Mach. Learn.*, 20 :273–297, 1995.

[20] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, pages 21–27, 1967.

[21] R. Cramer, D. Patterson, and J. Bunce. Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110 :5959–5967, 1988.

[22] R. Cramer, D. Patterson, and J. Bunce. Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110 :5959–5967, 1988.

[23] P. Crivori, G. Cruciani, P. Carrupt, and B. Testa. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.*, 47 :2204–2216, 2000.

[24] G. Cruciani, P. Crivori, P. Carrupt, and B. Testa. Molecular fields in quantitative structure-permeation relationships : the volsurf approach. *Teochem*, 503 :17–30, 2000.

[25] A. Dudek, T. Arodzb, and J. Gálvezc. Computational methods in developing quantitative structure-activity relationships (qsar) : A review. *Combin. Chem. High Throughput Screen.*, 9 :213–228, 2006.

[26] J. Durant, B. Leland, D. Henry, and J. Nourse. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(6) :1273–1280, 2002.

[27] H. Eckert and J. Bajorath. Molecular similarity analysis in virtual screening : foundations, limitations and novel approaches. *Drug Discov. Today*, 12(5-6) :225–233, 2007.

[28] T. Ewing, C. Baber, and M. Feher. Novel 2d fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.*, 46 :2423–2431, 2006.

[29] U. Fechner, L. Franke, S. Renner, P. Schneider, and G. Schneider. Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput.-Aided Mol. Des.*, 17 :687–698, 2003.

[30] U. Fechner, J. Paetz, and G. Schneider. Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR Comb. Sci.*, 24 :961–967, 2005.

[31] R. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7 :179–188, 1936.

[32] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, and G. Schneider. Extraction and visualization of potential pharmacophore points using support vector machines : application to ligand-based virtual screening for cox-2 inhibitors. *J. Med. Chem.*, 48 :6997–7004, 2005.

[33] J. Galvez, R. Garcia-Domenech, M. Salabert, and R. Soler. Charge indexes. new topological descriptors. *J. Chem. Inf. Comput. Sci.*, 34 :520–525, 1994.

[34] S. Gelfand, C. Ravishankar, and E. Delp. An iterative growing and pruning algorithm for classification tree design. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 163–174, 1991.

[35] V. Gillet, P. Willett, and J. Bradshaw. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.*, 43 :338–345, 2003.

[36] A. Gionis, P. Indyk, and M. R. Similarity search in high dimensions via hashing. *VLDB '99 Proceedings of the 25th International Conference on Very Large Data Bases*, 1999.

[37] D. Graham, C. Malarkey, and M. Schulmerich. Information content in organic molecules : Quantification and statistical structure via brownian processing. *J. Chem. Inf. Model.*, 46 :1601–1611, 2004.

[38] R. Guha and P. Jurs. Development of qsar models to predict and interpret the biological activity of artemisinin analogues. *J. Chem. Inf. Comput. Sci.*, 44 :1440–1449, 2004.

[39] I. Guyon and A. Elisseeff. Special issue on variable and feature selection. *J. Mach. Learn Res.*, 3 :1157–1182, 2003.

[40] M. Hann and R. Green. Chemoinformatics — a new name for an old problem ? *Curr. Opin. Chem. Biol.*, 3(4) :379–383, 1999.

[41] G. Hessler, M. Zimmermann, H. Matter, A. Evers, T. Naumann, T. Lengauer, and M. Rarey. Multiple-ligand-based virtual screening : Methods and applications of the mtree approach. *J. Med. Chem.*, 48 :6575–6584, 2005.

[42] J. Higo and N. Go. Algorithm for rapid calculation of excluded volume of large molecules. *J. Comp. Chem.*, 10 :376–379, 1989.

[43] D. Horvath. Compharm – automated comparative analysis of pharmacophoric patterns and derived qsar approaches, novel tools in high throughput drug discovery. a proof of concept study applied to farnesyl protein transferase inhibitor design. In M. Diudea, editor, *QSPR/QSAR Studies by Molecular Descriptors*, pages 395–439. Nova Science Publishers, Inc, New York, 2001.

[44] D. Horvath. High throughput conformational sampling & fuzzy similarity metrics : A novel approach to similarity searching and focused combinatorial library design and its role in the drug discovery laboratory. In V. V. Ghose, A.K., editor, *Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discovery.*, pages 429–472. Marcel Dekker, Inc., New York, 2001.

[45] D. Horvath and F. Barbosa. Neighborhood behavior – the relation between chemical similarity and property similarity. *Curr. Trends Med. Chem.*, 4 :589–600, 2004.

[46] D. Horvath and C. Jeandenans. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces - a novel understanding of the molecular

similarity principle in the context of multiple receptor binding profiles. *J. Comput. Inf. Comp. Sci.*, 43 :680–690, 2003.

[47] D. Horvath and C. Jeandenans. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces – a benchmark for neighborhood behavior assessment of different in silico similarity metrics. *J. Chem. Inf. Comput. Sci.*, 43 :691–698, 2003.

[48] D. Horvath, C. Koch, G. Schneider, G. Marcou, and A. Varnek. Local neighborhood behavior in a combinatorial library context. *J. Comput. Aided Mol. Des.*, 25(237-252), 2011.

[49] T. Hou, W. Zhang, K. Xia, X. Qiao, and X. Xu. Adme evaluation in drug discovery. 5. correlation of caco-2 permeation with simple molecular properties. *J. Chem. Inf. Model.*, pages 1585–1600, 2004.

[50] A. Jain, J. Mao, and K. Mohiuddin. Artificial neural networks : a tutorial. *Computer*, 29(3) :31–44, 1996.

[51] G. Jones, P. Willet, and R. Glen. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.*, 9 :532–549, 1995.

[52] A. Katritzky, L. Mu, V. Lobanov, and M. Karelson. Correlation of boiling points with molecular structure. 1. a training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Comp. Chem.*, 100 :10400, 1996.

[53] L. Kier and L. Hall. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.*, 70 :583–589, 1981.

[54] D. Kitchen, H. Decornez, J. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery : methods and applications. *Nat. Rev. Drug Discov.*, 3 :935–949, 2004.

[55] G. Klopman. Chemical reactivity and the concpt of charge and frontier controlled reactions. *J. Am. Chem. Soc.*, 90 :223–234, 1968.

[56] I. Kuntz, J. Blaney, S. Oatley, R. Langridge, and T. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161 :269–288, 1982.

[57] L. Kurunczi, E. Seclaman, T. Oprea, L. Crisan, and Z. Simon. Mtd-pls : A pls variant of the minimal topologic difference method. iii. mapping interactions between estradiol derivatives and the alpha estrogenic receptor. *J. Chem. Inf. Model.*, 45 :1275–1281, 2005.

[58] P. Labute. A widely applicable set of descriptors. *J. Mol. Graph. Model.*, 18 :464–477, 2000.

[59] T. Laidboeur, D. CabrolBass, and O. Ivanciuc. Determination of topo-geometrical equivalence classes of atoms. *J. Chem. Inf. Comput. Sci.*, 37 :87–91, 1997.

[60] T. Lin, H. Li, and K. Tsai. Implementing the fisher's discriminant ratio in a k-means clustering algorithm for feature selection and data set trimming. *J. Chem. Inf. Comput. Sci.*, 44(1) :76–87, 2004.

[61] C. Low, I. Buck, T. Cooke, J. Cushnir, S. Kalindjian, A. Kotecha, M. Pether, N. Shankley, J. Vinter, and L. Wright. Scaffold hopping with molecular field points : identification of a cholecystokinin-2 (cck2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based cck2 antagonists. *J. Med. Chem.*, 48 :6790–6802, 2005.

[62] P. Lyne. Structure-based virtual screening : an overview. *Drug Discov. Today*, 7(20) :1047–1055, 2002.

[63] Y. Martin, M. Bures, E. Danaher, J. Delazzer, I. Lico, and P. Pavlik. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.*, 7 :83–102, 1993.

[64] J. Mason, I. Morize, P. Menard, D. Cheney, C. Hulme, and R. Labaudiniere. New 4-point pharmacophore method for molecular similarity and diversity applications : Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.*, 38 :144–150, 1998.

[65] C. Merkwirth, H. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl, and T. Lengauer. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.*, 44(6) :1971–1978, 2004.

[66] M. Minailiuc, O.M.and Diudea. Ti-mtd model. applications in molecular design. In M. Diudea, editor, *QSPR/QSAR Studies by Molecular Descriptors.*, pages 363–388. Nova Science Publishers, Inc., New York, 2001.

[67] R. S. Mulliken. Electronic population analysis on lcao-mo molecular wave functions. i. *J. Chem. Phys.*, 23 :1833–1846, 1955.

[68] M. Olah, C. Bologa, and T. Oprea. An automated pls search for biologically relevant qsar descriptors. *J. Comput.-Aided Mol. Des.*, 18 :437–439, 2004.

[69] S. Oloff, R. Mailman, and A. Tropsha. Application of validated qsar models of d(1) dopaminergic antagonists for database mining. *J. Med. Chem.*, 48 :7322–7332, 2005.

[70] S. Pickett, J. Mason, and I. McLay. Diversity profiling and design using 3d pharmacophores : Pharmacophore-derived queries. *J. Chem. Inf. Comput. Sci.*, 36(1214-1223), 1996.

[71] S. Pickett, J. Mason, and I. McLay. Diversity profiling and design using 3d pharmacophores : Pharmacophore-derived queries. *J. Chem. Inf. Comput. Sci.*, 36(1214-1223), 1996.

[72] Y. Qi, R. Sadreyev, Y. Wang, B.-H. Kim, and N. Grishin. A comprehensive system for evaluation of remote sequence similarity detection. *BMC BIOINFORMATICS*, 8 :Art. No. 314, 2007.

[73] J. Quinlan. Induction of decision trees. *Mach. Learn.*, 1 :81–106, 1986.

[74] M. Randić. Characterization of molecular branching. *J. Am. Chem. Soc.*, 97(23) :6609–6615, 1975.

[75] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6) :386–408, 1958.

[76] E. Russo. Chemistry plans a structural overhaul. *Nature*, 419 :4–7, 2002.

[77] G. Schneider, W. Neidhart, T. Giller, and G. Schmid. "scaffold-hopping" by topological pharmacophore search : A contribution to virtual screening. *Angew. Chem. Int. Ed. Engl.*, 38 :2894–2896, 1999.

[78] H. Schultz. Topological organic chemistry. 1. graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, 29(3) :227–228, 1989.

[79] C. Senese, J. Duca, D. Pan, A. Hopfinger, and Y. Tseng. 4d-fingerprints, universal qsar and qspr descriptors. *J. Chem. Inf. Model.*, pages 1526–1539, 2004.

[80] R. Sheridan and S. Kearsley. Why do we need so many chemical similarity search methods ? *DDT*, 7(17) :903–911, 2002.

[81] B. Silverman and D. Platt. Comparative molecular moment analysis (comma) : 3d-qsar without molecular superposition. *J. Med. Chem.*, 39(11) :2129–2140, 1996.

[82] Z. Simon, A. Chiriac, S. Holban, D. Ciubotariu, and G. Mihalas. *Minimum Steric Difference. The MTD Method for QSAR Studies.* Research Studies Press. Letchworth and Wiley, New York, 1984.

[83] A. Smellie. Compressed binary bit trees : A new data structure for accelerating database searching. *J. Chem. Inf. Model.*, 49 :257–262, 2009.

[84] F. Stahura and J. Bajorath. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.*, 11(9) :1189–1202, 2005.

[85] D. Stanton. Evaluation and use of bcut descriptors in qsar and qspr studies. *J. Chem. Inf. Comput. Sci.*, 39 :11–20, 1999.

[86] D. Stanton, L. Egolf, P. Jurs, and M. Hicks. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *J. Chem. Inf. Comput. Sci.*, 32(4) :306–316, 1992.

[87] T. Steindl, D. Schuster, C. Laggner, and T. Langer. Parallel screening : a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.*, 46 :2146–2157, 2006.

[88] N. Stiefl, I. Watson, K. Baumann, and A. Zaliani. Erg : 2d pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.*, 46 :208–220, 2006.

[89] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. Sheridan, and Q. Song. Boosting : An ensemble learning tool for compound classification and qsar modeling. *J. Chem. Inf. Model.*, pages 786–799, 2005.

[90] D. C. I. Systems. Smarts. http ://www.daylight.com/dayhtml/doc/theory.smarts.html, 2007.

[91] S. Trohalaki, R. Pachter, K. Geiss, and J. Frazier. Halogenated aliphatic toxicity qsars employing metabolite descriptors. *J. Chem. Inf. Model.*, pages 1186–1192, 2004.

[92] W. Walters, M. Stahl, and M. Murko. Drug discovery — an overview. *Drug. Discov. Today*, 3 :160–178, 1998.

[93] W. Warr. Balancing the needs of the recruiters and the aims of the educators. In *Presented at 218th ACS National Meeting. New Orleans, Louisiana*, 1999.

[94] J. Weiser, A. Weiser, P. Shenkin, and W. Still. Neighbor-list reduction : Optimization for computation of molecular van der waals and solvent-accessible surface areas. *J. Comp. Chem.*, 19 :797–808, 1998.

[95] H. Wiener. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, 69(1) :17–20, 1947.

[96] P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Model.*, 38 :983–996, 1998.

[97] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression : a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58 :109–130, 2001.

[98] S. Word, C. Albano, D. I. W.J., K. Esbensen, S. Hellberg, E. Johansson, and H. Sjöström. Pattern recognition : finding and using regularities in multivariance data. *SIAM J. Sci. Stat. Comput.*, 5 :735–743, 1984.

[99] S. Yalkowsky, A. Sinkula, and S. Valvani. *Physical chemical properties of drugs.* Marcel Dekker, 1988.

[100] Z. Zhou and R. Parr. Activation hardness : New index for describing the orientation of electrophilic aromatic substitution. *J. Am. Chem. Soc.*, 112 :5720–5724, 1990.

[101] M. Zimmermann. *Rechnerunterstütze Analyse von HTS - Daten.* Mathematisch Natur-wissenschaftliche Fakultät ; Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 2003.