Développement de nouvelles méthodes de criblage in silico en chémogénomique

Jamel-Eddine Meslamani

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire d'Innovation Thérapeutique, UMR 7200

THÈSE

présentée par

Jamel Meslamani

soutenue le : **13 Septembre 2012**

pour obtenir le grade de

Docteur de l'Université de Strasbourg

Discipline / Spécialité : Chimie/Chémoinformatique

Développement de nouvelles méthodes de criblage in silico en chémogénomique

THÈSE dirigée par :

Dr. ROGNAN Didier	Directeur de recherche, CNRS

RAPPORTEURS :

Dr. MORELLI Xavier	Chargé de recherche, Centre de Recherche en Cancérologie de Marseille
Pr. MORIN-ALLORY Luc	Professeur, Université d'Orléans

MEMBRES DU JURY :

Dr. JACOBY Edgar	Senior Investigator, Novartis, Bâle
Pr. VARNEK Alexandre	Professeur, Université de Strasbourg

Remerciements

Je remercie tout d'abord le Pr. Marcel Hibert pour m'avoir accueilli au sein du Laboratoire d'Innovation Thérapeutique. Ensuite mon superviseur, le Dr. Didier Rognan, premièrement, parce qu'il a cru en moi, s'est débattu pour obtenir un financement pour la thèse et pour m'avoir encadré. C'est sans doute grâce à ses efforts et à son investissement personnel que j'ai eu la chance de faire cette thèse. Je ne le remercierai jamais assez pour ça. J'ai vraiment beaucoup appris au sein de son laboratoire autant sur le plan scientifique que personnel.

Je veux également remercier notre deuxième superviseur à nous tous au laboratoire, le Dr. Esther Kellenberger. Nous avons travaillé ensemble sur la base de données sc-PDB et c'était vraiment une expérience enrichissante. Je la remercie vivement car elle était tout le temps à l'écoute, me conseillait et m'orientait dans les différents projets.

Je tiens également à remercier les membres du jury, Dr. Xavier Morelli, Pr. Luc Morin-Allory, Pr. Alexandre Varnek et le Dr. Edgar Jacoby, pour avoir accepté de juger mon travail. Le Dr. Frédéric Bihel pour l'avoir fait au cours de la deuxième année. Je remercie autant toutes les personnes avec lesquelles j'ai pu collaborer tout au long de la thèse: Jérôme Pansanel de l'Institut Pluridisciplinaire Hubert Curien, Pascal Calvat du centre de calcul CC-IN2P3, Jiabo Li, Jon Sutter, Adrian Stevens, Hugues-Olivier Bertrand et Katalin Nadassy de la campanie Accelrys.

Je remercie ensuite tout spécialement mon ancien collègue, le Dr. Nathanaël Weill, pour ses conseils et pour son aide dès mon arrivée au laboratoire et qui continue de le faire d'ailleurs ! Ainsi que tous mes collègues à commencer par ceux qui sont déjà partis: Jérôme Hert, Claire Schalon, Chris de Graaf, Gwo-Yu Chuang, Vladimir Chupakhin et ceux qui sont encore présents et avec lesquels j'ai passé d'agréables moments : Anne-Marie Ray, Jérémy Désaphy et Noé Sturm ainsi que tous les membres de l'UMR pour l'ambiance fort chaleureuse.

J'aimerais également remercier tous les membres du laboratoire d'Infochimie à Strasbourg pour les moments passés ensemble lors de l'organisation des écoles d'été et les

différentes manifestations scientifiques. Une petite pensée pour Evgeny Kondratovich et sa famille, un collègue et ami apprécié de tous, qui est malheureusement parti très tôt.

Et enfin, je remercie tout particulièrement ma mère ainsi que toute ma famille qui m'ont toujours soutenu et encouragé tout le long.

Sommaire

Résumé

La chémoinformatique et la bioinformatique sont des disciplines devenues indispensables à la découverte de médicaments. De nos jours, les industries pharmaceutiques consacrent près de 10% de leur budget de recherche et développement, à la recherche de médicaments assisté par ordinateur (Kapetanovic 2008). Cette émergence peut s'expliquer à la fois par le développement des architectures de calculs mais aussi par le faible coup qu'engendrent des analyses in silico par rapport à des tests in-vitro.

Les essais biologiques qui ont été menés depuis des décennies afin d'identifier des médicaments potentiels, commencent à former une source très importante de données et plusieurs bases de données commencent à les répertorier. La disponibilité de ce type de données a favorisé le développement d'un nouvel axe de recherche appelé la "chémogénomique" et qui s'intéresse à l'étude et à l'identification des associations possibles entre plusieurs molécules et plusieurs cibles. Ainsi, la chémogénomique permet de déterminer le profil biologique d'une molécule et nous renseigne sur sa capacité à devenir une touche intéressante mais aussi à identifier ses possibles effets indésirables.

Des méthodes de chémoinformatique permettent d'utiliser ces sources de données à des fins d'apprentissage et établir des modèles prédictifs qui permettront par la suite de faire des prédictions pour connaitre l'activité d'une molécule.

Cette thèse a porté sur le développement et l'utilisation de méthodes de prédictions d'association protéine-ligand. La prédiction d'une association est importante en vue d'un criblage virtuel et peut s'effectuer à l'aide de plusieurs méthodes. Au sein du laboratoire, on s'intéresse plus particulièrement au profilage de bases de données de molécules (chimiothèques) contre une série de cibles afin d'établir leur profil biologique. J'ai donc essayé au cours de ma thèse de mettre au point des modèles prédictifs d'association protéine-ligand pour un grand nombre de cibles, valider des méthodes de criblage virtuel récentes à des fins de profilage mais aussi établir un protocole de profilage automatisé, qui décide du choix de la méthode de criblage la plus adaptée en s'appuyant sur les propriétés physico-chimiques du ligand à profiler et de l'éventuelle cible.

Au commencement de la thèse, nous nous sommes intéressés à l'identification des différents sites de liaison au sein d'une même protéine et ceci à travers celles présentes dans la base de données des sites sc-PDB. L'identification de ses sites nous permettra d'éliminer les redondances de touches lors de campagnes de criblages virtuels et de profilage biologique.

Dans un deuxième temps, des modèles d'apprentissage ont été générés afin de prédire l'association entre un ligand et une cible. Nous avons ainsi généré des modèles prédictifs pour toutes les cibles protéiques ayant des structures cristallographiques, et évalué leurs performances prédictives à travers l'utilisation de différents descripteurs de protéines. Nous avons prêté une attention particulière à l'apport d'un descripteur de cavité dans les performances de ces modèles chémogénomique prédictifs.

Dans un troisième temps, une méthode de criblage virtuel, basée sur les pharmacophores générés à partir des interactions identifiées entre une cible et son ligand, a été évaluée dans le cadre d'un profilage biologique. La méthode a été comparée aux méthodes basées sur les ligands à l'aide des similarités 2D et 3D, ainsi qu'à l'arrimage moléculaire afin de se positionner par rapport à leurs performances.

Dans la dernière partie de cette thèse, un protocole de profilage automatique hybride a été réalisé et validé. Le but de cette étude a été d'utiliser la méthode de criblage virtuel adéquate en fonction de la nature de la molécule à profiler ainsi que des propriétés de la cible.

Références :

Kapetanovic, I. M., Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chem Biol Interact* **2008,** 171, (2), 165-76

Summary

Chemoinformatics and bioinformatics methods are now necessary in every drug discovery program. Pharmaceutical industries dedicate more than 10% of their research and development investment in computer aided drug design (Kapetanovic 2008). The emergence of these tools can be explained by the increasing availability of high performance calculating machines and also by the low cost of in silico analysis compared to in vitro tests.

Biological tests that were performed over last decades are now a valuable source of information and a lot of databases are trying to list them. This huge amount of information led to the birth of a new research field called “chemogenomics”. The latter is focusing on the identification of all possible associations between all possible molecules and all possible targets. Thus, using chemogenomics approaches, one can obtain a biological profile of a molecule and even anticipate possible side effects.

This thesis was focused on the development of approaches that aim to predict the binding of molecules to targets. In our lab, we focus on profiling molecular databases in order to get their full biological profile. Thus, my main work was related to this context and I tried to develop predictive models to assess the binding of ligands to proteins, to validate some virtual screening methods for profiling purpose, and finally, I developed an automatic hybrid profiling workflow that selects the best fitted virtual screening approach to use according the ligand/target context.

In the first study, we were interested in identifying different binding sites within one target. Thus, we used our in house database of annotated binding sites sc-PDB. Identifying distinct binding sites for each target will allow us to perform a better analysis of virtual screening hits by discarding the redundancies of sc-PDB entries.

Second, chemogenomic SVM models were developed in order to predict ligand binding to several targets. The models were dedicated to crystallographic structures and we evaluated the model performance by using different target descriptors.

Third, we evaluated the performance of a new protein-ligand pharmacophore virtual screening method for the biological profiling. The latter was compared to 2D and 3D ligand-based methods and molecular docking as a structure-based method.

And last, we developed and validated a hybrid and efficient profiling protocol which will select the appropriate virtual screening method according to the profiled ligand and targets properties.

References:

Kapetanovic, I. M., Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chem Biol Interact* **2008,** 171, (2), 165-76

Acronymes

ADME : Absorption, distribution, métabolisme et excrétion

DUD : Directory of Useful Decoys

ECFP : Extended Connectivity Fingerprint

FCFP : Functional Connectivity Fingerprint

FPD : Feature-Pair Distribution

PDB : Protein Data Bank

PLS : Partial Least Square

RCPG : Récepteurs couplés aux protéines G

RMSD : Root Mean Square Deviation

sc-PDB : screening Protein Data Bank

SHED : Shannon Entropy Descriptors

SVM : Support Vector Machine

PHRAG : Pharmacophoric Fragments

Glossaire

Cible : Macromolécule biologique de type protéine (enzyme ou récepteur)

Ligand : Toute molécule de faible masse moléculaire capable de se lier à une cible.

Cible orpheline : Cible pour laquelle aucun ligand n'est connu.

Classe d'activité/effet pharmacologique : Quand un ligand se lie à une cible, il engendre une réponse qui peut être différente selon le type de ligand. Le ligand peut avoir soit un effet activateur ou inhibiteur. On distingue alors plusieurs classes d'activités différentes pour la même cible.

Site de liaison : Ensemble des acides aminés formant une cavité capable d'interagir avec un ligand.

Association : Interaction identifiée entre un ligand et une cible (peu importe le mode de liaison).

Pose d'arrimage : La conformation et la position prédite du ligand arrimé dans un site de liaison d'une cible donnée.

Modèle de classification : Tout modèle mathématique construit à l'aide de descripteurs et capable de classer une molécule en deux catégories (active, inactive) voir plusieurs.

Introduction

Mon travail de thèse a porté sur les méthodes de criblage en chémogénomique et plus particulièrement sur le profilage biologique. Disposer d'un profil biologique d'une molécule facilite le processus de découverte de médicaments car il permet d'anticiper des effets secondaires et éventuellement toxiques.
Les données de relations structure-activité sont une source primaire pour les études de chémogénomique. Celles-ci sont répertoriées dans des bases de données de bioactivité.

Nous allons exposer dans le premier chapitre les bases de données publiques de bioactivité et nous allons détailler celles qui répertorient les valeurs d'affinités entre des molécules et des cibles. Une attention particulière sera portée sur le contenu de ses bases publiques, leurs recouvrement avec les bases commerciales et enfin la fiabilité des données qu'elles contiennent.

Dans la deuxième partie du premier chapitre, nous allons exposer les différentes méthodes de profilage à l'aide de quelques exemples d'études. Nous allons aborder les approches basées sur les ligands, celles qui se basent sur les structures des cibles, celles qui portent sur une représentation hybride incluant à la fois le ligand et la cible, et enfin les approches qui exploitent les données expérimentales.

Nous aborderons ensuite dans le deuxième chapitre, une méthode d'identification des différents sites de liaisons pour chaque protéine dans la base de données cristallographique sc-PDB (Kellenberger *et al.* 2006). Cette base de données a été créée dans le but de fournir un ensemble de données adaptées au criblage virtuel. La multiplicité des entrées pour certaines cibles, nous a convaincu, lors d'analyses de criblages virtuels, qu'il serait intéressant d'éliminer cette redondance dans le but d'obtenir des listes où les cibles présentent une seule copie d'un de leurs sites de liaison.

Dans la troisième partie de la thèse, nous présenterons la génération de modèles de classification chémogénomique servant ainsi à du profilage biologique sur 87 cibles issues de la base sc-PDB. Nous avons utilisé différentes représentations de la cible, à savoir une représentation 1D en utilisant son nom Uniprot (UniProt 2012), une représentation 2D en utilisant le descripteur de sa séquence protéique (Leslie *et al.* 2002) et une représentation 3D en

utilisant un descripteur de sa cavité (Weill *et al.* 2010). Nous avons essayé de déterminer l'apport de l'utilisation d'un descripteur de cavité 3D dans les modèles chémogénomiques.

Dans la quatrième partie, nous traiterons de l'utilisation d'une nouvelle méthode de criblage virtuel utilisant les pharmacophores d'interactions protéine-ligand dans le cadre d'un profilage. La base de données Pharmadb incluant plus de 68000 pharmacophores a été créée pour la validation de ce profilage. 157 ligands divers ont été profilés sur cette base et les performances ont été comparées aux approches 2D et 3D basées sur les ligands ainsi qu'à l'arrimage moléculaire.

Et pour conclure, en se basant sur les résultats de profilages de l'étude précédente, nous avons établi un protocole automatique hybride dédié à cet effet. Celui-ci a pour but la sélection de la méthode de criblage virtuel la plus adaptée en fonction du ligand à profiler et de l'éventuelle cible. Ce protocole sera traité et la validation exposée dans le dernier chapitre.

Références

Kellenberger, E., P. Muller, C. Schalon, G. Bret, N. Foata and D. Rognan (2006). "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank." J Chem Inf Model **46**(2): 717-727.

Leslie, C., E. Eskin and W. S. Noble (2002). "The spectrum kernel: a string kernel for SVM protein classification." Pac Symp Biocomput: 564-575.

UniProt, C. (2012). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Res **40**(Database issue): D71-75.

Weill, N. and D. Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites." J Chem Inf Model **50**(1): 123-135.

Chapitre 1 :

Bases de données chémogénomiques et méthodes de profilage biologique.

1. Les bases de données publiques de bioactivité

Les données de relation structure-activité générées par les campagnes de criblages à haut débit et les programmes de chimie médicinale constituent une richesse qu'il est nécessaire d'exploiter. Cependant, l'archivage de ce type de résultats nécessite une harmonisation des données dûes à l'hétérogénéité induite par les conditions expérimentales et la technique de criblage ou de test utilisée.

Convaincu par l'importance de cette tâche, des initiatives ont été menées par des partenaires de recherche publique et/ou privées qui ont initié des programmes d'archivage. Les deux exemples d'instituts publics les plus connus qui investissent dans ce domaine sont le *National Center for Biotechnology Information* (NCBI) aux Etats Unis, et l'*Institut Européen de Bioinformatique* (EMBL-EBI) en Europe. Ces instituts financent des laboratoires et des infrastructures dédiés aux archivages de données de bioactivité.

L'accessibilité de la plupart de ces bases publiques n'est devenu effective qu'en 2007 lors de la mise à disposition de la base PubChem BioAssay (Wang *et al.* 2012) et la base ChEMBL (Gaulton *et al.* 2012) en 2009.

Les méthodes de profilage basées sur les ligands (Keiser *et al.* 2009), utilisent les ligands connus des cibles afin d'en identifier d'autres. Ces méthodes se basent sur les informations contenues dans ces bases de données. Elles ont contribué de manière significative à l'essor des études de profilage.

On ne détaillera que les sources primaires de données, car plusieurs autre bases de bioactivité utilisent les informations (Tableau 1) prélevées à partir d'autres bases souvent ChEMBL et PubChem BioAssay. Les bases de bioactivité qui sont considérées dans cette partie sont celles pour lesquelles des valeurs de constantes de dissociation (Kd) ou d'inhibition (Ki, IC50) sont disponibles.

Bien entendu, certaines bases de bioactivité (Tableau 1) ne répertorient que des résultats de criblage à haut débit alors que d'autres se focalisent sur une présentation chémogénomique regroupant ainsi plusieurs informations : informations chimiques (structure de molécules,

propriétés physicochimiques), biologiques (génomique, voie métabolique) et thérapeutiques (toxicologie, effets secondaires).

Base	Systèmes concernés	Information d'interaction	Sources	Informations disponibles	Lien
ChemBank (Seiler *et al.* 2008)	Tous	score-*Z* composite	Criblage à haut débit HTS	Interactions protéine-ligand	http://chembank.broadinstitute.org
Stitch (Kuhn *et al.* 2012)	Tous	score de confiance	ChEMBL (Gaulton *et al.* 2012), PDSP Ki (Roth *et al.* 2000), BindingDB (Liu *et al.* 2007), PDB (Berman *et al.* 2003), DrugBank (Knox *et al.* 2011), GLIDA (Okuno *et al.* 2008), CTD (Davis *et al.* 2011), KEGG (Tanabe *et al.* 2012)	Réseaux d'interaction protéine-ligand	http://stitch.embl.de
Comparative Toxicogenomics Database CTD (Davis *et al.* 2011)	Tous	score de confiance	Publications scientifiques	Interactions protéine-ligand, annotation de gènes, maladies, voies métaboliques	http://ctdbase.org
ChemProt (Taboureau *et al.* 2011)	Tous	Score de confiance ou valeur d'affinité	ChEMBL, WOMBAT, DrugBank, PubChem BioAssay, PDSP Ki, BindingDB, Stitch, PharmGKB, CTD	Interactions protéine-ligand, maladies et effet pharmacologique	http://www.cbs.dtu.dk/services/ChemProt
iPhace (Garcia-Serna *et al.* 2010)	Tous	Valeurs d'affinités	IUPHAR-DB, PDSP Ki	Interactions protéine-ligand, profil biologique	http://cgl.imim.es/iphace
TDR Targets (Magarinos *et al.* 2012)	Maladies tropicales	Valeurs d'affinités si disponibles	PubChem, DrugBank, criblage à haut débit	Interactions protéine-ligand, maladies et effet pharmacologique	http://tdrtargets.org
SuperTarget (Hecker *et al.* 2012)	Médicaments	Valeurs d'affinités si disponibles	BindingDB, DrugBank, SuperCyp, corrections manuelles	Interactions, voies métaboliques, ontologie, effets secondaires, classification ATC des médicaments (http://www.whocc.no/atc_ddd_index)	http://insilico.charite.de/supertarget

Promiscous (von Eichborn *et al.* 2011)	Médicaments + interactions protéine-protéine	Valeurs d'affinités si disponibles	BindingDB, DrugBank, SuperCyp, corrections manuelles	Interactions protéine-ligand et protéine-protéine, voies métaboliques, ontologie, effets secondaires, classification ATC des médicaments	http://bioinformatics.charite.de/promiscuous
PharmGKB (Hodge *et al.* 2007)	Médicaments	Indication sur la présence de l'interaction	Médicaments répertoriés par la FDA et publications scientifiques	Interactions protéine-ligand, interactions gène-médicament, maladies, voies métaboliques, effets secondaires	http://www.pharmgkb.org
DrugBank (Knox *et al.* 2011)	Médicaments	Indication sur la présence de l'interaction	Médicaments répertoriés par la FDA	Interactions protéine-ligand, interaction médicament-médicament, voies métaboliques, pharmacogénomique	http://www.drugbank.ca
eDrug3D (Pihan *et al.* 2012)	Médicaments	Indication sur la présence de l'interaction	Médicaments répertoriés par la FDA	Interactions protéine-ligand, pharmacologie, criblage virtuel possible	http://chemoinfo.ipmc.cnrs.fr/MOLDB/index.html
Sider (Kuhn *et al.* 2010)	Médicaments	Indication sur la présence de l'interaction	Médicaments répertoriés par la FDA	Interactions protéine-ligand, classification ATC des médicaments, effets secondaires	http://sideeffects.embl.de
Brenda (Scheer *et al.* 2011)	Enzymes	Valeurs d'affinités si disponibles	Publications scientifiques	Interactions protéine-ligand, maladies, voies métaboliques	http://www.brenda-enzymes.info

Tableau 1 : Bases de données de bioactivité dont certaines regroupent des informations de plusieurs bases et d'autres ne mentionnent que les interactions sans mesure d'affinité.

1.1. PubChem BioAssay

PubChem BioAssay (Wang *et al.* 2012) est sans doute l'une des bases de bioactivité les plus fournies. Elle répertorie des données d'activités issues de criblages à haut débit et de programmes de chimie médicinale. La base fournit également des outils d'analyses pour explorer et interpréter ces données. Cinquante-deux sources (Tableau 2) alimentent actuellement la base (Juin 2012) et ce nombre est en constante augmentation.

Source	Nombre d'essais
ChEMBL	615 740
Scripps Research Institute	1 101
NIH/NCGC	1 044
Sanford-Burnham Center	869
Broad Institute	689
NIH/NMMLSC	422
Southern Research Screening Center	292
Johns Hopkins Ion Channel Center	220
NIH/DTP-NCI	173
Vanderbilt Screening Center	138
ChemBank	106
Autres	675

Tableau 2 : Les plus importantes sources de la base PubChem BioAssay

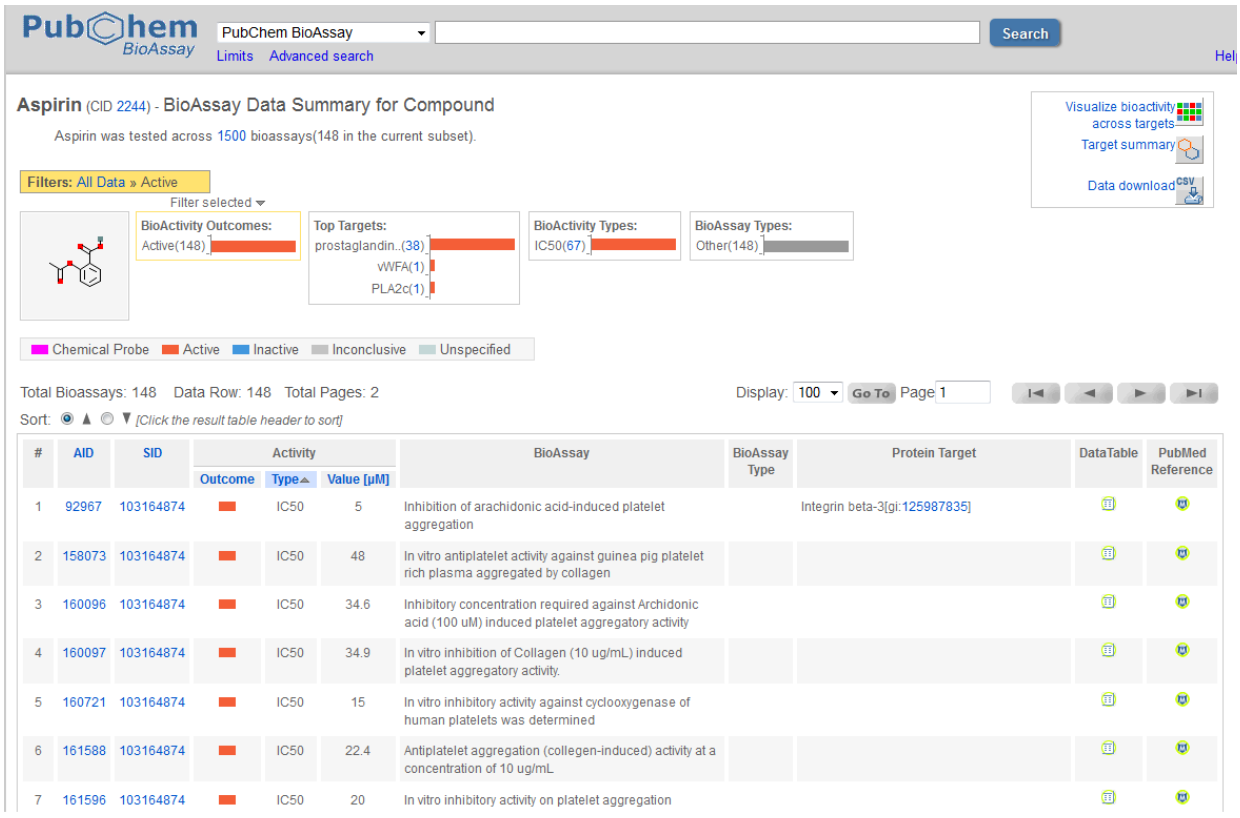
La base contient plus de 600 000 essais qui contiennent des activités entre une/plusieurs molécule(s) et une/plusieurs cible(s) ou cellule(s). Elle possède de plus des activités pour environ 5000 cibles différentes et plus de 130 millions de données de bioactivité. Environ 85% de ces données proviennent de criblages à haut débit et d'essais cellulaires et les 15% restant proviennent d'essais fonctionnels ou d'affinités sur des cibles (environ 800) pour lesquelles un identifiant PubMed (http://pubmed.gov) vers la publication qui expose les données est répertorié.

Une interface web permet de faire des requêtes personnalisées par molécules, protéine, gène ou sur les valeurs d'affinités. Pour chaque activité répertoriée, un lien vers la structure de la molécule dans la base "PubChem Substance" et/ou "PubChem Compound" est fourni. Si l'affinité est mesurée avec une cible ou un gène connu, un lien vers la base RefSeq (Benson

et al. 2012) est fourni à travers un numéro GI (*GI number*) afin d'obtenir une annotation plus complète de la cible ou du gène.

L'activité répertoriée dépend de la nature du test effectué. Dans le cas d'une mesure d'affinité, la concentration standard choisie est le micro-molaire (Figure 1). Pour les autres activités déterminées à partir d'autres méthodes, l'unité choisie est indiquée dans chaque fichier d'essai.

La base est téléchargeable sous trois différents formats. Le premier étant un fichier de type "ASN.1" qui est un format de fichier de norme ISO qui sert à produire des documents structurés. Le deuxième format proposé est le format "XML" qui est maintenant largement utilisé par les programmes et navigateurs du fait de sa simplicité. Le troisième format proposé est une combinaison de fichiers textes "CSV" contenant les données expérimentales et de fichiers "XML" contenant les en-têtes et définitions.



#	AID	SID	Activity			BioAssay	BioAssay Type	Protein Target	DataTable	PubMed Reference
			Outcome	Type	Value [µM]					
1	92967	103164874		IC50	5	Inhibition of arachidonic acid-induced platelet aggregation		Integrin beta-3[gi:125987835]		
2	158073	103164874		IC50	48	In vitro antiplatelet activity against guinea pig platelet rich plasma aggregated by collagen				
3	160096	103164874		IC50	34.6	Inhibitory concentration required against Archidonic acid (100 uM) induced platelet aggregatory activity				
4	160097	103164874		IC50	34.9	In vitro inhibition of Collagen (10 ug/mL) induced platelet aggregatory activity.				
5	160721	103164874		IC50	15	In vitro inhibitory activity against cyclooxygenase of human platelets was determined				
6	161588	103164874		IC50	22.4	Antiplatelet aggregation (collegen-induced) activity at a concentration of 10 ug/mL				
7	161596	103164874		IC50	20	In vitro inhibitory activity on platelet aggregation				

Figure 1 : Affichage des résultats d'activité pour la molécule d'aspirine à partir du site web PubChem BioAssay. Le tableau répertorie les différents essais (AID) et les valeurs d'activité correspondantes.

La base est également mise à jour quotidiennement et les activités sont rajoutées hebdomadairement.

Des efforts d'intégration d'autres bases de bioactivités publiques ont été entreprit par les équipes en charge de PubChem BioAssay. Des collaborations multilatérales ont permis l'intégration de grandes bases de bioactivité comme ChEMBL (Gaulton *et al.* 2012), IUPHAR-DB (Sharman *et al.* 2011) et BindingDB (Liu *et al.* 2007). Mais malheureusement, il est impossible de connaitre la version intégrée de ces bases. Il faut s'adresser directement au service support de PubChem afin d'obtenir l'information. Cependant, les politiques de stockage et de traitement de l'information (*data cleaning*) diffèrent d'une base à une autre. Par exemple, la base ChEMBL n'est intégrée qu'à moitié dans PubChem BioAssay. Seules les affinités ayant des scores de confiance (c.f. paragraphe 2.2) élevés sont retenues.

1.2. ChEMBL

Anciennement dénommée StARlite et initiée par la société Inpharmatica (Overington 2009), ChEMBL est sans doute la base de donnée de bioactivité la plus importante et la plus détaillée. Elle a été créée dans le but d'exploiter au mieux les relations de structures activités entre les molécules et les cibles. Elle a été acquise et maintenue par l'Institut Européen EMBL-EBI depuis 2008 et mise en ligne en Octobre 2009 dans sa première version.
ChEMBL inclut des essais contenant des affinités de molécules avec leurs cibles ou cellules, des essais fonctionnels et des propriétés ADMET. Les données sont extraites de publications issues des journaux scientifiques dédiés à la chimie médicinale (Gaulton *et al.* 2012). La base dans sa version actuelle ChEMBL_13 (Juin 2012) contient 6 933 068 données d'activités couvrant ainsi 8 845 cibles (Figure 2) et 1 143 682 molécules. Un identifiant de référence CiteXplore (outil bibliographique de référencement utilisé à l'EBI) est défini afin de faire un lien avec la publication scientifique qui contient les données. L'activité est répertoriée avec des concentrations nano-molaires (Figure 3).

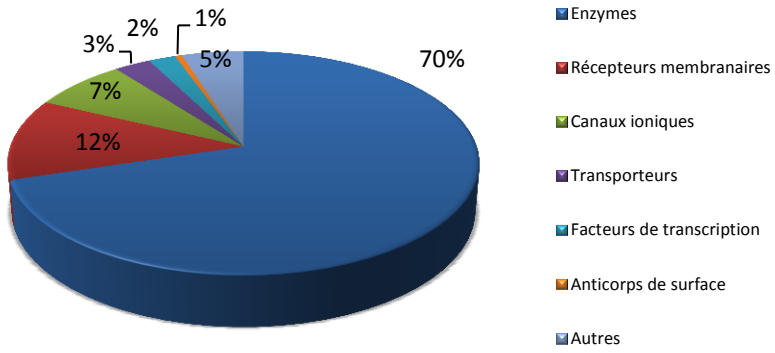
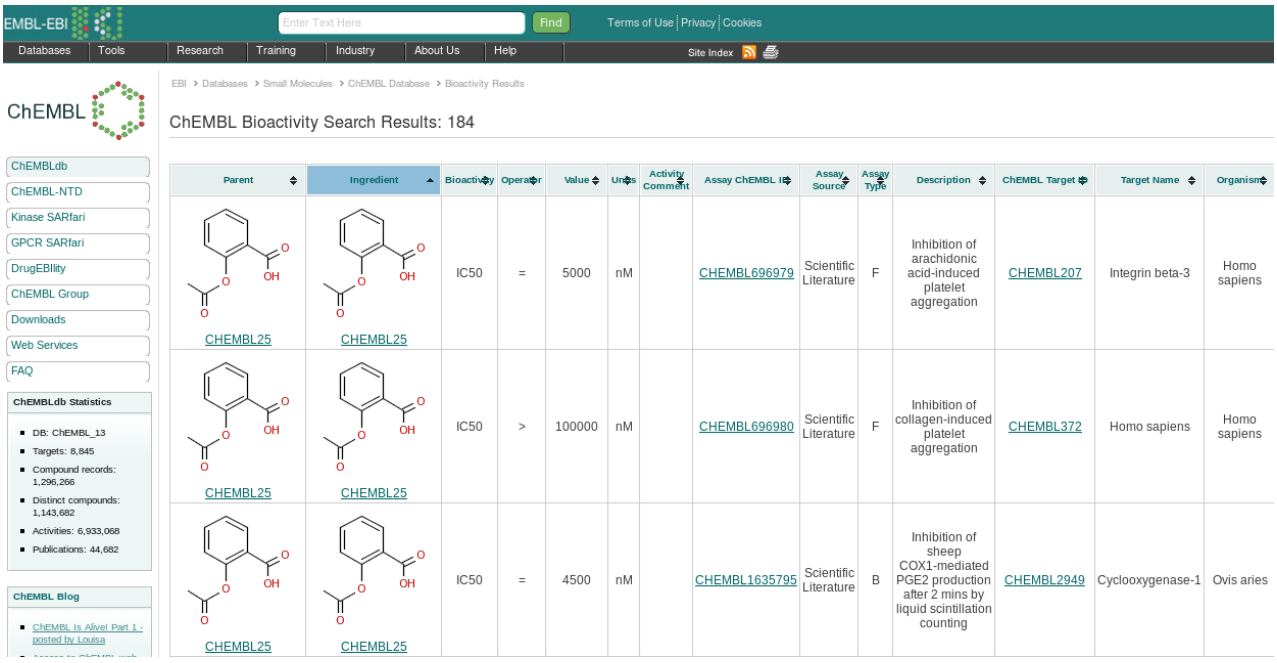


Figure 2 : Familles protéiques les plus importantes dans ChEMBL. Source : https://www.ebi.ac.uk/chembldb

ChEMBL contient également des informations sur les médicaments autorisés par l'agence de santé publique américaine (*Food and Drug Administration FDA*) à savoir la voie d'administration, informations sur le dosage, type de la molécule (synthétique, produit naturel, anticorps…).



Parent	Ingredient	Bioactivity	Operator	Value	Units	Activity Comment	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target	Target Name	Organism
CHEMBL25	CHEMBL25	IC50	=	5000	nM		CHEMBL696979	Scientific Literature	F	Inhibition of arachidonic acid-induced platelet aggregation	CHEMBL207	Integrin beta-3	Homo sapiens
CHEMBL25	CHEMBL25	IC50	>	100000	nM		CHEMBL696980	Scientific Literature	F	Inhibition of collagen-induced platelet aggregation	CHEMBL372	Homo sapiens	Homo sapiens
CHEMBL25	CHEMBL25	IC50	=	4500	nM		CHEMBL1635795	Scientific Literature	B	Inhibition of sheep COX1-mediated PGE2 production after 2 mins by liquid scintillation counting	CHEMBL2949	Cyclooxygenase-1	Ovis aries

Figure 3 : Affichage des résultats d'activité pour la molécule d'aspirine à partir du site web de ChEMBL. Le tableau répertorie les différents essais et les valeurs d'activités correspondantes.

La base est consultable à l'aide d'une interface web qui permet la recherche de molécules (par nom, code *smiles* ou structure), la recherche des cibles à l'aide de leurs noms et séquence, et l'exploration par familles protéiques ou à partir d'un arbre taxonomique.

Les cibles sont annotées à partir de la base Uniprot (UniProt 2012) et un lien vers celle-ci est répertorié pour permettre d'avoir une annotation plus complète. ChEMBL associe à chaque valeur d'activité un seuil de confiance. Ce seuil évalue la fiabilité de l'essai (conditions expérimentales) et la fiabilité sur l'annotation de la protéine. En effet, les cibles testées peuvent être insuffisamment décrites dans certaines publications scientifiques et des ambigüités peuvent subsister. Des protéines orthologues (transcrites par des gènes d'espèces différentes, descendants d'un même ancêtre commun et différenciés par un phénomène de spéciation) ou paralogues (transcrites par des gènes de même espèce, descendants d'un même ancêtre commun et différenciés par un phénomène de duplication) peuvent être mal distinguées lors de l'extraction de données des sources scientifiques.

Ce seuil de confiance varie de 0 à 9, et il est égal à 1 quand la cible est une cellule, et au moins égal à 4 si la cible est une protéine. Si celle-ci est bien définie, c'est-à-dire que la protéine est bien annotée, son espèce est définie et la paralogie/orthologie est distinguée, le seuil de confiance est alors supérieur à 7.

ChEMBL intègre les données d'affinité de PubChem BioAssay pour lesquelles des courbes de doses-réponses sont disponibles. Les équipes travaillent conjointement pour faciliter l'intégration des données.

La base est téléchargeable sous des formats de base de données relationnelle tels que *mysql* (http://www.mysql.com) et *Oracle database* (http://www.oracle.com).
ChEMBL est en moyenne mise à jour 3 à 4 fois par an et des améliorations ainsi que des corrections sont introduites dans chaque version.

1.3. IUPHAR-DB

Cette base de données a été mise à disposition par l'Union Internationale de Pharmacologie Fondamentale et Clinique (IUPHAR) qui est une association à but non lucratif dont l'objectif est de promouvoir la coopération internationale entre les pharmacologistes. La base est financée par des organismes internationaux comme l'UNESCO et plusieurs laboratoires pharmaceutiques privés. La base contient des informations d'affinité de molécules (de faible masse moléculaire et des peptides) sur des RCPG, des canaux ioniques, des récepteurs hormonaux nucléaires et quelques enzymes (Figure 4). Elle est plutôt focalisée sur les récepteurs car ils constituent les cibles de plus d'un tiers des médicaments (Sharman *et al.* 2011).

IUPHAR-DB est sans doute l'une des plus fiables car les données sont saisies par plus de 700 experts scientifiques coordonnés par l'IUPHAR. Les sources des données de la base sont des publications de revues médicales (Sharman *et al.* 2011). La base répertorie plus de 4000 molécules ayant des affinités pour plus de 500 cibles (Juin 2012). 99% des affinités présentent un lien PubMed vers la publication d'origine (Sharman *et al.* 2011).

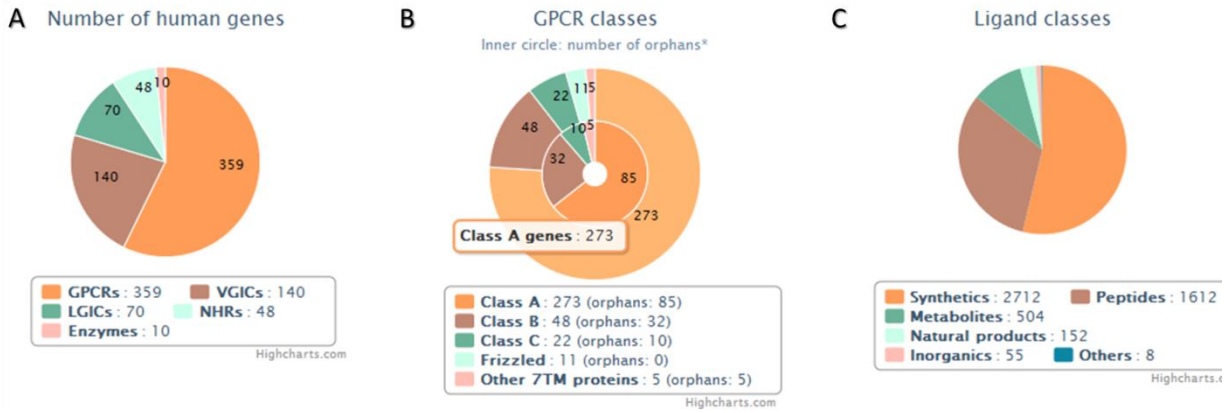


Figure 4 : A) Les grandes familles protéiques incluses dans IUPHAR-DB. **B)** Classification des RCPG ainsi que le contenu de chaque classe en nombre de protéines. **C)** Classification et nombre de molécules dans la base. Figure extraite du site de la base IUPHAR-DB (http://www.iuphar-db.org/index.jsp). Statistiques au mois de Juillet 2012.

Cette base est consultable via une interface web et la recherche peut s'effectuer à partir de la cible (par nom ou par identifiant d'autres banques génomiques mais pas de recherche par séquence) pour explorer tous ces ligands ou à partir des molécules (recherche par nom, structure et similarité) pour explorer leurs cibles. Plusieurs liens vers d'autres bases de données sont disponibles à la fois pour les ligands et pour les cibles afin de disposer d'une annotation plus complète (liste des bases disponible sur : http://www.iuphar-db.org/helpPage.jsp#databaseLinks). La base n'est téléchargeable que partiellement selon les résultats des requêtes saisies et il faut contacter les administrateurs de la base pour avoir une copie complète gratuite pour les chercheurs académiques (fichier *PostgreSQL*).

1.4. PDSP Ki

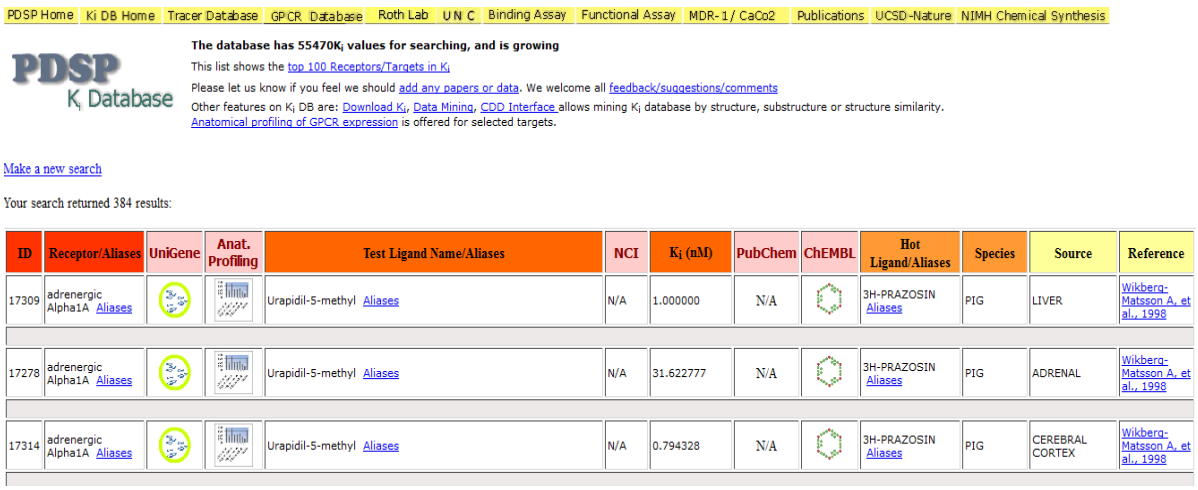
La base de données PDSP Ki est maintenue par l'Université de Caroline du Nord à Chapel Hill depuis 1999 et fait partie du programme de dépistage des molécules psychoactives (PDSP). La base de données tire ses fonds de ce programme initié par l'Institut National de la

Santé Mentale Américain (NIMH) et d'une donation de l'institut de recherche Heffter sur les psychédéliques.

PDSP Ki regroupe 55 000 valeurs d'activité pour un peu plus de 10 000 ligands et environ 750 cibles différentes dont un tiers sont d'origine humaine et 40% d'origine murine. Les valeurs d'activité sont exclusivement des valeurs de constantes d'inhibition exprimées à l'aide de concentrations exprimées en nano-molaire. Les cibles incluses dans la base appartiennent aux plus grandes familles protéiques dont certaines relèvent du système nerveux central.

La base est consultable en ligne via un portail web (http://pdsp.med.unc.edu) et les requêtes peuvent s'effectuer par nom de cible ou nom de la molécule. Cependant on ne peut effectuer que des recherches textuelles. Des recherches par similarité, sous-structure ou séquence sont donc à exclure. Pour chaque valeur d'affinité (Figure 5), une référence vers la base UniGene (Sayers *et al.* 2012) est indiquée ainsi qu'une référence vers les bases PubChem et ChEMBL est indiquée pour les ligands. Un lien PubMed est aussi présent pour accéder à la publication qui expose les résultats. On trouve aussi un lien vers un histogramme qui indique l'expression relative du gène codant pour la protéine dans les organes humains.

La base PDSP est téléchargeable en un seul fichier texte d'où il est possible d'extraire les affinités.


PDSP Home | Ki DB Home | Tracer Database | GPCR Database | Roth Lab | UNC | Binding Assay | Functional Assay | MDR-1/CaCo2 | Publications | UCSD-Nature | NIMH Chemical Synthesis

PDSP Ki Database

The database has 55470K_i values for searching, and is growing

This list shows the top 100 Receptors/Targets in K_i

Please let us know if you feel we should add any papers or data. We welcome all feedback/suggestions/comments

Other features on K_i DB are: Download K_i, Data Mining, CDD Interface allows mining K_i database by structure, substructure or structure similarity.

Anatomical profiling of GPCR expression is offered for selected targets.

Make a new search

Your search returned 384 results:

ID	Receptor/Aliases	UniGene	Anat. Profiling	Test Ligand Name/Aliases	NCI	K_i (nM)	PubChem	ChEMBL	Hot Ligand/Aliases	Species	Source	Reference
17309	adrenergic Alpha1A Aliases			Urapidil-5-methyl Aliases	N/A	1.000000	N/A		3H-PRAZOSIN Aliases	PIG	LIVER	Wikberg-Matsson A. et al., 1998
17278	adrenergic Alpha1A Aliases			Urapidil-5-methyl Aliases	N/A	31.622777	N/A		3H-PRAZOSIN Aliases	PIG	ADRENAL	Wikberg-Matsson A. et al., 1998
17314	adrenergic Alpha1A Aliases			Urapidil-5-methyl Aliases	N/A	0.794328	N/A		3H-PRAZOSIN Aliases	PIG	CEREBRAL CORTEX	Wikberg-Matsson A. et al., 1998

Figure 5: Résultat d'affinités pour le médicament Prazosin à partir du site web de la base.

1.5. Binding DB

Binding DB est une base de données de bioactivité maintenue par la faculté de pharmacie *Skaggs* de l'université de San Diego. A l'origine, Cette base était destinée à ne répertorier que les valeurs d'affinités pour des complexes protéine-ligand cristallisés et présent dans la base

PDB (Liu *et al.* 2007). Elle a étendu son champ d'application en intégrant des données d'autres bases de bioactivité à savoir ChEMBL (essais sur les protéines), PubChem BioAssay (essais sur les protéines pour lesquels une courbe de dose réponse est présente) et la base entière PDSP Ki (http://www.bindingdb.org/bind/info.jsp). Cependant il est impossible de connaitre les versions des bases incluses et à quelle date elles ont été intégrées. Les données répertoriées directement par les responsables de la base sont extraites des publications scientifiques et pour les cibles d'intérêt thérapeutique. La base contient au mois de juillet 2012 environ 832 773 affinités correspondant à 5 765 protéines et 362 123 molécules. Des affinités pour 6 401 complexes sont issues de la base PDB. L'unité standard pour les valeurs d'affinités (Ki, Kd, IC50) est exprimée à l'aide d'une concentration en nano-molaire (Figure 6) .

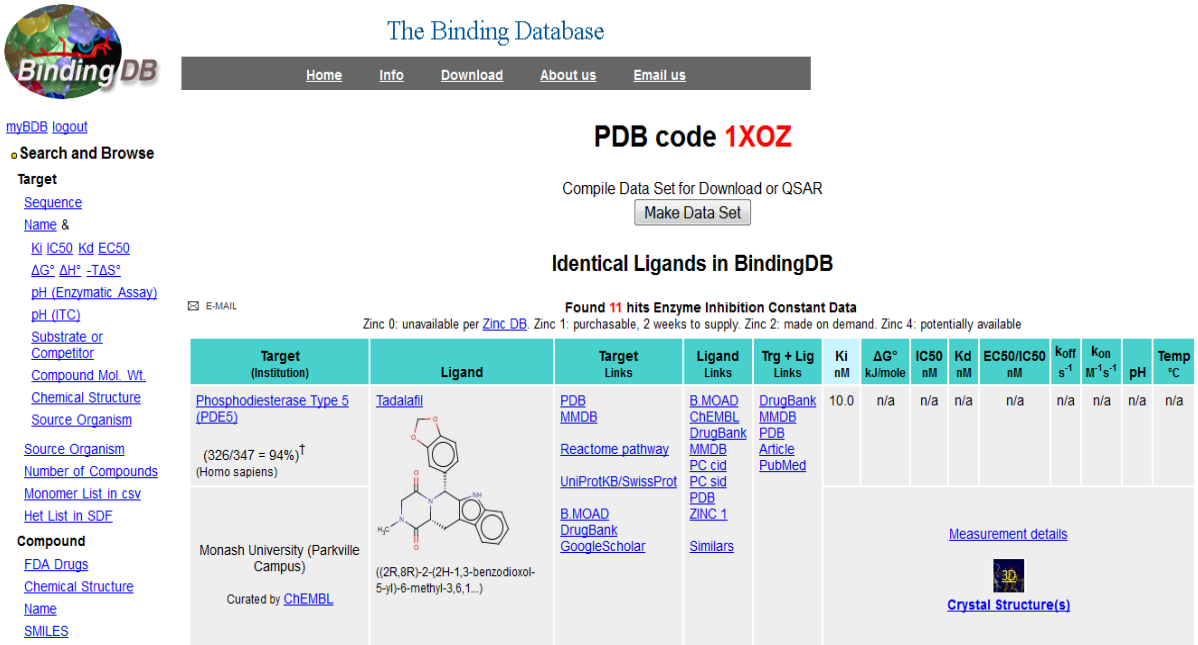


Figure 6 : Valeurs d'affinité pour le complexe 1xoz entre le Tadalafil et la protéine Phosphodiesterase 5.

Plusieurs références à d'autre bases sont également disponibles (PDB, ChEMBL, Uniprot, DrugBank). La base permet de faire des consultations par structure chimique pour les molécules (également recherche par similarité et sous-structure) ou par nom ou séquence pour les protéines.

La base est disponible au téléchargement à l'aide d'un fichier SDF ou d'un fichier texte CSV.

1.6. Les bases de données de bioactivité de complexes cristallographiques

Certaines bases de données se focalisent sur le partage d'informations d'activités mais que pour les complexes cristallographiques. Ces bases (Tableau 3) répertorient des valeurs d'affinités extraites manuellement de la littérature scientifique et dont la structure est répertoriée dans la base PDB.

Base de données	Nombre de complexes	Valeurs d'affinités	Portail web
Binding MOAD	16 984	5 630	http://bindingmoad.org
PDBbind-CN	7 986	7 986	http://pdbbind.org.cn
AffinDB	474	748 (plusieurs par complexe)	http://pc1664.pharmazie.uni-marburg.de/affinity
Protein Ligand Database (PLD)	359	+500 (plusieurs par complexe)	http://chemistry.st-andrews.ac.uk/staff/jbom/group/PLD.xls

Tableau 3 : Les bases de données de bioactivité de complexes cristallographiques.

1.6.1. Binding MOAD

Cette base de données est maintenue par l'Université de Michigan depuis 2004 (Benson, et al., 2008). Elle répertorie des complexes cristallographique protéine-ligand avec leurs affinités si celle-ci sont disponibles dans la littérature. Les complexes retenus dans Binding MOAD doivent posséder une résolution cristallographique inférieure ou égale à 2.5Å et le ligand ne doit pas être lié covalament à la cible. Le ligand peut être un peptide de moins de 10 acides aminés ou un oligonucléotide de moins de 4 paires de bases. Les molécules telles que les solvants et les sels sont écartés. Les cibles répertoriés sont classées par familles au sein desquelles les protéines partagent au minimum 90% d'identité de séquence, et classées par fonction dans le cas des enzymes en utilisant la classification enzymatique (http://enzyme.expasy.org/enzyme-byclass.html).

La base contient 16 984 complexes dont 5 630 pour lesquelles des valeurs d'affinité sont répertoriées. Un lien vers la base PDB est aussi indiqué ainsi que le lien PubMed pour la publication d'origine (Figure 7).

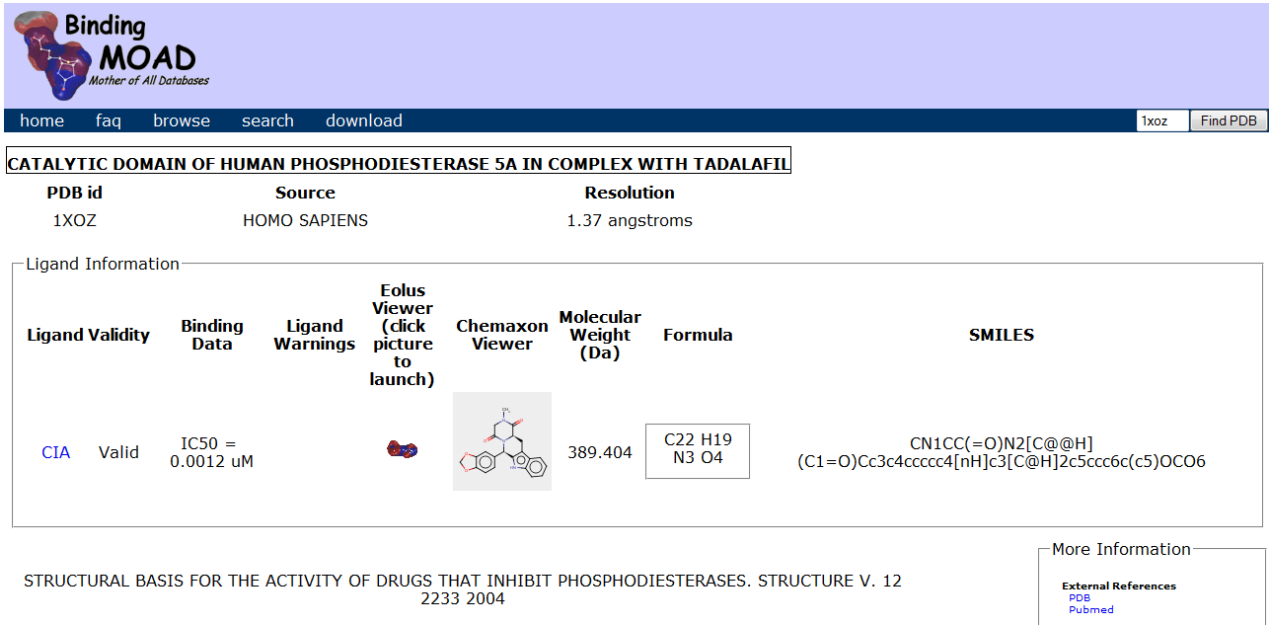


Figure 7 : Valeur d'affinité du complexe 1xoz entre le Tadalafil et la protéine Phosphodiesterase 5A.

Elle est consultable en ligne (http://bindingmoad.org) à l'aide de recherches textuelles en indiquant le code d'accession PDB ou bien à l'aide d'une exploration des cibles à partir de leur classification enzymatique. Le ligand pour lequel l'affinité est mesuré est représenté par son code HET et sa structure chimique (Figure 7).
Binding MOAD est disponible au téléchargement grâce à un fichier texte qui contient toutes les informations.

1.6.2. PDBbind-CN

PDBbind est aussi une base de données de bioactivité publiée en 2004 et développée grâce à une collaboration entre l'institut de chimie organique de Shanghai (Chine) et l'université du Michigan (Etats-Unis). Il existe deux miroirs pour la base, cependant celui qui est hébergé aux Etats-Unis n'est plus mis à jour depuis 2007 contrairement au site miroir chinois dont la dernière mise à jour date du mois de Septembre 2011.

La base répertorie des valeurs d'affinités (Ki, Kd et IC50) pour 7 986 complexes issus de la base PDB et consultable en ligne. La base peut être téléchargée sous format excel "XLS".
La disparité des données qui existent dans PDBbind et Binding MOAD s'explique par le fait que les bases de données ne traitent pas toutes les mêmes journaux et ni de la même façon. Binding MOAD utilise par exemple un algorithme de fouille de texte pour extraire les valeurs

d'affinités entre une molécule et une cible et c'est par la suite qu'une vérification par un expert est réalisée pour valider l'information extraite. PDBbind utilise une procédure manuelle où les experts sont en charge de l'identification de la publication et l'extraction des informations d'affinités. Néanmoins, il n'y a pas beaucoup de différences entre les données des deux bases. Elles se regroupent entre elles et le pourcentage de disparité entre les données est estimé à 3% seulement par les auteurs de PDBbind (Wang *et al.* 2005). Ces disparités peuvent subvenir car différentes sources sont utilisées dans le report des activités.
Les auteurs ont aussi estimé que le taux d'erreurs dans l'extraction et l'enregistrement des valeurs d'affinités dans PDBbind était de 1% (Wang *et al.* 2005).

1.6.3. AffinDB

AffinDB est à l'origine une base de données interne d'un laboratoire de l'université de Marburg (Allemagne). Les auteurs ont décidé d'en faire une base publique dès lors que les données commençaient à former une masse importante (Block *et al.* 2006). AffinDB contient des valeurs d'affinités pour des complexes connus pour calibrer les fonctions de scores d'arrimage moléculaire qui ont été supplémentées par la suite par quelques familles protéiques très étudiée comme les serines protéases et les anhydrases carboniques (Block *et al.* 2006). Les valeurs d'affinités sont extraites de publications scientifiques. La base contient aussi quelques valeurs issues de tests au sein même du laboratoire hôte (http://pc1664.pharmazie.uni-marburg.de).

La base est consultable en ligne à l'aide de recherche textuelle et des références sur les conditions expérimentales du test sont répertoriées si elles sont connues. Un lien de la publication originale vers PubMed est aussi disponible.

1.7. Evaluation du contenu des bases de données de bioactivité

Comme mentionné précédemment, plusieurs bases de bioactivité sont maintenant disponibles. En regardant de plus près les statistiques sur le nombre de molécules, de protéines et d'activités répertoriées dans ces bases, on s'aperçoit que le volume de données est assez conséquent.

Dès lors on peut se poser la question sur la qualité des données répertoriées et sur leur utilisation lors d'analyses de relation structure-activité :

- ✓ Est-il possible d'évaluer le pourcentage d'erreur dans ces bases ?
- ✓ À quel type d'erreurs doit-on faire face ?
- ✓ Quels est le pourcentage d'association ou d'interactions vraiment exploitable dans ces bases ?
- ✓ Peut-on évaluer le pourcentage de recouvrement entre celle-ci sachant qu'elles se basent pratiquement toutes sur des sources communes comme les journaux scientifiques?

1.7.1. Homogénéisation des données des bases de bioactivité

Afin de pouvoir exploiter les données issues des bases de données de bioactivité, il faut définir quelques règles selon lesquelles les données vont être standardisées afin qu'elles soient uniformes.

Les molécules doivent être représentées de la même façon, les valeurs d'activités doivent être exprimées à l'aide de la même unité et les cibles annotées à partir d'une base de données protéique (Uniprot ou GenBank).

Les molécules sont enregistrées dans les bases à l'aide d'une chaîne *smiles* ou un fichier de coordonnées sdf. La stéréo-isomérie est en général définie. Mais il se peut que cette information soit absente. Dès lors il faut écarter les molécules pour lesquelles cette information est manquante si on utilise des descripteurs 3D pour décrire les molécules, ou bien garder toutes les molécules si on utilise des descripteurs 2D. L'équilibre tautomérique doit être de même pris en considération. Il est envisageable que deux sources différentes citent des affinités en représentant certains tautomères d'une même molécule. Il est donc nécessaire de garder une seule forme, celle qui est la plus abondante par exemple, sinon le risque d'obtenir des doublons dans les analyses devient très probable.

Les protéines doivent être représentées à l'aide d'une annotation complète et le nom des protéines doit être harmonisé. Il est préférable d'utiliser les noms de protéines issus de bases comme Uniprot (UniProt 2012) ou RefSeq (Sayers *et al.* 2012) mais de ne pas les employer simultanément. Certaines bases de données utilisent les numéros d'accession Uniprot et d'autres utilisent des numéros gi (RefSeq). L'homogénéité des noms de protéines peut se faire alors à travers des fichiers mis à disposition dans ces bases et qui permettent de faire le

lien entre les numéros d'accession de différentes bases d'annotations protéiques (c.f. ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz). Il devient possible à ce moment-là d'utiliser une annotation protéique homogène.

Les valeurs d'affinités doivent être aussi converties en une concentration standard. Le plus souvent, ces activités se présentent sous forme de pKi, pKd, ou pIC50.

1.7.2. Analyse du contenu des bases de bioactivité

Nous avons extrait au laboratoire les informations d'affinités entre des molécules et des cibles dans le but de les utiliser pour du profilage sur ces cibles en utilisant l'information de leurs ligands. Ce profilage biologique se focalisera ainsi sur un grand nombre de cibles issues de ces bases (ce point sera abordé et discuté dans le chapitre 5).

Pour cette étude, on s'est intéressé aux plus grandes bases de données ChEMBL, PubChem BioAssay et IUPHAR-DB. Au commencement de ce projet en décembre 2011, la base ChEMBL était sous sa version 12, et la base PubChem BioAssay n'intégrait à l'époque que la version 11 de ChEMBL. Sachant que les procédures de traitement de données différent d'une base à une autre, nous avons téléchargé entièrement les trois bases et on a pris le soin d'exclure les données issues de PubChem et intégrées dans ChEMBL, les données de ChEMBL intégrées dans PubChem et les données d'IUPHAR-DB intégrées dans PubChem. Dès lors on est en mesure d'évaluer le contenu propre à chaque base.

Nous allons tenter de répondre aux questions énoncées au début du paragraphe grâce à l'analyse des données de bioactivité issues de ces trois bases. Il serait intéressant de savoir si les bases de données les plus importantes se recouvrent entre elles. Si c'est le cas, il est inutile de s'intéresser à plusieurs bases si une seule suffit à extraire les informations intéressantes.

Pour pouvoir effectuer des analyses, nous avons utilisé les activités extraites sous forme de valeurs de concentration d'inhibition IC50, constante d'association Ki et la constante de dissociation Kd. Ces constantes sont considérées comme des bons indicateurs de liaison entre un ligand et une protéine.

Le tableau 4 résume le nombre d'associations utilisables à partir de ces bases de données.

Base	Nombre d'activités total	Pourcentage d'affinités exploitables(*)
ChEMBL	6 000 000	7.5%
PubChem BioAssay	8 000 000	0.9 %
IUPHAR-DB	8 000	40.8 %

Tableau 4 : Nombre d'activités total sans traitement des données et pourcentage d'affinités exploitables parmi toutes les activités répertoriées dans la base. (*) Les affinités exploitables (valeurs de Ki, Kd, IC50) doivent correspondre à une structure bien définie de la molécule et une annotation non ambigüe de la cible. Les doublons sont écartés de l'analyse.

Cela peut surprendre, mais le pourcentage d'affinités exploitable (affinités exprimées en Ki, Kd et IC50) dans les deux plus grandes bases est en fait très mince.

Pour la base PubChem BioAssay, cela peut s'expliquer du fait qu'elle répertorie beaucoup plus d'essais issus de criblage à haut débit et que les données d'affinités sont très maigre.

Concernant la base ChEMBL, plusieurs valeurs d'associations sont inexploitables. C'est en général l'unité qui pose problème. On peut y trouver des concentrations exprimées en kilomètre/heure (et bien d'autres…), bref des données erronées qu'il faut écarter. Par la suite, si on ne garde que les cibles qui sont bien annotées (seuil de confiance ≥ 7), on se retrouve avec 7.5% des données qui sont réellement exploitables.

La base IUPHAR-DB quant à elle, ne contient pas un grand nombre d'activité mais c'est la base pour laquelle la moitié de données est utilisable. A l'évidence cette base privilégie la qualité à la quantité.

Si on regarde de plus près le recouvrement entre ses bases (Figure 8), on s'aperçoit que le nombre de molécules identiques entre les trois bases ne représente que 0.05% par rapport aux nombre total de molécules. Le nombre de cibles communes ne représente lui que 1.5% du nombre total.

Les articles scientifiques qui sont citées par les trois bases sont à l'évidence différents. Même si certains journaux sont indexés par les trois bases de données à la fois, Tiikkaienen et al. (Tiikkainen *et al.* 2012) ont remarqué que les périodes d'indexations n'étaient pas identiques. Ceci justifie le peu de recouvrement des données d'affinités (Figure 8 C, D). Par conséquent, les trois bases ne contiennent pas les mêmes informations, c'est ce qui justifie les différentes initiatives entre PubChem et ChEMBL afin de partager les données communes.

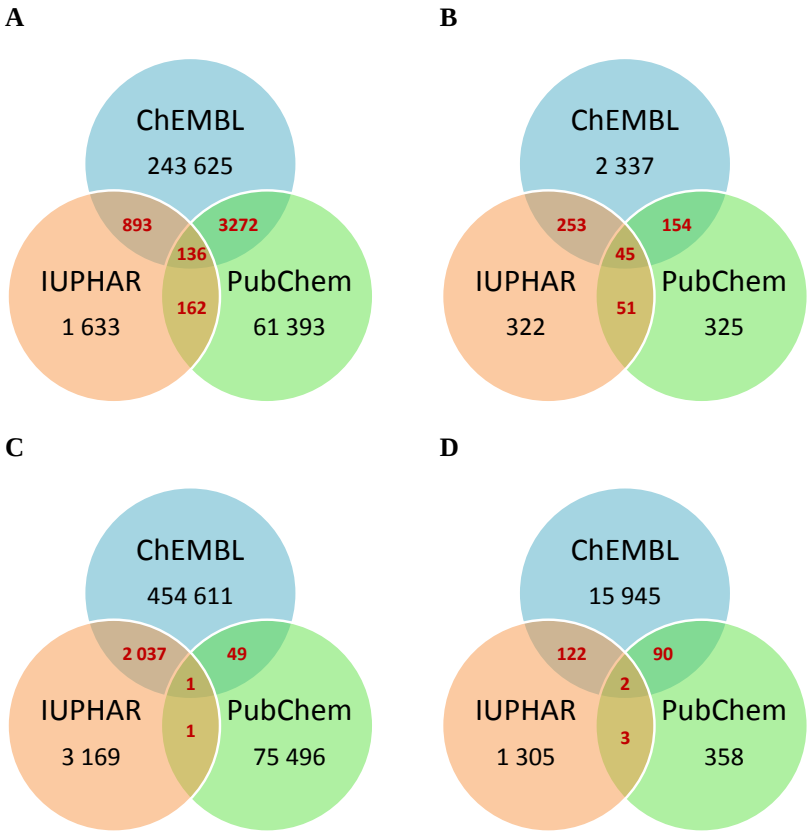


Figure 8 : Contenu des 3 bases et leur intersection en nombre de molécules, protéines et affinités. **(A)** intersections en nombre de molécules. **(B)** intersections en nombre de protéines. **(C)** intersections en nombre de valeurs d'affinité. **(D)** intersections des publications scientifiques d'où sont extraites les données.

Les cinq plus grandes familles de protéines à savoir les enzymes, les RCPG, les canaux ioniques, les récepteurs nucléaires et les transporteurs sont présentes avec les même proportions dans les bases PubChem BioAssay et ChEMBL. Contrairement à IUPHAR-DB où 69% des cibles sont principalement des RCPG (Figure 9).

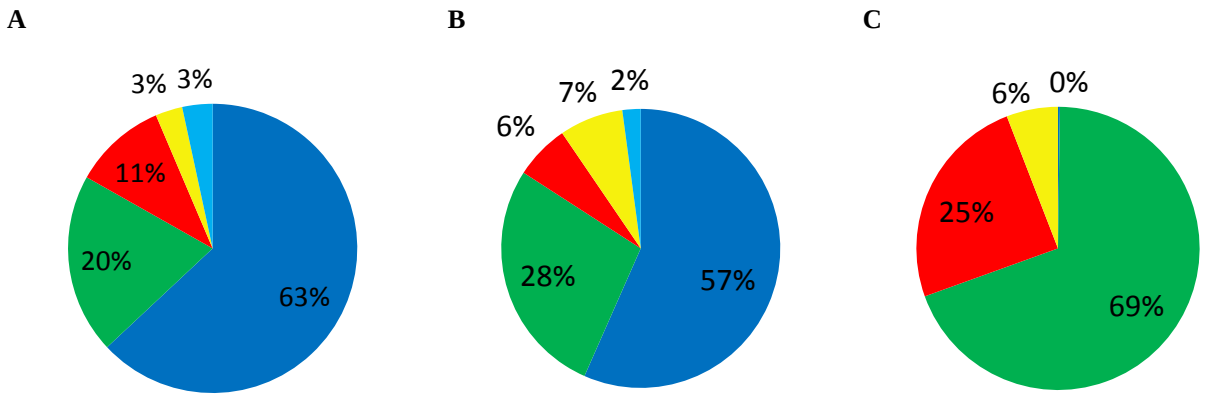


Figure 9 : Répartition de 5 grandes familles protéique dans les 3 bases. (Bleu : enzymes ; vert : RCPG, rouge : canaux ioniques ; jaune : récepteurs nucléaires ; cyan : transporteurs).
(A) ChEMBL. **(B)** PubChem BioAssay. **(C)** IUPHAR-DB.

1.7.3. Disparité entre les bases de bioactivité publiques et privées

Les bases de données publiques n'ont fait leur apparition que depuis 5 ou 6 ans. Autrefois, seules les bases de données commerciales étaient disponibles : Thomson Reuters Integrity (https://integrity.thomson-pharma.com/integrity/xmlxsl/) , Accelrys MDDR (http://accelrys.com/products/databases/bioactivity/mddr.html), WOMBAT (Olah *et al.* 2008) et Evolvus (http://www.evolvus.com/Products/Databases.html). Ces bases répertorient des associations protéine-ligand et certaines indiquent les valeurs d'affinités correspondantes. Il est donc intéressant de savoir si ces données commerciales se retrouvent dans les bases de données publiques telles que ChEMBL ou PubChem BioAssay.

Tiikkainen et al. (Tiikkainen *et al.* 2012) ont essayé de comparer les données extraites des bases publiques ChEMBL, PubChem BioAssay et PDSP Ki avec les données d'affinités répertoriées dans les bases commerciales Evolvus et WOMBAT. Les molécules et protéines extraites de toutes les bases ont été standardisées de la même manière qu'énoncée précédemment. En rassemblant toutes les données d'activités, les auteurs ont remarqués que 44,6% de ces activités proviennent des bases de données commerciales Evolvus et WOMBAT. Ceci nous révèle que les données issues des bases de données publiques et commerciales sont complémentaires. Cette observation s'explique entre autres par le fait qu'Evolvus est la seule base à indexer les brevets dans cette étude, même si les données d'affinités qui y sont extrait ne représentent que 36% de données contenu dans la base

Evolvus. De plus, les valeurs d'affinités que contiennent les bases de données commerciales sont relevées des journaux scientifiques depuis les années 1990, tandis que les bases publiques ne les répertorient en général qu'à partir des années 2000.

Il est à l'évidence conseillé de disposer des grandes bases publiques et commerciales lors d'études de relations structure-activité.

1.7.4. Erreurs inhérentes aux bases de données de bioactivité

a. Les erreurs de structures

Une des erreurs les plus fréquentes dans les bases de données est l'erreur sur la structure chimique des molécules. Certaines molécules possèdent des *smiles* erronés et donc illisibles lors des traitements. D'autres contiennent des valences incorrectes, des atomes manquants. Certaines molécules chirales ne possèdent pas de stéréo-isomérie définie. Ce type d'erreur doit être anticipé avant les analyses et il faut standardiser la représentation selon le type de descripteur qu'on va utiliser par la suite.

Voici un exemple d'erreur (Figure 10) où la structure répertoriée est fausse (Tiikkainen *et al.* 2012). Cette molécule qu'on nommera CHEMBL611822 est active sur le récepteur A1 bovin de l'adenosine (numéro d'accession Uniprot : P28190) avec une constante d'association Ki=0.085µM (Cappellacci *et al.* 2005). La structure mentionnée dans la publication originale (Cappellacci *et al.* 2005) est celle répertoriée par la base WOMBAT. Cette structure possède un atome de Chlore en position *meta* de la pyrimidine. La structure répertoriée dans la base ChEMBL est fausse car le Chlore est manquant. Quant à la structure répertoriée par la base Evolvus, la stéréo-isomérie n'est pas définie. Cet exemple permet d'illustrer la difficulté d'extraction des informations des publications scientifiques ou des brevets car les molécules sont représentées à l'aide de structures de Markush (Barnard 1991) et les substituants sont indiqués soit dans des tableaux ou dans le texte. Ceci nécessite une vérification très minutieuse pour éviter ce type d'erreur.

A) ChEMBL **B)** WOMBAT **C)** Evolvus

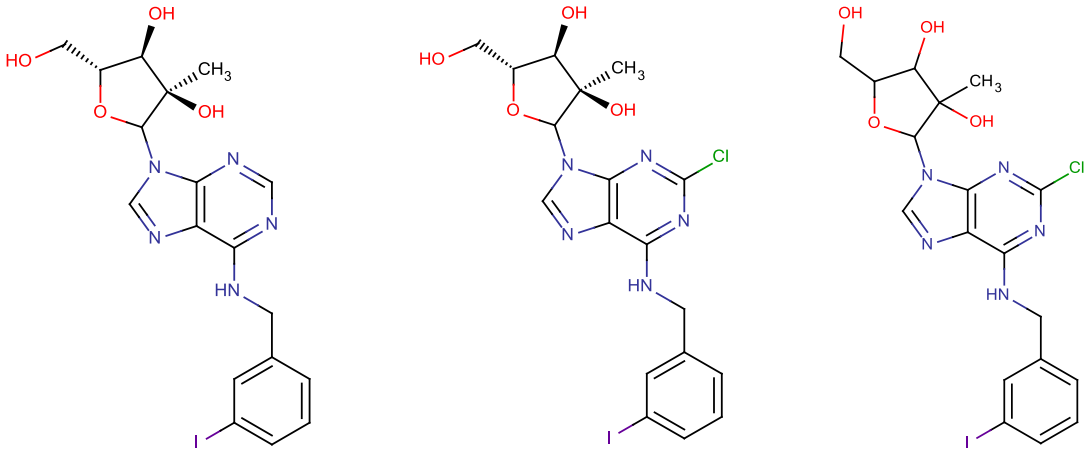


Figure 10 : Exemple d'un cas où la représentation des molécules diffère selon la base.

b. Les erreurs d'annotation des protéines

Lors de l'extraction des données, certaines erreurs se produisent concernant l'annotation de la protéine. Les espèces peuvent être modifiées et les protéines homologues mal identifiées.

En voici un exemple prélevé de la publication de (Tiikkainen *et al.* 2012) avec la molécule CHEMBL368061 qui se lie au récepteur 5- hydroxytryptamine-2A (5-HT-2A ou récepteur de Sérotonine 2A) bovin (el Ahmad *et al.* 1997). On remarque à partir de la figure 11 que l'annotation des récepteurs est fausse dans deux cas pour ChEMBL, fausse pour WOMBAT et correcte pour Evolvus. La base WOMBAT répertorie un autre orthologue du récepteur 5-HT-2A qui appartient à l'espèce des rats, et la base ChEMBL l'associe avec deux autres paralogues 5-HT-2B et 5-HT-2C de l'espèce humaine qui ne sont pas mentionnés dans la publication (el Ahmad *et al.* 1997). C'est pour quoi on recommande de ne considérer que des associations pour lesquelles un seuil de confiance est supérieur ou égal à 7, même si on risque d'exclure certaines bonnes associations, ceci permet de travailler sur des données fiables et correctes.

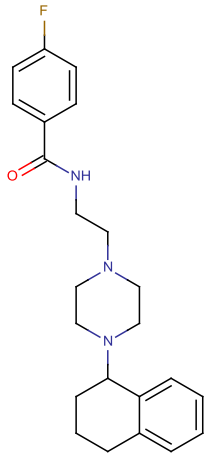
Molécule CHEMBL368061	Cibles associés avec Ki = 1.38 µM		
	**ChEMBL**	**WOMBAT**	**Evolvus**
	✖ P28335 : Nom : 5-HT-2C Seuil de confiance = 6 Espèce : Humaine ✖ P41595 : Nom : 5-HT-2B Seuil de confiance = 6 Espèce : Humaine ✓ Q75Z89 : Nom : 5-HT-2A Seuil de confiance = 6 Espèce : Bovine	✖ P14842 : Nom : 5-HT-2A Espèce : Rat	✓ Q75Z89 : Nom : 5-HT-2A Espèce : Bovine

Figure 11 : Exemple d'un cas où les cibles diffèrent selon la base de données et citant la même source (el Ahmad *et al.* 1997)

c. L'incertitude expérimentale

Au-delà des erreurs de saisies lors de l'enregistrement des informations dans les bases de données de bioactivité, le taux d'erreurs expérimentales demeure inaccessible et non quantifié. Ceci peut impacter la qualité des données dans les bases en particulier pour les mesures qui sont faites dans des laboratoires différents et avec des conditions expérimentales différentes.

Une récente étude (Kramer *et al.* 2012) a permis de lever le voile sur cette incertitude expérimentale en se basant sur les données d'activités répertoriées dans la base ChEMBL (version 12). Toutes les données de constante d'association exprimées en pKi de paires d'interaction protéine-ligand de la base ChEMBL ont été extraites. Les auteurs stipulent que l'incertitude expérimentale est la même quelle que soit l'origine de la mesure et qu'elle est indépendante des molécules ou de l'annotation biologique de la cible.

Premièrement, les paires de mesures pour lesquelles la différence de pKi était de 3 et de 6 ont été écartées afin d'éviter la comparaison de valeurs assurément erronées pour des erreurs de saisies sur les unités des concentrations (exemple : µM en nM). Deuxièmement, les paires de mesures pour lesquelles la différence de pKi était inférieure ou égal à 0,02 ont été écartées en vue d'exclure les mesures pour lesquelles une erreur de précision dans l'enregistrement des valeurs s'est produite. Enfin, un ensemble de 11 621 paires de mesures d'affinité

représentant 2 540 complexes protéine-ligand a été retenu pour déterminer l'incertitude expérimentale à l'aide des paramètres statistiques (Kramer *et al.* 2012) du tableau 5.

Paramètre statistique	Formule
Erreur moyenne absolue	$MUE = \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \lvert y_{i,1} - y_{i,2} \rvert$
Déviation standard	$\sigma_E = \sqrt{\frac{1}{2(n-1)} \sum_{i=1}^{n} (y_{i,1} - y_{i,2})^2}$
Coefficient de Pearson	$R^2_{Pearson} = \sum_{i=1}^{N} (y_{i,1} - \overline{y_1})(y_{i,2} - \overline{y_2}) \Big/ \left[\sqrt{\sum_{i=1}^{N} (y_{i,1} - \overline{y_1})^2} \times \sqrt{\sum_{i=1}^{N} (y_{i,2} - \overline{y_2})^2} \right]$

Tableau 5 : Paramètres statistiques pour l'évaluation de l'incertitude expérimentale.

L'erreur expérimentale moyenne estimée à partir de cet ensemble est de 0.44 unité de pKi pour une déviation standard de 0.56. (R^2=0.82). Cette estimation est très importante surtout lorsqu'on développe des modèles QSAR prédictifs. Lors de l'évaluation de la performance des modèles générés, les modèles possédant des erreurs moyenne absolue inférieures à 0.44 unité de pKi sont probablement des modèles surentrainés qu'il faudra écarter ou modifier.
En regardant de plus près la distribution des variations sur les valeurs d'affinité en chaque point de l'intervalle des mesures (pKi ∈ [3,12]), on s'aperçoit que les molécules de basse et haute affinité disposaient du même taux d'incertitude.

Une autre observation révélée dans cette étude était que le taux de variation sur les mesures augmenterait en fonction de l'augmentation de la surface polaire de la molécule, de sa masse moléculaire et si son coefficient de partage ClogP était soit très petit, soit très grand. En l'occurrence la taille et les propriétés physicochimiques de molécules testées ont un effet sur la mesure d'affinité.

1.8. Conclusion

Nous venons d'énumérer les bases de données de bioactivité publique et nous avons détaillé leur contenu. Une question se pose alors sur l'utilité de l'utilisation de toutes ces bases. Les plus importantes comme ChEMBL et PubChem BioAssay (hors données HTS) commencent à fusionner leur contenu et les deux bases ressembleront à deux miroirs d'ici quelques années.

De notre point de vue et à partir des analyses que nous avons réalisé précédemment (cf. 2.8.2), il suffirait de se limiter aux données de bioactivité répertoriée dans ChEMBL et ceci pour plusieurs raisons. Premièrement, cette base possède une structure de données basée sur un schéma relationnel conçu à partir d'un modèle MVC (Modèle-Vue-Contrôleur : http://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller). Ce ci rend les données très facile à extraire contrairement à la base PubChem BioAssay qui malgré les efforts de standardisation, répertorie les essais sous un format de données qui n'est généralement pas respecté par tous les fournisseurs de la base. Dès lors il devient très fastidieux de récupérer les informations des fichiers d'essais. Ensuite, le seuil de confiance répertoriée dans ChEMBL permet d'exclure les mesures d'affinité incertaines pour les protéines dont l'annotation biologique est ambiguë. Il n'y a malheureusement pas d'équivalent dans la base PubChem BioAssay.

Les utilisateurs de ses bases doivent être très vigilants lors de la procédure d'extraction des données et nous recommandons de suivre les étapes décrites dans le paragraphe 1.7.1.

1.9. Références

Barnard, J. M. (1991). "A comparison of different approaches to Markush structure handling." J Chem Inf Comput Sci **31**(1): 64-68.

Benson, D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell and E. W. Sayers (2012). "GenBank." Nucleic Acids Research **40**(D1): D48-D53.

Berman, H., K. Henrick and H. Nakamura (2003). "Announcing the worldwide Protein Data Bank." Nat Struct Biol **10**(12): 980.

Block, P., C. A. Sotriffer, I. Dramburg and G. Klebe (2006). "AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB." Nucleic Acids Research **34**(Database issue): D522-526.

Cappellacci, L., P. Franchetti, M. Pasqualini, R. Petrelli, P. Vita, A. Lavecchia, E. Novellino, B. Costa, C. Martini, K. N. Klotz and M. Grifantini (2005). "Synthesis, biological evaluation, and molecular modeling of ribose-modified adenosine analogues as adenosine receptor agonists." Journal of Medicinal Chemistry **48**(5): 1550-1562.

Davis, A. P., B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wiegers and C. J. Mattingly (2011). "The Comparative Toxicogenomics Database: update 2011." Nucleic Acids Research **39**(Database issue): D1067-1072.

el Ahmad, Y., E. Laurent, P. Maillet, A. Talab, J. F. Teste, R. Dokhan, G. Tran and R. Ollivier (1997). "New benzocycloalkylpiperazines, potent and selective 5-HT1A receptor ligands." Journal of Medicinal Chemistry **40**(6): 952-960.

Garcia-Serna, R., O. Ursu, T. I. Oprea and J. Mestres (2010). "iPHACE: integrative navigation in pharmacological space." Bioinformatics **26**(7): 985-986.

Gaulton, A., L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Research **40**(Database issue): D1100-1107.

Hecker, N., J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, M. K. Gilson, P. E. Bourne and R. Preissner (2012). "SuperTarget goes quantitative: update on drug-target interactions." Nucleic Acids Research **40**(Database issue): D1113-1117.

Hodge, A. E., R. B. Altman and T. E. Klein (2007). "The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge." Clinical Pharmacology & Therapeutics **81**(1): 21-24.

Keiser, M. J., V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth (2009). "Predicting new molecular targets for known drugs." Nature **462**(7270): 175-181.

Knox, C., V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart (2011). "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic Acids Research **39**(Database issue): D1035-1041.

Kramer, C., T. Kalliokoski, P. Gedeck and A. Vulpetti (2012). "The Experimental Uncertainty of Heterogeneous Public K(i) Data." Journal of Medicinal Chemistry **55**(11): 5165-5173.

Kuhn, M., M. Campillos, I. Letunic, L. J. Jensen and P. Bork (2010). "A side effect resource to capture phenotypic effects of drugs." Mol Syst Biol **6**.

Kuhn, M., D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen and P. Bork (2012). "STITCH 3: zooming in on protein-chemical interactions." Nucleic Acids Research **40**(Database issue): D876-880.

Liu, T., Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic Acids Research **35**(Database issue): D198-201.

Magarinos, M. P., S. J. Carmona, G. J. Crowther, S. A. Ralph, D. S. Roos, D. Shanmugam, W. C. Van Voorhis and F. Aguero (2012). "TDR Targets: a chemogenomics resource for neglected diseases." Nucleic Acids Research **40**(Database issue): D1118-1127.

Okuno, Y., A. Tamon, H. Yabuuchi, S. Niijima, Y. Minowa, K. Tonomura, R. Kunimoto and C. Feng (2008). "GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update." Nucleic Acids Research **36**(Database issue): D907-912.

Olah, M., R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mractc and T. I. Oprea (2008). WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. Chemical Biology, Wiley-VCH Verlag GmbH**:** 760-786.

Overington, J. (2009). "ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European

Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr." J Comput Aided Mol Des **23**(4): 195-198.

Pihan, E., L. Colliandre, J. F. Guichou and D. Douguet (2012). "e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design." Bioinformatics **28**(11): 1540-1541.

Roth, B. L., E. Lopez, S. Patel and W. K. Kroeze (2000). "The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches?" Neuroscientist **6**(4): 252-262.

Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko and J. Ye (2012). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research **40**(Database issue): D13-25.

Scheer, M., A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele and D. Schomburg (2011). "BRENDA, the enzyme information system in 2011." Nucleic Acids Research **39**(Database issue): D670-676.

Seiler, K. P., G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons (2008). "ChemBank: a small-molecule screening and cheminformatics resource database." Nucleic Acids Research **36**(Database issue): D351-359.

Sharman, J. L., C. P. Mpamhanga, M. Spedding, P. Germain, B. Staels, C. Dacquet, V. Laudet, A. J. Harmar and I. Nc (2011). "IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data." Nucleic Acids Research **39**(Database issue): D534-538.

Taboureau, O., S. K. Nielsen, K. Audouze, N. Weinhold, D. Edsgard, F. S. Roque, I. Kouskoumvekaki, A. Bora, R. Curpan, T. S. Jensen, S. Brunak and T. I. Oprea (2011). "ChemProt: a disease chemical biology database." Nucleic Acids Research **39**(Database issue): D367-372.

Tanabe, M. and M. Kanehisa (2012). "Using the KEGG Database Resource." Curr Protoc Bioinformatics **Chapter 1**: Unit1 12.

Tiikkainen, P. and L. Franke (2012). "Analysis of commercial and public bioactivity databases." J Chem Inf Model **52**(2): 319-326.

UniProt, C. (2012). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Research **40**(Database issue): D71-75.

von Eichborn, J., M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne and R. Preissner (2011). "PROMISCUOUS: a database for network-based drug-repositioning." Nucleic Acids Research **39**(Database issue): D1060-1066.

Wang, R., X. Fang, Y. Lu, C. Y. Yang and S. Wang (2005). "The PDBbind database: methodologies and updates." Journal of Medicinal Chemistry **48**(12): 4111-4119.

Wang, Y. L., J. W. Xiao, T. O. Suzek, J. Zhang, J. Y. Wang, Z. G. Zhou, L. Y. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant (2012). "PubChem's BioAssay Database." Nucleic Acids Research **40**(D1): D400-D412.

2. Méthodes *in silico* pour le profilage biologique de petites molécules

Un médicament est toute molécule qui engendre à travers son mode d'action, un changement phénotypique chez un individu en réponse à une pathologie. Il se peut que ce dernier puisse entrainer en outre des changements phénotypiques inattendus et indésirables qu'on nomme effets secondaires. Ces derniers constituent l'une des principales causes de rejet de molécules candidats-médicament en phase clinique. Il est donc nécessaire d'anticiper et d'identifier ces derniers afin d'éviter des tests cliniques inutiles et coûteux. Les effets secondaires sont divers et variés et sont principalement dûs à la non sélectivité du médicament. On peut quantifier la sélectivité d'une molécule par le nombre de cibles à laquelle elle se lie. Par conséquent, il est nécessaire de connaitre toutes les possibilités d'association entre cette molécule et les cibles présentes dans l'organisme.

Avant la commercialisation de tout médicament, des tests *in vitro* sont réalisés sur une catégorie de cibles particulières pour évaluer sa toxicité ou ses éventuels effets secondaires. Le coût des tests *in vitro* reste cependant très élevé en particulier lors de criblages à haut débit où l'on teste plusieurs milliers de molécules. C'est à cette étape qu'il est nécessaire d'intervenir afin de diminuer ces coûts en écartant au préalable les molécules pour lesquelles on pourrait prédire un effet toxique ou indésirable.

Les outils de chémoinformatique (Varnek *et al.* 2011) interviennent dans la phase préclinique et permettent l'élimination de molécules susceptibles d'être rejetées en phase clinique à cause de leur toxicité ou de leur non spécificité. Grâce à ces outils, des criblages virtuels peuvent être réalisé afin d'identifier des touches intéressantes (Sotriffer C. 2011). Elles ont entre autres facilité la construction de chimiothèques focalisées par cible thérapeutique, l'analyse des données de relation structure-activité mais aussi la prédiction des propriétés ADME et de toxicité.

Le criblage virtuel (Figure 12) est couramment utilisé afin d'identifier des molécules actives et constitue un réel complément au criblage à haut débit (Schneider 2010). Ce processus

implique qu'on dispose d'une chimiothèque que l'on crible sur une cible pour identifier des possibilités d'association.

Le criblage virtuel inverse (Figure 12) quant à lui, nous renseigne sur le profil biologique d'une molécule. Cette dernière est criblée sur plusieurs cibles et une liste de cibles potentielles est obtenue. L'identification de ces associations constitue une recherche importante car elles permettent l'anticipation de certains effets secondaires susceptibles d'être observés lors d'essais cliniques, mais également d'évaluer la toxicité potentielle d'une molécule. Dans le cadre d'un profilage de médicament, l'identification d'une autre cible peut induire à son utilisation pour un autre effet thérapeutique. C'est ce qu'on appelle le repositionnement de médicaments.

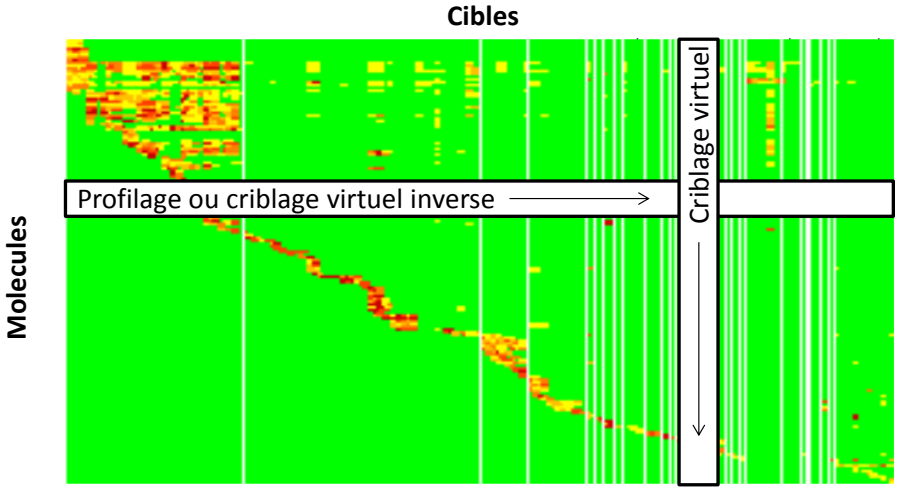


Figure 12 : Matrice chémogénomique extraite de la base iPhace (Garcia-Serna *et al.* 2010). Les molécules sont représentées en ligne et les cibles en colonne. Chaque point de la matrice correspond à une association entre une molécule et une cible. L'affinité des associations décroit de la couleur rouge à la couleur jaune, les points verts représentent des associations non identifiées expérimentalement.

Le profil biologique d'une molécule peut être obtenu d'une manière expérimentale (*tests in vitro*) ou d'une manière virtuelle (*méthodes in silico*).

A l'aide de méthodes in silico, le profil d'une molécule peut être établi en cherchant des molécules qui lui sont similaires et dont on connait les cibles, puis supposer que cette dernière se lie à ces mêmes cibles. L'hypothèse ainsi formulée stipule que les molécules similaires se lient à des cibles similaires (Keiser *et al.* 2007; Klabunde 2007). Certes, ce principe est une vision sommaire. En effet, il est connu que certaines molécules similaires n'interagissent pas de la même manière avec une cible donnée (Martin *et al.* 2002).

On peut également établir le profil biologique d'une molécule en établissant la liste des cibles similaires aux cibles déjà connues pour celle-ci. Le principe se résume au fait que les cibles similaires doivent reconnaitre les mêmes ligands (Klabunde 2007).

Néanmoins, le profil d'une molécule peut également être obtenu en évaluant son affinité avec des cibles quelconques. Cette association peut être prédite à travers plusieurs méthodes en utilisant à la fois la structure de la molécule profilée et la structure de la cible.

2.1. Profilage utilisant les méthodes basées sur les ligands

Une association entre une molécule et une cible thérapeutique peut être prédite d'une manière déductive en se basant sur la structure des molécules se liant à celle-ci. En effet, les molécules de structures similaires tendent à avoir des activités biologiques similaires dans la majorité des cas.

La similarité entre molécules se mesure de plusieurs façons et selon les descripteurs utilisés (Bender *et al.* 2004). Une similarité des propriétés physicochimiques ou une similarité structurale peut ainsi être ciblée selon le type de descripteur choisi (Todeschini R. 2000). Le choix des descripteurs est à l'évidence très important dans tout type d'étude et il doit être adapté à la problématique adressée.

Les approches les plus utilisées pour la similarité entre molécules peuvent être groupées dans deux catégories : une similarité basée sur le graphe moléculaire défini par la table de connectivité qu'on désigne par les approches à deux dimensions (*2D*) ; et la similarité basée sur les coordonnées atomiques des molécules qu'on désigne par les approches à trois dimensions (*3D*). Il existe certes des descripteurs à une dimension (compte du nombre d'atomes, formule brute, etc…) mais ceux-ci sont inutilisables lors d'un criblage virtuel ou lors d'un profilage car la similarité qu'ils définissent n'est pas adaptée à ce genre de problématique.

2.1.1. Profilage utilisant des descripteurs 2D

Les descripteurs 2D se classent dans deux grandes catégories. La première appartient à la catégorie de ceux qui découlent du graphe moléculaire, et la seconde est définie par les

propriétés physico-chimiques de celle-ci. Les descripteurs qui découlent du graphe moléculaire peuvent être représentés sous forme numérique à l'instar des descripteurs topologiques mais aussi sous forme d'une suite d'entiers qu'on appelle généralement l'empreinte moléculaire. Dans les deux cas, une métrique est utilisée pour mesurer la similarité selon le type de descripteur.

Dans le cas des empreintes moléculaires, le coefficient de Tanimoto (Tableau 1) est généralement utilisé pour mesurer la similarité entre deux empreintes, et la distance Euclidienne ou le coefficient de Pearson (Tableau 6) dans le cas où des descripteurs numériques sont utilisés. Evidemment, il en existe bien d'autres et une revue détaillée sur les métriques ainsi que leurs applications a été publiée par Willett et al. (Willett *et al.* 1998).

Métrique	Formule	Détails
Tanimoto	$c/(a+b-c)$	c: bits communs allumés sur empreintes 1 et 2 a: bits allumés sur l'empreinte 1 b: bits allumés sur l'empreinte 2
Distance Euclidienne	$\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}$	x_i : descripteur i de la molécule 1 y_i : descripteur i de la molécule 2
Coefficient de Pearson	$\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \Big/ \left[\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\right]$	x_i : descripteur i de la molécule 1 y_i : descripteur i de la molécule 2

Tableau 6 : Métriques fréquemment utilisées pour mesurer la similarité entre deux molécules

En règle générale, la similarité entre la molécule *i* à profiler est calculée avec toutes les molécules actives *j* d'une cible *A*. Si une des molécules actives *j* lui est similaire, alors la cible *A* est une cible potentielle pour la molécule *i* (Figure 13). Ceci implique la définition d'un seuil de similarité qui est en général déterminé empiriquement ou bien à partir de la distribution des similarités des molécules actives de la cible.

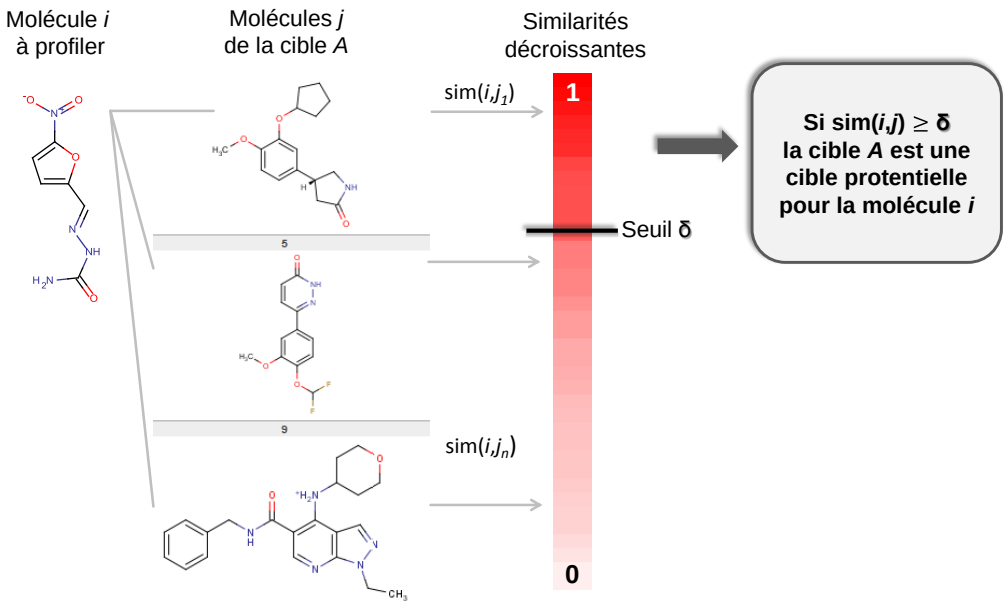


Figure 13 : Principe du profilage en utilisant la similarité des ligands

Le profilage en utilisant la similarité 2D est certainement la méthode de profilage la plus utilisée. Plusieurs méthodes existent (Tableau 7), la plus connue étant la méthode SEA (*Similarity Ensemble Approach*) publiée par Keiser et al. (Keiser *et al.* 2007).

Méthode	Description	Descripteurs
Bender et al. (Bender *et al.* 2006)	Probabilité d'activité évaluée par des modèles bayésiens multi-classes	Empreintes moléculaires ECFP_4
Similarity Ensemble Search, SEA (Keiser *et al.* 2007)	Valeur d'expectation calculée à partir d'une classe de molécules actives.	v1: Empreintes moléculaires Daylight v2: Empreintes moléculaires ECFP_4
Inverse Distance Weighting, IDW (Vidal *et al.* 2010)	Prédiction d'affinité par les N plus proches voisins	Empreintes moléculaires SHED
Profile-QSAR (Martin *et al.* 2011)	Modèles PLS pour prédire l'affinité	Empreintes moléculaires FCFP_6, aLogP, masse moléculaire, nombre de donneurs/accepteurs de liaisons H, nombre d'angles de torsion

Tableau 7 : Sélection de méthodes de profilage pharmacologique utilisant des descripteurs 2D de ligands.

Bender et al. (Bender *et al.* 2006) ont extrait des activités pour 1 003 classes d'activités à partir de la base de bioactivité WOMBAT (Olah *et al.* 2008). Les molécules actives pour une

classe sont utilisées pour construire un modèle Bayésien naïf de classification. Ce dernier utilise les molécules issues d'autres classes d'activités comme des molécules inactives. C'est de là que découle l'appellation de modèle Bayésien multi-classes ou multi-catégories (Nidhi *et al.* 2006). Un modèle pour chaque classe d'activité a été construit à l'aide d'empreintes moléculaires ECFP_4 (Rogers *et al.* 2010). Un rappel de 35% a été obtenu en moyenne sur les classes d'activité étudiées. Les auteurs ne se sont pas vraiment focalisés sur le profilage pharmacologique à l'aide des modèles générés, mais ils ont utilisées les prédictions de ces modèles pour créer une empreinte moléculaire basée sur la probabilité d'association de la molécule sur chacune de ces classes d'activités. Les empreintes en question ont été appelées empreintes moléculaires d'affinités. Ces empreintes permettent d'obtenir de meilleurs enrichissements (en moyenne ± 23.6% en valeur relative) par rapport aux empreintes structurales ECFP_4 surtout pour certaines classes d'activités où les molécules sont très diverses.

Keiser et al. ont introduit la méthode SEA (Keiser *et al.* 2007) qui se base sur un principe statistique utilisé par BLAST (Altschul *et al.* 1990), qui est un outil de bioinformatique pour les alignements de séquences. Cette méthode utilise des espérances mathématiques pour évaluer la similarité entre deux ensembles de ligands. Des variables centrées réduites (*score-Z*) de similarité sont calculées pour deux ensembles de ligands en y incluant un ensemble aléatoire de ligands, ce qui permet de contourner le poids que porte le nombre de ligands présents dans chaque ensemble. La matrice des scores-*Z* obtenue est modélisée par une loi des valeurs extrêmes et un seuil permettant l'obtention du meilleur ajustement par rapport à cette distribution est retenu. Ce seuil est enfin traduit en espérance mathématique qui traduit la probabilité d'observer aléatoirement une similarité entre les deux ensembles. Les cibles retenues dans la première version de SEA sont au nombre de 246, principalement des enzymes, des RCPG, des canaux ioniques et des récepteurs nucléaires issus de la base de bioactivité Accelrys MDDR (http://accelrys.com/products/databases/bioactivity/mddr.html consulté en Juin 2012). Néanmoins, la version actuelle inclut des cibles des bases ChEMBL (Gaulton *et al.* 2012), WOMBAT (Olah *et al.* 2008), MDDR (http://accelrys.com/products/databases/bioactivity/mddr.html)(http://accelrys.com/products/databases/bioactivity/mddr.html) et de la KEGG (Kanehisa *et al.* 2012) et leur nombre est estimé à environ 2 000 (Laggner *et al.* 2012). Seules les cibles qui possèdent plus de 5 ligands différents

sont prises en compte par la méthode. Celle-ci a été utilisée afin de soutenir l'hypothèse permettant de relier les cibles entre elles à l'aide de la similarité de leurs ligands. Ce concept a été exprimé par Paolini et al. (Paolini *et al.* 2006) qui ont défini l'espace pharmacologique humain à l'aide d'un réseau de graphes reliant et regroupant ainsi les cibles similaires. En effet, pour la méthode SEA, les auteurs ont construit un réseau de graphes où chaque nœud correspond à une cible. Les nœuds sont reliés entre eux à l'aide de la similarité de leurs ligands et permettent de démontrer que les familles de protéines se regroupent convenablement (Figure 14).

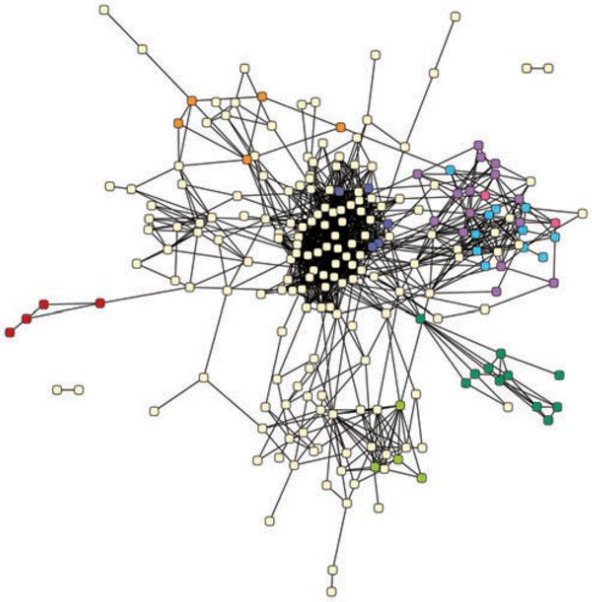


Figure 14 : Représentation en réseau de l'espace pharmacologique définit dans SEA. Les nœuds représentent les cibles coloriées par famille protéique : les antifolates (rouge), les phosphodiésterases (orange), les opioïdes (bleu), les bêta-lactames antibiotiques (vert foncé), les sérotonergiques métabotropiques (violet), les sérotonergiques ionotropiques (rose), les adrénergiques (cyan) et les modulateurs d'estrogène (vert clair).
Figure extraite de l'article (Keiser et al. 2007)

De plus, cette méthode permet d'identifier des cibles nouvelles pour une molécule en la profilant sur l'ensemble des ligands disponibles. En effet, en s'appuyant sur cette méthode, les auteurs ont identifié 23 nouvelles interactions confirmées expérimentalement entre des médicaments déjà sur le marché et des RCPG aminergiques (Keiser *et al.* 2009).

Dans une seconde étude, les auteurs (Laggner *et al.* 2012) ont pu identifier et confirmer 11 autres associations. Un ensemble de composés neuro-actifs sur le modèle du poisson zèbre ont été profilé sur 2 000 cibles que propose la méthode SEA. Après identification de nouvelles

associations prédites, les auteurs ont validé expérimentalement 11 associations sur des RCPG, des neurotransmetteurs, des canaux de potassiques et des kinases.

Cette méthode bien qu'elle a prouvé son efficacité à travers les applications citées, souffre d'une limite. Le fait de choisir un seuil d'espérance pour lequel on décide de la séparation entre des associations possibles et des associations aberrantes peut conduire à quelques faux positifs ou à quelques faux négatifs. SEA enfin est une méthode qui ne s'applique pas aux cibles orphelines ou pour lesquelles très peu de ligands sont disponibles.

Vidal et al. (Vidal *et al.* 2010) ont introduit une méthode appelée *Inverse Distance Weighting* (IDW) qui consiste à prédire l'affinité d'une molécule à partir de ses *k* plus proches voisins (Figure 15). L'affinité pour une cible est définie comme la somme pondérée des affinités des molécules voisines dans l'espace chimique (Equation 1).

$$F_i = \frac{\sum_j^k \left(d_{ij}^{-2} \times F_j\right)}{\sum_j^k d_{ij}^{-2}} \qquad (1)$$

F_i : affinité de la molécule *i* à prédire
F_j : affinité de la molécule *j* déjà connue
d_{ij} : distance entre la molécule *i* et la molécule *j*
k : nombre de plus proches voisins

La méthode IDW a été validée en profilant 13 médicaments antipsychotiques sur 34 cibles (Roth *et al.* 2004) dont 30 RCPG et 3 neurotransmetteurs. Les auteurs ont utilisé trois types d'empreintes moléculaires pour définit l'espace chimique : SHED (Gregori-Puigjane *et al.* 2006), FPD (Vidal *et al.* 2011) et PHRAG (Vidal *et al.* 2011). La combinaison des trois empreintes pour identifier les *k* proches voisins de la molécule à profiler améliore la prédiction des valeurs d'affinités.

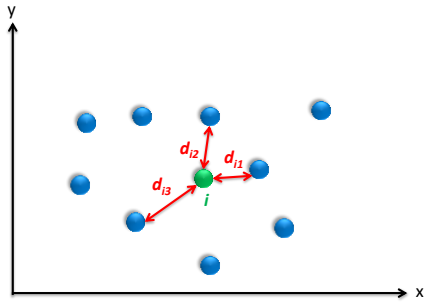


Figure 15 : Prédiction de la valeur d'affinité de la molécule *i* à partir des *k* proches voisins (*k*=3 sur la figure). La molécule *i* à profiler est en vert. Les molécules de l'ensemble d'entrainement sont représentées en bleu. Les axes x et y représentent un espace chimique sur lequel ces molécules sont définies.

En effet, en utilisant cette approche, les auteurs ont eu la possibilité de faire des prédictions pour 70% de l'ensemble de la matrice d'association. Parmi ces prédictions, 65% des affinités étaient prédites avec une erreur inférieure à une unité logarithmique (pKi) et avec une précision de 93%.

Martin et al. (Martin *et al.* 2011) ont utilisé des modèles Bayésiens naïfs de classification pour 92 protéines kinases afin de prédire l'activité de nouvelles molécules sur une nouvelle protéine kinase (non communiquée). 130 000 molécules ont été décrites avec des empreintes moléculaires FCFP_6 (Rogers *et al.* 2010) ainsi que cinq autres descripteurs à savoir aLogP, la masse moléculaire, le nombre de donneurs/accepteurs de liaisons hydrogène et le nombre d'angles de torsion. Ces descripteurs ont été utilisés pour construire cinq modèles Bayésiens afin de modéliser l'affinité pour chacune des 92 protéines kinases. La médiane des coefficients de corrélation sur un ensemble de données externes représentant 25% des données d'interactions était de 0.59. Ces modèles peuvent servir pour un profilage pharmacologique classique sur les kinases, mais les auteurs les ont également utilisés pour modéliser l'activité à l'aide d'une régression PLS d'une autre protéine kinase. Cette technique est identique à celle utilisée précédemment par Bender et al. (Bender *et al.* 2006).

Les auteurs ont également étendu cette approche pour prédire des activités cellulaires en entrainant les modèles Bayésiens non pas sur les valeurs d'affinités mais sur des valeurs d'efficacité pEC50 de 42 essais cellulaires de protéines kinases. Un ensemble de données externes contenant 25% des données récoltés a été utilisé pour évaluer ces modèles. Un coefficient de corrélation médian R^2=0.58 a été obtenue pour 24 essais cellulaires de modulation (*target modulation assay*) et un coefficient médian R^2=0.41 obtenue pour 18 essais cellulaires de prolifération (*cell proliferation assay*).

L'avantage de cette méthode est que le profil engendré pour chaque molécule sur les 92 protéines kinases peut nous permettre d'évaluer sa sélectivité sur cette portion du kinôme. Du fait que les sites des protéines kinases se ressemblent, prédire la sélectivité des inhibiteurs de protéines kinases est un axe de recherche très convoité car beaucoup d'entre elles sont impliquées dans des maladies comme le cancer.

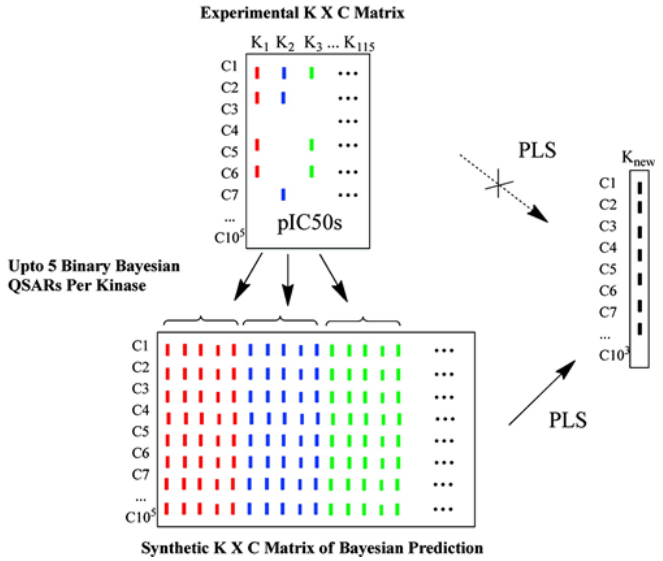


Figure 16 : Méthode Profile-QSAR. La matrice du haut est la matrice expérimentale d'association entre les molécules et les protéines kinases du jeu d'entrainement. Cette matrice sert à produire 5 modèles Bayésiens pour chaque protéine kinase et ce profil est utilisé à l'aide d'une régression PLS pour prédire l'association à une nouvelle protéine kinase (K_{new}).
Figure extraite de l'article de (Martin *et al.* 2011).

Les méthodes de profilage utilisant la similarité 2D des ligands se distinguent par leur simplicité et leur rapidité. Il est possible de profiler des millions de molécules par jour sur une seule machine.

En revanche, comme toutes autres méthodes, ce type de profilage a des inconvénients. Lorsqu'on utilise des descripteurs structuraux, on n'arrive à établir un profil biologique que pour les ligands dont les structures sont similaires. Si le châssis moléculaire de la molécule à profiler présente une faible similarité structurale mais quelques points pharmacophoriques identiques, celle-ci peut s'avérer active sur la cible mais ne sera pas identifié comme telle par la méthode.

De plus le principal inconvénient réside évidemment dans le choix du seuil de similarité à définir pour discriminer les molécules actives et inactives. Même si des critères statistiques permettent le choix d'un seuil adapté, celui-ci reste toujours dépendant des structures des ligands déjà connus pour la cible à profiler.

D'autre part, ces méthodes ne sont performantes que lorsqu'un nombre important de ligands divers est connu. Certaines cibles thérapeutiques intéressantes ont peu de chémotypes connus et

l'application de ces méthodes peut présenter un risque d'échec dans l'identification de nouvelles molécules avec un châssis différent.

2.1.2. Profilage utilisant des descripteurs 3D

Les descripteurs 3D utilisent les coordonnées atomiques de la molécule pour la décrire. On peut les regrouper en trois catégories : les descripteurs géométriques (n-tuplets de pharmacophores), les descripteurs de formes (représentées par des Gaussiennes ou des harmoniques sphériques) et les descripteurs de grilles (CoMFA, VolSurf). La plupart des études de profilage utilisent soit les descripteurs de forme (volume occupé dans l'espace) ou les descripteurs de grilles (Tableau 8).

Le principe du profilage reste le même que celui utilisant la structure 2D (Figure 13). Une molécule est comparée à des molécules annotées et si l'une d'entre elle lui est similaire, la cible correspondante est identifiée comme cible potentielle.

Méthode	Description	Descripteurs
ROCS (AbdulHameed *et al.* 2012)	Profilage à l'aide d'une représentation Gaussienne de la forme moléculaire	Volume à partir de la somme de sphères Gaussiennes atomiques
PARASURF (Perez-Nueno *et al.* 2011)	Profilage à l'aide d'une représentation en harmoniques sphériques de la forme moléculaire	Volume à partir de la somme de sphères harmoniques atomiques
Topomer/CoMFA (Wendt *et al.* 2011)	Profilage à l'aide d'une représentation en grille	Fragments Topomer sur une grille CoMFA représentant des descripteurs stériques et électrostatiques.
ReverseScreen3D (Kinnings *et al.* 2011)	Profilage à l'aide d'une représentation floue du graphe moléculaire	Graphe moléculaire traduit en combinaison de triplets codant le type atomique.

Tableau 8 : Sélection de méthodes de profilage pharmacologique utilisant des descripteurs 2D de ligands

Abdulhameed et al. (AbdulHameed, et al., 2012) utilisent le programme ROCS (OpenEye) pour identifier les cibles secondaires à partir d'un ensemble de 245 cibles décrites par leurs 1150 ligands (appelés *représentatifs*) connus extraits de la base DrugBank (Knox *et al.* 2011). Les cibles incluses sont des RCPG, des canaux ioniques, des récepteurs nucléaires, des transporteurs

et des enzymes. Les auteurs ont tout d'abord validé le fait que ces ligands *représentatifs* arrivaient bien à discriminer les molécules actives de chaque cible en criblant les molécules actives et inactives de 40 cibles issues de la base de données DUD (Huang *et al.* 2006).

Dans un premier temps, un criblage virtuel classique des molécules actives et inactives a été réalisé indépendamment pour chaque cible. Et dans un second temps, un criblage virtuel est effectué sur une chimiothèque où les molécules actives sont mélangées à toutes les molécules inactives des autres cibles. Dans 77% des cas, le criblage est réussi et présente une aire sous la courbe ROC supérieur à 0.7. La performance de ces criblages étant satisfaisantes, les auteurs ont par la suite profilé 14 médicaments dont les cibles secondaires étaient répertoriées dans la littérature. Le critère utilisé pour évaluer le succès de la méthode est le rang des cibles secondaires parmi la liste des 245 cibles. Si la cible secondaire est répertoriée parmi les 5% premières de la liste (12 premières cibles), le profilage est jugé comme performant. Le succès du profilage était de 74% soit 10 cibles secondaires identifiées sur un total de 14.

Perez-Nueno et al. (Perez-Nueno and Ritchie, 2011) ont évalué la promiscuité des 2950 ligands de la base de données DUD sur les 40 cibles de celle-ci. Une matrice de comparaisons des ligands a été générée en comparant les formes issues d'un calcul des sphères harmoniques atomiques. Un algorithme de classification a été ensuite utilisé pour permettre la formation de classes de ligands de formes similaires. Cette procédure permet d'identifier les ligands promiscuitaires, ceux pour lesquels la forme est similaire à beaucoup d'autres ligands extraits de cibles différentes. Cependant, cette étude était beaucoup plus focalisée sur la comparaison des performances de cette méthode avec celle du logiciel ROCS (OpenEye) et d'arrimage avec le logiciel GOLD (Jones *et al.* 1997).

Wendt et al. (Wendt *et al.* 2011) ont construit des modèles PLS pour prédire l'affinité à partir de descripteurs électrostatiques et stériques issues de la méthode CoMFA (Cramer *et al.* 1988). Des interactions ont été traitées à partir de 7 bases de bioactivités publiques : PubChem (Li *et al.* 2010), ChemBank (Seiler *et al.* 2008), ChEMBL (Gaulton *et al.* 2012), Binding DB (Liu *et al.* 2007), PDSP Ki (http://pdsp.med.unc.edu/pdsp.php consulté en Juin 2012), Binding MOAD (Benson *et al.* 2008) et AffinDB (Block *et al.* 2006). Pour chaque cible, une molécule *centroïde* est définie comme étant la plus similaire aux autres et fragmentée en énumérant tous les doublets

possibles de fragments *topomers* (fragments composés d'au moins trois atomes). Ces doublets sont par la suite comparés à tous les doublets de fragments *topomers* des autres molécules de la même cible et ceux qui leurs sont similaires sont retenus. Des descripteurs CoMFA sont par la suite projetés sur les doublets de fragments *topomers* retenus. Finalement, ces descripteurs sont ensuite utilisés pour construire un modèle de régression PLS en les corrélant avec l'affinité des molécules. Sur 5 718 cibles, les auteurs ont gardé les modèles qui ont un coefficient de corrélation supérieur à 0.2 conduisant donc à des modèles pour 1 795 cibles.

Quatre médicaments ont été profilés et l'identification des propriétés des fragments *topomers* responsables dans l'identification de la cible connue ont été analysés. Malheureusement, les auteurs n'ont pas discuté et quantifié le nombre de cibles retrouvées pour chaque molécule. Ils se sont plutôt focalisés sur l'analyse des fragments *topomers* et leurs propriétés en relation avec l'activité. Ceci est difficilement saisissable avec des descripteurs 2D car la méthode Topomer/CoMFA arrive à identifier les fragments avec leurs contributions stériques et électrostatiques responsables de l'activité.

Kinnings et al. (Kinnings *et al.* 2011) utilisent une représentation triangulaire du graphe moléculaire pour chaque conformation. Tous les triplets pharmacophoriques (type atomique) sont énumérés en appliquant des seuils de distances et comparés avec ceux d'une autre molécule. La similarité est évaluée à l'aide d'un coefficient de Tanimoto. Les auteurs ont validé la méthode en profilant 20 inhibiteurs extraient de la base DrugBank sur un ensemble de 6 041 cibles issues de la PDB. Une comparaison des enrichissements de la méthode *ReverseScreen3D* avec la méthode *ReverseScreen2D* (utilisation d'empreintes moléculaires Daylight (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html consulté en Juin 2012)) a été effectuée. L'enrichissement à 1% des cibles profilées pour les 20 molécules était de 34 pour la méthode *ReverseScreen2D* et de 41 pour la méthode *ReverseScreen3D*. Les enrichissements à 2%, 5% et 10% étaient équivalents pour les deux méthodes : 22, 11 et 7. La méthode *ReverseScreen3D* arrivait à mieux classer les cibles connues par rapport à la méthode *ReverseScreen2D* grâce à la similarité 3D.

Cette observation est en réalité un peu biaisée par le descripteur 2D utilisé à savoir l'empreinte moléculaire Daylight. En effet, cette empreinte est très utile dans le cas de recherches sous-

structurales et a tendance à bien évaluée des petits châssis moléculaires du fait de leur présence dans une molécule plus grande.

Il faut souligner que le dernier point observé par Wendt et al. (Wendt *et al.* 2011) est très important. La similarité à l'aide de descripteurs 3D permet d'identifier plus facilement les cibles secondaires par rapport à une similarité utilisant des descripteurs 2D. Les molécules qui sont structuralement similaires (similarité 2D) et qui ont une forme similaire (similarité 3D) vont avoir le même effet pharmacologique et probablement les mêmes points d'ancrage et d'interactions. Les molécules ne partageant pas de similarités structurales 2D mais qui partageant une similarité selon des descripteurs 3D (forme et propriétés pharmacophoriques), auront tout de même tendance à partager la même activité pharmacologique.

Lors d'un profilage, il serait donc plus facile d'identifier des cibles secondaires dans certains cas à l'aide d'une similarité 3D. Cette observation a été démontrée un peu plus tard par Yera et al. (Yera *et al.* 2011) en profilant 358 médicaments sur une série de RCPG et de canaux ioniques. Les auteurs avaient observés que la similarité 2D et 3D avaient des performances similaires lors d'identification des cibles principales, mais qu'un apport notable de la similarité 3D a été observé pour l'identification de cibles secondaires. Les auteurs concluent qu'il est toujours préférable d'associer les deux approches lors d'un profilage et que la fusion des probabilités des cibles potentielles serait plus judicieuse.

Évidemment, bien qu'elles soient performantes, ces méthodes ont quelques limites. Premièrement, leur domaine d'applicabilité se limite aux cibles qui ont des ligands connus. Deuxièmement, le fait de fixer un seuil de similarité peut parfois être contraignant. Même si certaines méthodes contournent le problème en transformant la similarité en probabilité par rapport à une distribution de similarité aléatoire.

Et enfin, une similarité 3D nécessite le calcul de conformères et un alignement 3D des molécules ce qui peut être plus long qu'une simple similarité 2D. Sachant que pour les molécules possédant plus d'une dizaine d'angles de torsions, l'espace conformationnel n'est en général pas très bien couvert, ceci limite le succès des méthodes de profilage 3D à retrouver les bonnes cibles.

2.2. Profilage utilisant les méthodes basées sur les cibles

La polypharmacologie d'un ligand peut également être prédite seulement à partir de sa cible principale ou l'une de ces cibles déjà connues.

Le principe consiste à admettre que des cibles similaires auront tendance à partager des ligands similaires (Kuhn *et al.* 2008). Cette hypothèse a été vérifiée et quantifiée pour 140 protéines hélicoïdales (1er classe de la classification hiérarchique CATH (Greene *et al.* 2007) de domaines protéiques) issues de la PDB (Mitchell 2001).

Le pourcentage d'identité et de similarité de ces protéines corrélait assez bien avec la similarité de leurs ligands respectifs (Figure 17).

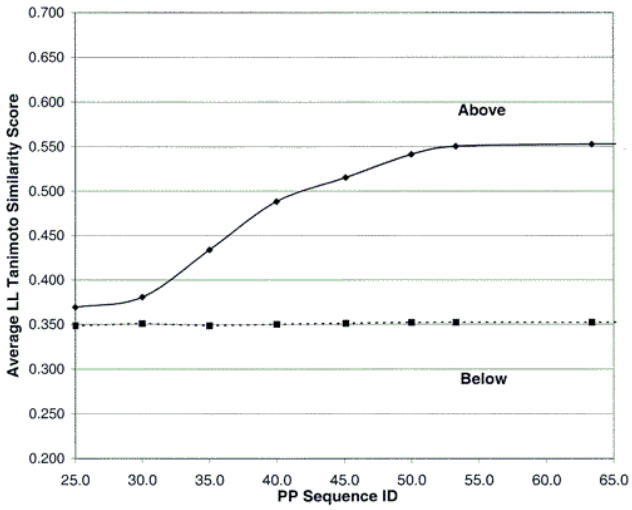


Figure 17 : Moyenne des similarités des paires de ligands au-dessus (ligne continue) et en dessous (ligne discontinue) des pourcentages d'identité de séquences correspondant (axe des abscisses).
Figure extraite de l'article (Mitchell 2001)

Les principes de la théorie d'évolution peuvent aussi expliquer cette observation. Les protéines qui partagent un ancêtre commun très proche, vont conserver une forte identité de séquence et vont présenter des changements très mineurs concernant la spécificité de leurs ligands.

Ceci est également très visible dans la famille des RCPG où la probabilité pour qu'un même ligand se lie à deux RCPG non similaires est quasi nulle particulièrement quand le pourcentage de leurs identité de séquence est inférieur à 30% (Kuhn *et al.* 2008). On peut alors profiler une série de ligands lorsqu'on connait une de leurs cibles. La cible connue est ainsi utilisée pour en

rechercher d'autres qui lui sont similaires et qui auront une forte probabilité de lier le ligand profilé.

La similarité entre plusieurs cibles peut être évaluée par deux approches : la première en s'appuyant sur la similarité de leur séquence protéique (structure primaire représentée par une liste d'acides aminés formant la séquence/similarité globale), et la deuxième approche en se basant sur la structure tertiaire ou structure tridimensionnelle de la cible (site de liaison/similarité locale). Dans le cadre du profilage, la deuxième approche a été utilisée maintes fois avec succès (Tableau 9).

Méthode	Description	Descripteurs	Alignement Structural
SOIPPA (Xie *et al.* 2009)	Profilage à l'aide de la similarité structurale des sites de liaison	Site représentée par un graphe issu d'une triangulation de Delaunay	nécessaire
FlapSite (Milletti *et al.* 2010)	Profilage à l'aide de la similarité structurale des sites de liaisons	Empreintes pour chaque atome du site de liaison codant des propriétés pharmacophoriques	nécessaire
SiteAlign (Defranchi *et al.* 2010)	Profilage à l'aide de la similarité structurale des sites de liaisons	Empreintes de sites de liaisons codant des propriétés pharmacophoriques et géométriques	nécessaire

Tableau 9 : Sélection de méthodes de profilage utilisant la similarité des protéines

Xie et al. (Xie *et al.* 2009) utilisent un graphe pour représenter une protéine. Chaque nœud du graphe représente un carbone C_α d'un résidu de la chaine protéique. Deux protéines sont comparées à l'aide d'un alignement qui optimise la correspondance des sous graphes communs. Une fonction de densité Gaussienne est utilisée pour évaluer la similarité entre deux protéines.
Les auteurs ont profilé 5985 structures issues de la base PDB en utilisant la protéine de transfert du cholestérol estérifié (*Cholesteryl Ester Transfer Protein CETP*) comme référence (code pdb : 2obd). Ceci a permis l'identification de 276 protéines similaires à la protéine CETP et susceptibles de contenir des inhibiteurs potentiels pour celle-ci. Les auteurs ont utilisé l'arrimage moléculaire (*molecular docking*) pour évaluer la possibilité d'association de trois inhibiteurs connus (Figure 18) de la protéine CETP avec les cibles potentielles dans le but d'identifier les effets indésirables que peuvent causer ses inhibiteurs. Les auteurs n'ont pas analysé toutes les

cibles secondaires obtenues mais ils ont identifié celles qui étaient responsables de trois effets cliniques connus à savoir l'hypertension, l'inflammation et le cancer. Ces effets ont été reliés à leurs voies biologiques (*biological pathways*) à travers les cibles secondaires identifiées.

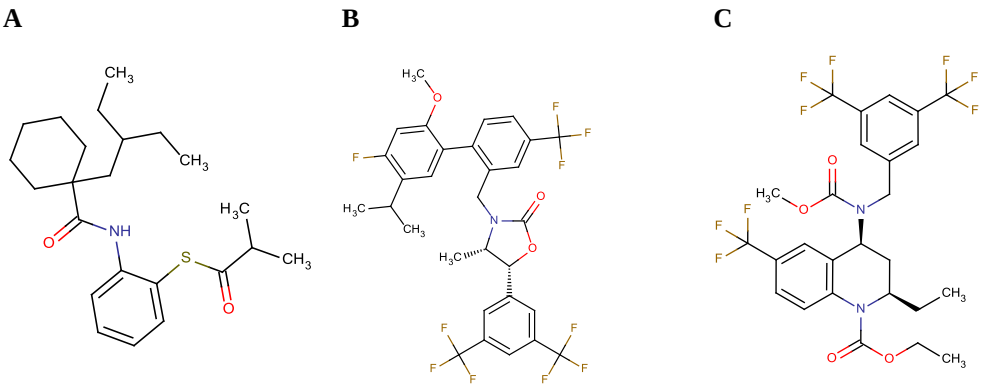


Figure 18 : Trois inhibiteurs connus de la CETP. **(A)** Dalcetrapib ou JTT-705. **(B)** Anacetrapib. **(C)** Torcetrapib

Milletti et al. (Milletti *et al.* 2010) ont profilé 17 inhibiteurs de protéines kinases sur 189 protéines kinases. Cet ensemble a été recueilli à partir d'un profilage expérimental de 38 inhibiteurs sur environ la moitié du kinôme humain à savoir 317 protéines kinases (Karaman *et al.* 2008).

Les 189 protéines kinases retenues pour l'étude appartiennent à 1 647 entrées de la base de données cristallographiques PDB (Berman *et al.* 2003) et rassemblent 3 957 sites de liaisons différents déterminés à l'aide du programme FlapSite (MolecularDiscovery).

Une empreinte est générée pour chaque atome du site (sans molécules d'eau) selon les propriétés pharmacophoriques des atomes voisins (hydrophobes, aromatiques, charges, donneur/accepteur de liaisons hydrogène). Les atomes voisins sont déterminés à partir de leur appartenance à 13 sphères ayant pour centre l'atome à coder, et pour rayon de 2 Å (1ère sphère) jusqu'à 16.8 Å (13ème sphère). Une métrique de similarité incluant la taille des sites à comparer et le RMSD calculé à partir de leur alignement permettent d'évaluer leur similarité. Les 17 sites des inhibiteurs de protéines kinases ont servi de références pour les comparaisons avec les 3 957 autres sites. Des courbes ROC ont été utilisées pour mesurer la performance de l'approche. Les vrais positifs ont été définis comme étant les sites d'inhibition présentant une affinité K_d <

10µM, et les faux positifs comme étant les sites d'inhibition présentant une affinité $K_d \geq 10\mu M$. A 10% d'enrichissement ROC (taux de vrais positifs à 10% de faux positifs), 37% des protéines kinases de haute affinité ont été retrouvées.

Les auteurs ont également profilé la molécule d'ATP (adénosine triphosphate) à l'aide de six représentations de sites de liaisons différents issus de protéines différentes sur toutes les cavités détectées à partir des entrées PDB. Ils ont pu observer ainsi, que l'ATP se liait à des sites de cibles similaires issus de conformations de structures secondaires différentes.

Defranchi et al. (Defranchi *et al.* 2010) ont profilé la Staurosporine (Figure 19) en utilisant comme référence le site de liaison de la Pim-1 kinase (code pdb: 1yhs) sur un total de 6 415 sites de liaisons extraits de la base de données cristallographiques sc-PDB (Meslamani *et al.* 2011). Cet inhibiteur étant connu pour être un inhibiteur non sélectif de protéines kinases, celles-ci ont été écartées des comparaisons avec la Pim-1 kinase. La liste des sites similaires est obtenue avec le programme SiteAlign (Schalon *et al.* 2008). Ce dernier positionne une sphère discrétisée en 80 triangles sur lesquels sont projetés des descripteurs pharmacophoriques et géométriques qui vont former une empreinte décrivant le site.

A **B**

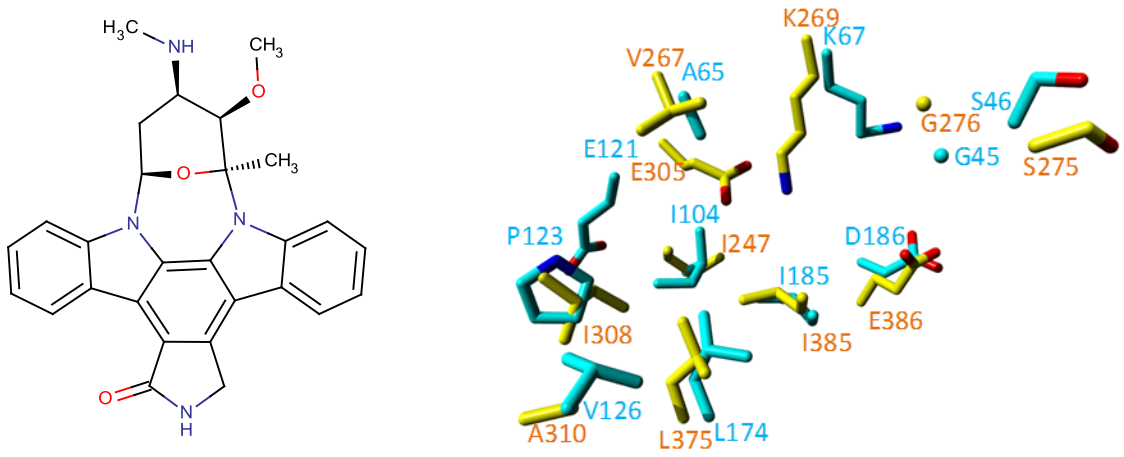


Figure 19 : **(A)** Structure de la Staurosporine. **(B)** superposition des sites de liaisons de la Pim-1 kinase (1yhs, cyan) et de la synapsine I (1aux, jaune). Figure B est extraite de l'article (Defranchi et al. 2010).

Lors d'une comparaison entre deux sites, ceux-ci sont alignés afin de maximiser la similarité calculée à partir des deux empreintes. Le meilleur alignement est celui qui correspond à la

similarité maximale des deux sites. Parmi toute la liste des sites de liaisons comparés, une protéine disposant d'un site de liaison similaire à celui de la Pim-1 kinase a retenu l'attention des auteurs. Ce site est celui de la Synapsine-1 qui appartient à la famille des Synapsines. Cette protéine se trouve dans le système nerveux central et périphérique et a pour rôle la régulation des neurotransmissions aux synapses. L'affinité de la Staurosporine avec la Synapsine-1 a été validée expérimentalement à l'aide d'un test in vitro indiquant une IC_{50} de 0.3µM. Les auteurs ont testé par la suite huit autres inhibiteurs de protéines kinases dont les sites de liaisons étaient plus ou moins similaires à celui de la Synapsine-1. Ces derniers présentaient une affinité à la Synapsine-1 corrélant bien avec le degré de similarité entre les sites de liaisons. Les inhibiteurs de sites très similaires à celui de la Synapsine-1 (CDK2, Caséine kinase II, Pim-1) présentent des affinités nanomolaires pour la Synapsine-1, tandis que les inhibiteurs de sites distants (ex : Checkpoint kinase 1, HSP-90 alpha, Diacylglycerol kinase) ne se lient pas à la contre-cible.

Le profilage utilisant les sites de liaison est donc très performant. Ces méthodes sont intéressantes car elles ne dépendent pas des ligands et sont par conséquent applicables à des cibles orphelines à partir de moment où l'on connait leur séquence protéique ou leur structure tridimensionnelle.

Cependant elles ne sont applicables qu'à une petite partie du protéome car beaucoup de protéines ne disposent pas de structures cristallographiques.

Une autre limite réside dans la sensibilité de la méthode à un éventuel changement conformationnel adopté par la cible lors d'une association. Ce changement conformationnel peut induire l'apparition d'un site de liaison complétement différent de celui déterminé à partir d'un autre ligand.

Ensuite, les chaines latérales susceptibles d'être sélectionnées pour définir un site de liaison peuvent induire du bruit lors de la comparaison entre deux sites, et par conséquent la similarité calculée va être tronquée. Une attention particulière doit être accordée lors de la sélection des résidus qui définissent le site de liaison. On peut néanmoins s'affranchir de cette limite en utilisant des méthodes de comparaisons qui se basent sur les atomes C_α des résidus (Weill *et al.* 2010).

Enfin, plusieurs protéines disposent de plusieurs sites de liaisons possibles (exemple des protéines kinases avec un site catalytique et un site allostérique). Il faut inclure les deux sites lors d'un profilage pour permettre de tester toutes les hypothèses possibles de liaisons.

2.3.Profilage utilisant les méthodes hybrides et chémogénomiques

Il est également possible de profiler un ligand sur une série de cibles en se basant à la fois sur sa représentation et sur celles des cibles à identifier (Rognan 2010).

Lors du profilage d'un ligand, toutes les cibles sont criblées indépendamment les unes des autres, et les associations possibles sont ainsi identifiées. L'arrimage moléculaire (*molecular docking*) ou les modèles de pharmacophores 3D en sont un bon exemple. A grande échelle ces méthodes s'apparentent aux approches chémogénomiques (Rognan 2007) où toutes les associations entre ligands et cibles sont investiguées dans le but d'être déterminées.

Les machines d'apprentissage permettent d'utiliser des représentations de plusieurs complexes protéines-ligands pour identifier d'éventuelles associations entre un nouveau ligand et des cibles. Les complexes sont représentés à l'aide des descripteurs de leurs ligands et de leurs cibles, et la machine d'apprentissage est entrainée dans le but de créer un ou plusieurs modèles prédictifs. Cette approche vise à réunir l'espace chimique du ligand et celui des cibles *in toto*.

Différentes études ont utilisées les approches citées, et nous allons commenter celles qui ont été réalisées pour un profilage à grande échelle.

2.3.1. Profilage utilisant l'arrimage moléculaire

Bien que l'arrimage moléculaire ait été longtemps considéré comme une méthode très coûteuse en temps de calcul lorsqu'on l'applique à grande échelle, les infrastructures de calcul à haut débit nous permettent aujourd'hui d'arrimer plusieurs centaines de molécules sur des milliers de cibles en quelques jours seulement. Ceci permet d'utiliser cette méthode à des fins de profilage biologique de petites molécules.

Le principe de l'arrimage réside en la prédiction du mode de liaison de la molécule avec la cible (Figure 19). Cette prédiction est évaluée par une fonction de score qui tient compte de la

nature des interactions non-covalentes et de l'énergie libre de liaison (dans certains cas) de la molécule avec les résidus du site de liaison de la cible (Rognan 2011).

Certaines études ont utilisé cette approche avec succès. Yang et al. (Yang *et al.* 2009) dédient un serveur web pour l'identification des effets secondaires à partir de matrices d'arrimage moléculaire à 79 cibles répertoriées comme responsables de ses effets.

A l'évidence, l'arrimage moléculaire souffre d'une limite particulière dûe à la fonction qui évalue l'association entre la molécule et la cible. Ces fonctions sont généralement construites et validées à partir d'un ensemble de cibles choisies pour cet effet. Dès lors qu'elles sont utilisées sur des cibles différentes et par conséquent en dehors du domaine d'applicabilité, le comportement de ces fonctions est incertain et la prédiction du mode de liaison trouvé n'est pas certaine d'être bien évaluée.

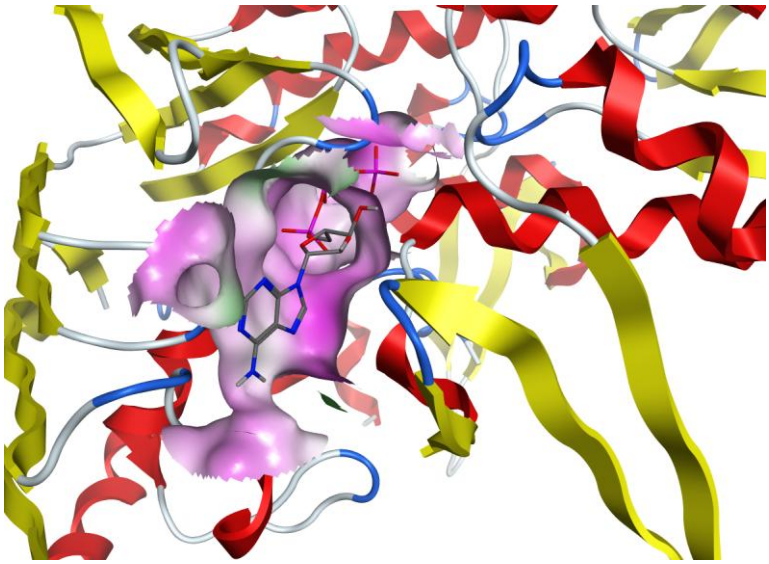


Figure 19: Exemple d'un arrimage moléculaire du ligand ATP dans le site de liaison de la protéine Rad50 (code pdb : 1f2u). Ici la surface moléculaire représentée délimite les résidus du site de liaison qui interagissent avec la molécule d'ATP.

Lors d'un profilage utilisant l'arrimage sur plusieurs cibles, on se retrouve dans des cas où certaines cibles vont être bien évalués et d'autres beaucoup moins. Pour contourner ce problème, on peut effectuer des transformations sur les scores bruts issus de l'arrimage afin de les ajuster et éviter le biais original.

Yang et al. (Yang *et al.* 2009) ont proposé une méthode utilisant une double transformation des scores appelée 2DIZ (*2-directional Z-transforamtion*). Des scores-*Z* (*Z*′) sont calculés à partir d'une normalisation des scores d'arrimages bruts de chaque association en les déterminant pour chaque ligand. Ensuite, un autre score-*Z* (*Z*′′) est calculé pour chaque interaction (en utilisant le score-*Z*′ calculé précédemment) induisant une normalisation pour chaque cible (Figure 20).

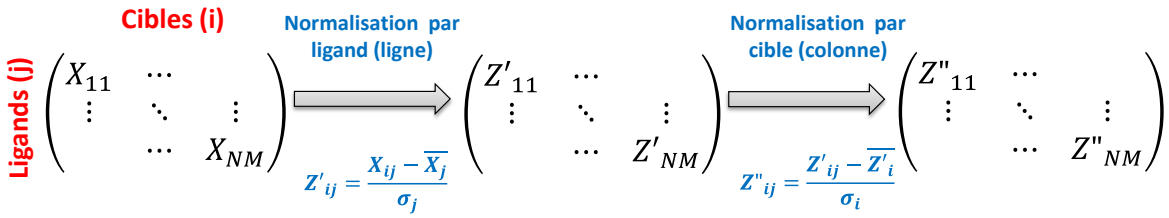


Figure 20 : Double transformation pour corriger et normaliser les scores d'arrimage brut. X_{ij} sont les scores d'arrimage brut. *Z*′ et *Z*′′ sont les scores-*Z* calculés pour chaque association. σ_j et σ_i correspondent aux déviations standards calculées pour les ligands j et les cibles i respectivement. N est le nombre total de ligands et M est le nombre total de cibles.

Il faut noter que cette transformation n'est applicable que pour une distribution de scores qui suivent une loi normale. Elle a été testée sur un ensemble de 79 cibles responsables d'effets secondaires connus et extraites de la PDB. Le profilage a été réalisé en utilisant 86 ligands connus de ces cibles et en se servant du programme DOCK. Les auteurs ont ainsi constaté un gain de 10% sur la sensibilité du profilage à retrouver les cibles connues par rapport à l'utilisation de scores bruts d'arrimage.

Li et al. (Li *et al.* 2011) ont profilé 4 621 médicaments sur 252 cibles extraites de la base DrugBank. Les auteurs utilisent des seuils de scores pour sélectionner les associations prédites comme correctes. Les scores obtenus lors du profilage sont triés du meilleur au plus mauvais. Le seuil de score choisi est utilisé pour écarter les mauvaises poses d'arrimage. Un pourcentage *P* des associations connues par rapport aux associations gardées (après avoir appliqué le seuil) est utilisé pour évaluer la méthode.

Une combinaison de ces seuils de scores avec des seuils appliqués sur le rang trouvé de la vraie cible (parmi toutes les cibles profilées) et le rang du ligand à profiler (parmi tous les ligands profilés sur la même cible) permettent de mieux écarter les mauvaises interactions et ont permis

d'obtenir un pourcentage *P* d'environ 40% mais en diminuant de 90% la taille de la liste à garder. Le seul inconvénient est sans doute le risque de passer à côté de plusieurs associations possibles, mais gardons à l'esprit que dans le cadre d'un profilage, le point le plus important n'est pas d'obtenir des listes exhaustives de cibles potentielles mais d'obtenir des listes de cibles avec le moins de faux positifs possibles.

L'arrimage moléculaire est donc une approche adaptée au profilage de petites molécules si toutefois on dispose d'une bonne stratégie pour identifier les bonnes associations. C'est également une méthode qui permet de trouver des molécules avec de nouveaux châssis moléculaires mais également de prédire leur mode d'interaction et ainsi de faciliter une optimisation ultérieure.

Mais l'arrimage moléculaire a aussi des inconvénients. Tout d'abord, il est nécessaire de disposer de plusieurs processeurs (une centaine voir plus) afin de pouvoir l'effectuer sur des milliers de cibles. Ensuite, les fonctions de scores sous-estiment ou surestiment certaines association entre molécules et cibles. Il est nécessaire d'appliquer une correction dans le but de s'affranchir de cette limite.

Une question se pose aussi sur le nombre de sites de liaisons à choisir à partir des structures cristallographiques disponibles pour chaque cible en vue de permettre l'identification d'une éventuelle association avec un ligand. Bien entendu ceci est étroitement lié avec la définition même d'un site de liaison. Si un site de liaison est défini en fonction du ligand qui lui est rattaché, il est plus judicieux d'inclure tous les sites liés aux ligands cristallisés avec cette cible. Mais cette remarque est surtout valable pour des cibles comme certaines protéines kinases qui peuvent adopter plusieurs conformations selon le type de son ligand. L'exemple le plus connu est celui de la protéine kinase *Abl* qui peut adopter une conformation DFG-in (conformation active) et DFG-out (conformation inactive) selon la nature de son inhibiteur, il faut donc prêter une attention particulière au choix du site (Zhou *et al.* 2010).

2.3.2. Profilage utilisant les pharmacophores 3D

Un pharmacophore est un ensemble de propriétés stériques et électroniques défini à partir d'une interaction entre deux entités moléculaires et nécessaire pour induire la réponse biologique souhaitée.

La représentation informatique d'un pharmacophore 3D (Figure 21) se résume à situer des motifs (sphères colorées sur la figure 21) sur le site de liaison selon la nature de l'interaction inter-moléculaire non covalente impliquée avec le ligand lié. Cette représentation dérive directement de la structure du complexe protéine-ligand. Mais on peut également la déterminer à partir d'une liste de ligands connus pour une cible. Les conformations de ses ligands sont superposées à l'aide d'un algorithme d'alignement en maximisant le nombre de propriétés pharmacophoriques communes qui sont ensuite identifiées et définies comme motifs pharmacophoriques.

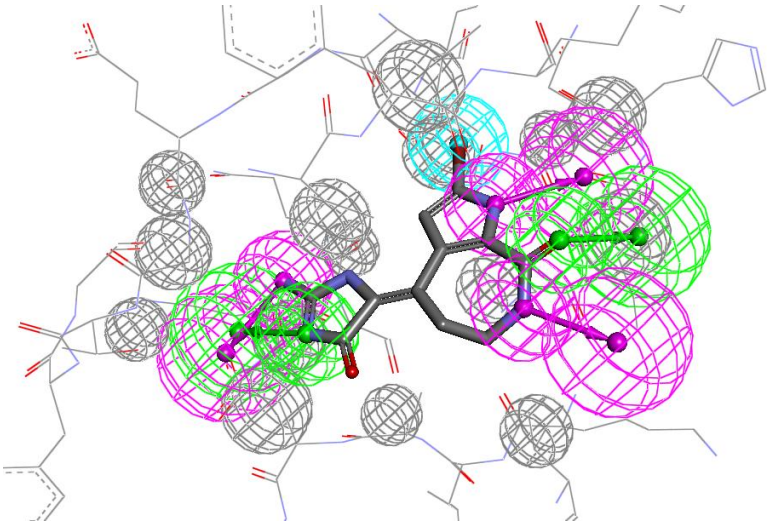


Figure 21 : Représentation d'un pharmacophore 3D dérivé du complexe cristallographique 1dm2. Le ligand est représenté avec des bâtons épais et les résidus du site de liaison avec des bâtons simples.
sphères grises: sphères d'exclusions ; sphères bleu: hydrophobes ; sphères magenta: donneurs de liaison H ; sphères verte: accepteurs de liaison H.

Les motifs pharmacophoriques les plus utilisés sont les interactions de type accepteur/donneur de liaisons hydrogène, interactions lipophiles et les interactions ioniques

(charge positive/négative). Il est aussi commun d'associer des sphères d'exclusion (sphères grises sur la figure 21) sur les atomes des résidus du site de liaison pour écarter tout contact stérique avec ces résidus lors d'un criblage. D'autres représentations peuvent délimiter le volume que doit occuper la conformation d'un ligand criblé afin de limiter la recherche sur un volume donné du site de liaison. Il est important de souligner que la représentation d'un pharmacophore 3D est très intuitive et facile à interpréter par les chimistes médicinaux d'où sa croissante utilisation dans le criblage virtuel au cours de cette décennie (Wolber *et al.* 2008).

Lors d'un criblage virtuel pharmacophorique 3D, une conformation d'un ligand est alignée pour maximiser le nombre de correspondance entre ses propriétés pharmacophoriques et les motifs complémentaires. Un score de correspondance (généralement appelé *fit*) est ainsi obtenu pour évaluer cette association. Une cible peut être représentée par plusieurs pharmacophores qui peuvent avoir un nombre de motifs différents. Le choix du nombre de motifs est généralement suivi d'une étude de sélectivité du modèle pharmacophorique généré. Des molécules quelconques sont criblées et une spécificité est calculée pour évaluer le pharmacophore. Les pharmacophores les plus spécifiques sont ceux qui sont retenus pour les campagnes de criblages.

Dans le cadre d'un profilage, des molécules sont criblées sur une série de pharmacophores qui représentent plusieurs cibles. D'un point de vue pratique, il est difficile de constituer une base de données de pharmacophores à des fins de profilage, car il est difficile d'automatiser la procédure de génération à grande échelle. Plusieurs problèmes se posent alors pour l'évaluation des modèles générés automatiquement dont la sélectivité des modèles.

Deux grandes collections répertorient des pharmacophores pour plusieurs cibles. PharmTargetDB (Liu *et al.* 2010) est une base de 7 000 pharmacophores de cibles extraites des bases DrugBank (Knox *et al.* 2011), BindingDB (Liu *et al.* 2007), PDBBind (Wang *et al.* 2005) et PDTD (Gao *et al.* 2008) et dont les structures cristallographiques protéine-ligand sont disponibles. Notons que la base inclut 2 241 pharmacophores issus de complexes avec des protéines humaines. Les pharmacophores ont été générés manuellement à l'aide du logiciel LigandScout (Wolber *et al.* 2005) et les auteurs ont mis à disposition un serveur web dédié au profilage (http://59.78.96.61/pharmmapper). La validation de ce protocole a été menée sur le tamoxifène, médicament utilisé pour le carcinome mammaire. 71% des 14 cibles connues ont été retrouvées parmi les 300 cibles les mieux scorées lors du profilage.

La deuxième base est celle commercialisée par la société Inte:Ligand (http://www.inteligand.com/pharmdb/) et qui contient plus de 2 500 pharmacophores pour 300 cibles thérapeutiques. Pharmdb est construite manuellement par des experts et les pharmacophores sont déduis à partir des structures cristallographiques de cibles des plus grandes classes thérapeutiques (anti-infectieux, cardiovasculaires, immunologiques, métaboliques, neurologiques…).

Steindl et al. (Steindl *et al.* 2006) ont profilé 100 composés antiviraux sur 50 pharmacophores de 5 protéines virales (HIV protéase, reverse transcriptase virus HIV, neuraminidase du virus Influenza, Protéine Rhinovirus coat et l'hépatite C RNA polymérase) issus de Pharmdb. Les 100 composés représentent la somme de 20 inhibiteurs connus pour chacune de ses protéines virales. Une identification du profil biologique définit par les 5 protéines virales est évalué. Les profils biologiques obtenus étaient bien prédits à 90%.
Les composés pour lesquels le profilage a échoué étaient des inhibiteurs de la *neuraminidase du virus Influenza*. Ces composés ont montré qu'ils se liaient à la *reverse transcriptase HIV* du fait de leurs propriétés pharmacophoriques communes. Les auteurs ont aussi remarqué que les pharmacophores de la *reverse transcriptase HIV* étaient peu sélectifs car cette protéine se lie à différents types de châssis moléculaires. La base Pharmdb a aussi servi dans une étude rétrospective pour l'identification de nouvelles cibles pour des métabolites secondaires de la plante *Ruta graveolens* (Rollinger *et al.* 2009).

L'utilisation des pharmacophores 3D à des fins de profilage présente plusieurs avantages. Cette technique est rapide et permet d'identifier des molécules avec des châssis moléculaires différents de ceux déjà connus pour une cible donnée.

Mais cette méthode a évidemment ses limites. La première réside dans l'échantillonnage conformationnel nécessaire pour le criblage d'une molécule. Si celle-ci est très flexible, l'échantillonnage conformationnel risque d'être incomplet. La deuxième limite se situe dans l'identification du seuil à choisir pour la séparation entre les molécules actives et les molécules inactives. Habituellement, ce seuil est choisi de façon à tenir compte de la correspondance de plus de la moitié des motifs du pharmacophore avec la conformation de la molécule ou en

sélectionnant des conformations de molécules qui correspondent à des motifs définis comme nécessaires pour toute association.

La troisième limite est l'évaluation de la sélectivité de chaque pharmacophore généré. Il est très important de ne sélectionner que les pharmacophores sélectifs pour des études de profilage au risque d'obtenir beaucoup de faux positifs. La sélectivité d'un modèle est généralement évaluée en criblant une chimiothèque référence, ce qui constitue un facteur limitant car demandant beaucoup de temps de calcul pour la génération d'une base de pharmacophores à des fins de profilage. Nous nous sommes intéressés à ce point dans le troisième chapitre et nous avons proposé une stratégie pour l'évaluation de la sélectivité à travers un modèle prédictif.

2.3.3. Profilage en utilisant les machines d'apprentissages

Les machines d'apprentissage ont longtemps été utilisées pour créer des modèles qui séparent les molécules actives des molécules inactives ou pour prédire une affinité mais par rapport à une seule cible en n'utilisant que les descripteurs de ses ligands (Geppert *et al.* 2010).

Désormais avec le développement des descripteurs de protéines, il est d'ores et déjà possible de combiner les descripteurs de ligands et ceux des protéines dans le but de créer des modèles chémogénomiques pour la prédiction simultanée de molécules sur plusieurs cibles (Figure 22).
Les modèles chémogénomiques issus des machines d'apprentissage, permettent la détection et la reconnaissance de sous-structures moléculaires responsable de l'activité pour chaque cible dans l'espace chémogénomique et les regroupent d'une manière informelle pour mieux les distinguer. L'apprentissage sur les données se fait d'une manière coopérative et les cibles partagent l'information de leurs ligands respectifs.

Les machines d'apprentissage les plus utilisées pour générer des modèles chémogénomiques sont les forêts d'arbres décisionnels (*random forest*), les séparateurs à vaste marge (SVM) et les régressions PLS. Une récente revue énumère les différentes machines d'apprentissage avec leurs exemples d'application en chémogénomique (van Westen *et al.* 2011).

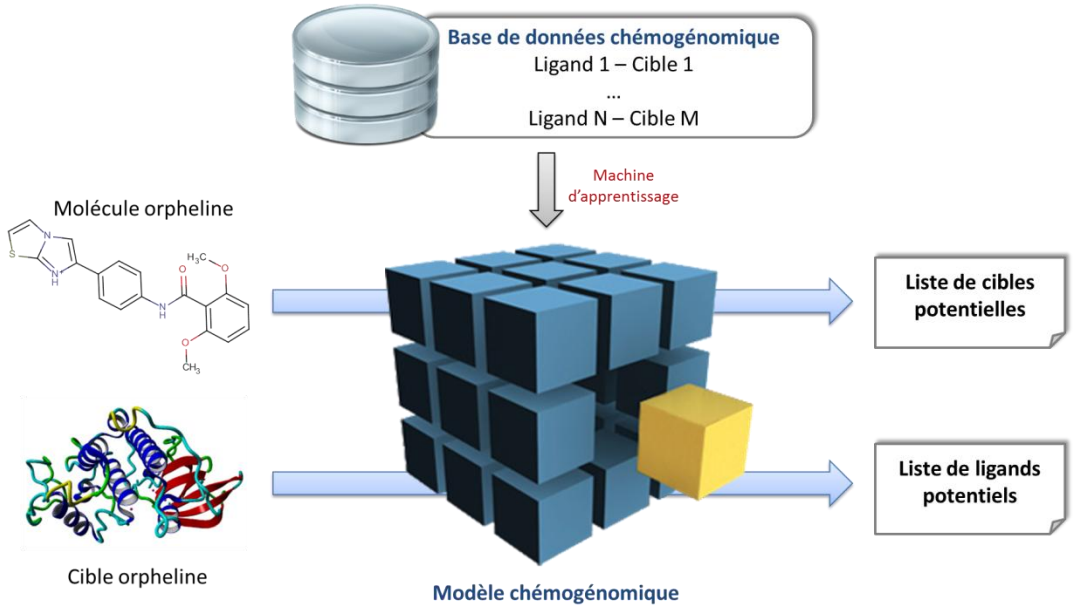


Figure 22 : Illustration du fonctionnement d'un modèle chémogénomique développé à partir des descripteurs de ligands et des protéines. La molécule et la cible orpheline doivent être incluses dans le domaine d'applicabilité du modèle pour pouvoir établir des prédictions.

Nous discuterons ici de deux études réalisées sur un grand ensemble de données et qui permettent le profilage sur un nombre important de cibles.

Strömbergsson et al. (Strömbergsson *et al.* 2010) ont construit un modèle de classification chémogénomique en utilisant les complexes protéines-ligands extraits de la base de donnée de bioactivité BindingDB. Un modèle contenant 7 087 complexes (2 721 ligands et 340 cibles) a été généré en combinant une série de descripteurs physicochimiques de ligands Dragon (http://www.talete.mi.it/products/dragon_description.htm) et de descripteurs physico-chimiques de protéines PROFEAT (Rao *et al.* 2011). Deux classes d'activités ont été définies, une classe de haute affinité (pKi $\geq$ 7) et une de basse affinité (pKi < 7). Deux arbres de décision ont été générés à partir de tous les descripteurs (ligand + protéine), le premier contenant 513 nœuds et le deuxième contenant 49 nœuds issus d'un élagage du premier arbre. Ces arbres de décision ont permis d'obtenir une précision de 82% pour le premier arbre et 75% pour le deuxième dans le cadre d'une validation sur un ensemble externe de données. Les auteurs ont utilisé deux procédures de validation croisée, la première consiste à construire un ensemble de test externe pour chaque protéine (extraction de tous les ligands associés à chaque protéine), et un ensemble

de test externe pour chaque ligand (extraction de toutes les associations où le ligand est impliquée). Le reste de l'ensemble est utilisé pour l'entrainement du modèle. Au cours de cette validation, les auteurs ont observé que le premier cas de validation (prédiction de protéines) était moins performant que le deuxième (prédiction de ligands). Cette tendance avait été observée dans une précédente étude de modèles de classifications chémogénomiques de RCPG (Weill *et al.* 2009). Ceci peut s'expliquer par le simple fait que le nombre de ligands inclus dans le modèle est en général beaucoup plus grand que le nombre de cibles, et par conséquent, le modèle généré aura plus de facilité à associer un nouveau ligand à une cible que l'inverse.
Ce modèle générée par Strömbergsson et al. permet de prédire une association à partir de simples descripteurs 2D de ligands et de protéines. Il n'est donc pas nécessaire de connaitre la structure tridimensionnelle du ligand, de la protéine ou du complexe ce qui élargit son champ d'application. Les principales cibles que l'on peut profiler à l'aide de ce modèle sont des récepteurs membranaires, des enzymes, des canaux ioniques et des récepteurs nucléaires, ce qui constitue les principales cibles de médicaments.

Wang et al. (Wang *et al.* 2011) ont utilisé 26 225 ligands et 626 cibles extraites de BindingDB (Liu *et al.* 2007). Les complexes sont représentés par des descripteurs physicochimiques 2D pour les ligands (Accelrys Discovery Studio) et une pseudo composition d'acides aminés qui tient compte de l'effet d'ordre de ces derniers dans la séquence. Cette composition appelée PseAAC (Du *et al.* 2012) se traduit par une suite de 20 coefficients décrivant chaque protéine.
Les descripteurs sont concaténés en un seul vecteur et un modèle SVM est généré à partir de l'ensemble des données. Un ensemble de validation externe a été construit pour deux bases : la base Accelrys MDDR et la base NCDS (*National Center for Drug Screening database*, Chine). Une précision de 82% a été observée sur le jeu de test externe. Les auteurs ont donc appliqué le modèle pour identifier des inhibiteurs de quatre protéines non présentes dans le jeu d'entrainement : un récepteur couplé aux protéines G GPR40, la déacétylase sirtuin-1 SIRT1, la protéine kinase p38 et la protéine kinase GSK-3β. Ces protéines sont impliquées dans plusieurs maladies dont la maladie d'Alzheimer et le diabète. La base des composés commerciaux de la société SPECS (http://www.specs.net) a été criblée en associant chaque descripteur des quatre

protéines à chaque molécule (191 407 au total), et le modèle a évalué la possibilité de leur association (Tableau 9).

Protéine	Nombre de molécules prédites actives par le modèle	Nombre de molécules actives confirmées par tests in vitro	Kd ou IC50 µM
GPR40	27	1	22
P38	47	2	4,72 / 4,76
SIRT1	55	5	5,7 / 16 / 27,6 / 28,6 / 50,2
GSK-3β	50	1	0,5

Tableau 9 : Nombre de molécules prédites et testées expérimentalement sur les 4 cibles

Neuf touches ont été confirmées expérimentalement sur les quatre cibles possédant des affinités micro-molaires.

L'avantage de ces modèles chémogénomiques réside dans le fait qu'un profilage de milliers de molécules en quelques heures est maintenant réalisable. Grâce à ses méthodes, on peut trouver des ligands pour des cibles orphelines dont il suffit simplement de connaitre la séquence d'acides aminées. Ces approches ont connu une progression d'utilisation et des cas d'application avec succès ont été publiés notamment pour identifier de nouveaux ligands non-peptidiques du récepteur de l'Oxytocine (Weill *et al.* 2011).

Le principal inconvénient des deux approches citées précédemment est le fait de l'utilisation de la séquence entière d'acides aminés de la protéine. L'association prédite est donc incomplète. Certaines protéines ont plusieurs sites de liaison, il n'est donc pas possible de connaitre quel site est impliqué dans l'association prédite. Il serait donc plus intéressant d'utiliser cette approche en incluant des descripteurs 3D de sites de liaisons. Ce point fera l'objet d'une étude qui sera détaillée dans le prochain chapitre et la performance de l'utilisation de différents descripteurs de cibles sera évaluée.

Il faut souligner que le domaine d'applicabilité pour les modèles chémogénomique doit être défini. Il est malheureusement très peu abordé dans ces études. Le fait d'inclure plusieurs cibles avec leurs ligands dans le modèle rend la formulation d'une définition générale un peu difficile. Ce point sera aussi abordé dans le prochain chapitre et on proposera une stratégie pour définir un domaine d'applicabilité pour les modèles chémogénomiques.

2.4. Profilage utilisant les approches expérimentales

Comme détaillé précédemment, la prédiction des associations entre des molécules et des cibles sont généralement identifiées à partir de la structure chimique des molécules et/ou celles des cibles.

Cependant, il est possible de les identifier en exploitant les caractéristiques pharmacologiques qui découlent de ces associations. En effet, les médicaments ont un effet thérapeutique souvent accompagné de quelques effets indésirables. La cible principale du médicament est la cible responsable de l'effet thérapeutique et du changement phénotypique désiré, mais les effets secondaires sont dus à la liaison de ce médicament à d'autres cibles.

Il a été démontré auparavant que des médicaments de profil biologique in vitro similaires ont tendance à avoir des effets secondaires similaires (Fliri *et al.* 2005). On peut alors essayer de relier les médicaments entre eux à l'aide de la similarité de leurs effets secondaires dans le but d'identifier des cibles secondaires communes (Figure 23).

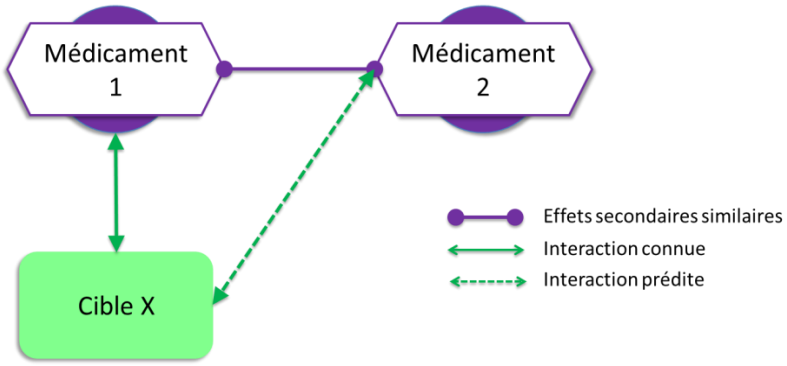


Figure 23 : Principe de prédiction d'une nouvelle interaction entre un médicament et une cible.

Campillos et al. (Campillos *et al.* 2008) ont identifié des nouvelles interactions entre des médicaments déjà disponibles sur le marché et des cibles à partir des effets secondaires répertoriées pour ces derniers. Les auteurs ont utilisé 502 médicaments avec leurs annotations connues de cibles humaines pour montrer que les ligands qui partagent les mêmes cibles, sont aussi responsables d'effets secondaires similaires (Figure 24).

Bien entendu, certains effets secondaires sont très fréquents quel que soit le médicament (nausées, fièvres, vomissements…) alors que d'autres le sont beaucoup moins (fibrose rénale, anémie). Il a fallu à l'évidence attribuer des poids aux fréquences d'occurrences des effets secondaires mais également les relier entre eux, car certains effets sont des conséquences directes d'autres effets (la nausée peut engendrer des vomissements).

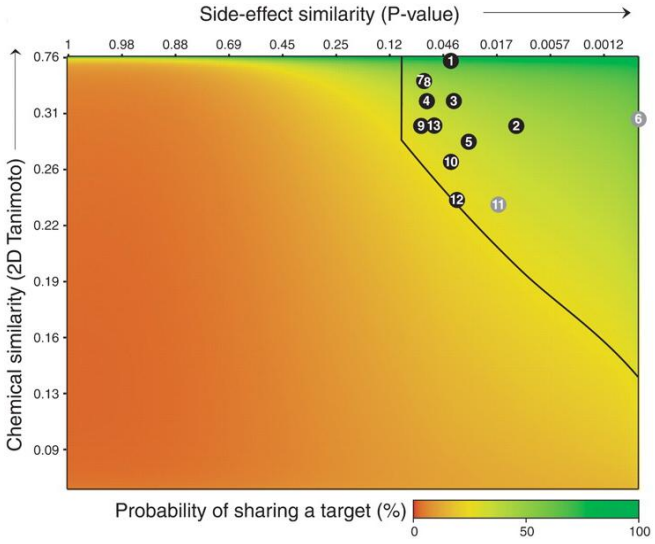


Figure 24 : Les ligands qui partagent les mêmes cibles (couleur verte) présente une forte similarité chimique mais aussi une forte similarité des effets secondaires.
Figure extraite de l'article (Campillos *et al.* 2008).

Cependant, en associant l'information de la structure chimique des molécules à leurs effets secondaires, les auteurs ont remarqué que la spécificité des prédictions d'association médicament-cible augmentait de 10% en moyenne par rapport à l'utilisation de la similarité des effets secondaires. Dès lors, ils ont fait le choix d'associer les deux représentations pour essayer de trouver de nouvelles interactions pour 746 médicaments. Les auteurs ont analysé les interactions prédites et identifiées à partir des paires de médicaments d'effets thérapeutiques différents. Les résultats obtenus étaient très satisfaisants, et ils ont pu obtenir des affinités à des concentrations micro-molaires pour 13 interactions à l'aide de tests in vitro et 11 interactions à l'aide de tests in-vivo.

Cette approche est donc très intéressante notamment dans le cadre du repositionnement des médicaments. Un médicament peut interagir avec une cible liée à un effet thérapeutique autre que l'effet initial ciblée par ce dernier. Ceci permet d'éviter des phases de tests cliniques

précoces déjà effectués et d'obtenir un médicament commercialisable avec un gain de temps considérable.

Cet avantage peut être aussi un inconvénient, car cette méthode n'est applicable qu'aux médicaments déjà commercialisés ou aux molécules ayant des effets secondaires connus. Ceci limite grandement son domaine d'application.

2.5. Références

AbdulHameed, M. D. M., S. Chaudhury, N. Singh, H. Sun, A. Wallqvist and G. J. Tawa (2012). "Exploring Polypharmacology Using a ROCS-Based Target Fishing Approach." J Chem Inf Model **52**(2): 492-505.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology **215**(3): 403-410.

Arakawa, K., N. Kono, Y. Yamada, H. Mori and M. Tomita (2005). "KEGG-based pathway visualization tool for complex omics data." In Silico Biol **5**(4): 419-423.

Bender, A. and R. C. Glen (2004). "Molecular similarity: a key technique in molecular informatics." Org Biomol Chem **2**(22): 3204-3218.

Bender, A., J. L. Jenkins, M. Glick, Z. Deng, J. H. Nettles and J. W. Davies (2006). ""Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept?" J Chem Inf Model **46**(6): 2445-2456.

Benson, M. L., R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin and H. A. Carlson (2008). "Binding MOAD, a high-quality protein-ligand database." Nucleic Acids Research **36**(Database issue): D674-678.

Berman, H., K. Henrick and H. Nakamura (2003). "Announcing the worldwide Protein Data Bank." Nat Struct Biol **10**(12): 980.

Block, P., C. A. Sotriffer, I. Dramburg and G. Klebe (2006). "AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB." Nucleic Acids Research **34**(Database issue): D522-526.

Campillos, M., M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork (2008). "Drug target identification using side-effect similarity." Science **321**(5886): 263-266.

Cramer, R. D., D. E. Patterson and J. D. Bunce (1988). "Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins." J Am Chem Soc **110**(18): 5959-5967.

Defranchi, E., C. Schalon, M. Messa, F. Onofri, F. Benfenati and D. Rognan (2010). "Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements." PLoS One **5**(8): e12214.

Du, P., X. Wang, C. Xu and Y. Gao (2012). "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions." Anal Biochem **425**(2): 117-119.

Fliri, A. F., W. T. Loging, P. F. Thadeio and R. A. Volkmann (2005). "Analysis of drug-induced effect patterns to link structure and side effects of medicines." Nature Chemical Biology **1**(7): 389-397.

Gao, Z., H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, W. Zhu, K. Chen, X. Wang and H. Jiang (2008). "PDTD: a web-accessible protein database for drug target identification." Bmc Bioinformatics **9**: 104.

Garcia-Serna, R., O. Ursu, T. I. Oprea and J. Mestres (2010). "iPHACE: integrative navigation in pharmacological space." Bioinformatics **26**(7): 985-986.

Gaulton, A., L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Research **40**(D1): D1100-D1107.

Geppert, H., M. Vogt and J. Bajorath (2010). "Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation." J Chem Inf Model **50**(2): 205-216.

Greene, L. H., T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton and C. A. Orengo (2007). "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution." Nucleic Acids Research **35**(Database issue): D291-297.

Gregori-Puigjane, E. and J. Mestres (2006). "SHED: Shannon entropy descriptors from topological feature distributions." J Chem Inf Model **46**(4): 1615-1622.

http://accelrys.com/products/databases/bioactivity/mddr.html. (consulté en Juin 2012). "Accelrys MDDR."

http://pdsp.med.unc.edu/pdsp.php. (consulté en Juin 2012). "PDSP Ki database."

http://www.daylight.com/dayhtml/doc/theory/theory.finger.html. (consulté en Juin 2012). "Daylight Fingerprints."

Huang, N., B. K. Shoichet and J. J. Irwin (2006). "Benchmarking sets for molecular docking." Journal of Medicinal Chemistry **49**(23): 6789-6801.

Jones, G., P. Willett, R. C. Glen, A. R. Leach and R. Taylor (1997). "Development and validation of a genetic algorithm for flexible docking." Journal of Molecular Biology **267**(3): 727-748.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi and M. Tanabe (2012). "KEGG for integration and interpretation of large-scale molecular data sets." Nucleic Acids Research **40**(D1): D109-D114.

Karaman, M. W., S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka and P. P. Zarrinkar (2008). "A quantitative analysis of kinase inhibitor selectivity." Nature Biotechnology **26**(1): 127-132.

Keiser, M. J., B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet (2007). "Relating protein pharmacology by ligand chemistry." Nature Biotechnology **25**(2): 197-206.

Keiser, M. J., V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth (2009). "Predicting new molecular targets for known drugs." Nature **462**(7270): 175-U148.

Kinnings, S. L. and R. M. Jackson (2011). "ReverseScreen3D: a structure-based ligand matching method to identify protein targets." J Chem Inf Model **51**(3): 624-634.

Klabunde, T. (2007). "Chemogenomic approaches to drug discovery: similar receptors bind similar ligands." Br J Pharmacol **152**(1): 5-7.

Knox, C., V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart (2011). "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs." Nucleic Acids Research **39**: D1035-D1041.

Kuhn, M., M. Campillos, P. Gonzalez, L. J. Jensen and P. Bork (2008). "Large-scale prediction of drug-target relationships." Febs Letters **582**(8): 1283-1290.

Laggner, C., D. Kokel, V. Setola, A. Tolia, H. Lin, J. J. Irwin, M. J. Keiser, C. Y. J. Cheung, D. L. Minor, B. L. Roth, R. T. Peterson and B. K. Shoichet (2012). "Chemical informatics and target identification in a zebrafish phenotypic screen." Nature Chemical Biology **8**(2): 144-146.

Li, Q., T. Cheng, Y. Wang and S. H. Bryant (2010). "PubChem as a public resource for drug discovery." Drug Discov Today **15**(23-24): 1052-1057.

Li, Y. Y., J. H. An and S. J. M. Jones (2011). "A Computational Approach to Finding Novel Targets for Existing Drugs." Plos Computational Biology **7**(9).

Liu, T., Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic Acids Research **35**(Database issue): D198-201.

Liu, X., S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, S. Zheng, Z. Li, H. Li and H. Jiang (2010). "PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach." Nucleic Acids Research **38**(Web Server issue): W609-614.

Martin, E., P. Mukherjee, D. Sullivan and J. Jansen (2011). "Profile-QSAR: A Novel meta-QSAR Method that Combines Activities across the Kinase Family To Accurately Predict Affinity, Selectivity, and Cellular Activity." J Chem Inf Model **51**(8): 1942-1956.

Martin, E., P. Mukherjee, D. Sullivan and J. Jansen (2011). "Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity." J Chem Inf Model **51**(8): 1942-1956.

Martin, Y. C., J. L. Kofron and L. M. Traphagen (2002). "Do structurally similar molecules have similar biological activity?" Journal of Medicinal Chemistry **45**(19): 4350-4358.

Meslamani, J., D. Rognan and E. Kellenberger (2011). "sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins." Bioinformatics **27**(9): 1324-1326.

Milletti, F. and A. Vulpetti (2010). "Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe." J Chem Inf Model **50**(8): 1418-1431.

Mitchell, J. B. O. (2001). "The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands." J Chem Inf Comput Sci **41**(6): 1617-1622.

MolecularDiscovery Flap; Molecular Discovery Ltd: Middlesex, U.K.

Nidhi, M. Glick, J. W. Davies and J. L. Jenkins (2006). "Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases." J Chem Inf Model **46**(3): 1124-1133.

Olah, M., R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mractc and T. I. Oprea (2008). WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. Chemical Biology, Wiley-VCH Verlag GmbH**:** 760-786.

OpenEye ROCS v.3.1.2; OpenEye Scientific Software: Santa Fe, NM 87507.

Paolini, G. V., R. H. B. Shapland, W. P. van Hoorn, J. S. Mason and A. L. Hopkins (2006). "Global mapping of pharmacological space." Nature Biotechnology **24**(7): 805-815.

Perez-Nueno, V. I. and D. W. Ritchie (2011). "Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening." J Chem Inf Model **51**(6): 1233-1248.

Rao, H. B., F. Zhu, G. B. Yang, Z. R. Li and Y. Z. Chen (2011). "Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence." Nucleic Acids Research **39**(Web Server issue): W385-390.

Rogers, D. and M. Hahn (2010). "Extended-connectivity fingerprints." J Chem Inf Model **50**(5): 742-754.

Rognan, D. (2007). "Chemogenomic approaches to rational drug design." Br J Pharmacol **152**(1): 38-52.

Rognan, D. (2010). "Structure-Based Approaches to Target Fishing and Ligand Profiling." Molecular Informatics **29**(3): 176-187.

Rognan, D. (2011). Docking Methods for Virtual Screening: Principles and Recent Advances. Virtual Screening, Wiley-VCH Verlag GmbH & Co. KGaA**:** 153-176.

Rollinger, J. M., D. Schuster, B. Danzl, S. Schwaiger, P. Markt, M. Schmidtke, J. Gertsch, S. Raduner, G. Wolber, T. Langer and H. Stuppner (2009). "In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens." Planta Med **75**(3): 195-204.

Roth, B. L., D. J. Sheffler and W. K. Kroeze (2004). "Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia." Nat Rev Drug Discov **3**(4): 353-359.

Schalon, C., J. S. Surgand, E. Kellenberger and D. Rognan (2008). "A simple and fuzzy method to align and compare druggable ligand-binding sites." Proteins **71**(4): 1755-1778.

Schneider, G. (2010). "Virtual screening: an endless staircase?" Nat Rev Drug Discov **9**(4): 273-276.

Seiler, K. P., G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons (2008). "ChemBank: a small-molecule screening and cheminformatics resource database." Nucleic Acids Research **36**(Database issue): D351-359.

Sotriffer C., M. R., Kubinyi H., Folkers G. (2011). Virtual Screening: Principles, Challenges, and Practical Guidelines, Volume 48, WILEY-VCH.

Steindl, T. M., D. Schuster, G. Wolber, C. Laggner and T. Langer (2006). "High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening." J Comput Aided Mol Des **20**(12): 703-715.

Strömbergsson, H., M. Lapins, G. J. Kleywegt and J. E. S. Wikberg (2010). "Towards Proteome–Wide Interaction Models Using the Proteochemometrics Approach." Molecular Informatics **29**(6-7): 499-508.

Todeschini R., C. V. (2000). Handbook of Molecular Descriptors, Volume 11, WILEY-VCH.

van Westen, G. J. P., J. K. Wegner, A. P. IJzerman, H. W. T. van Vlijmen and A. Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets." Medchemcomm **2**(1): 16-30.

Varnek, A. and I. I. Baskin (2011). "Chemoinformatics as a Theoretical Chemistry Discipline." Molecular Informatics **30**(1): 20-32.

Vidal, D., R. Garcia-Serna and J. Mestres (2011). "Ligand-based approaches to in silico pharmacology." Methods Mol Biol **672**: 489-502.

Vidal, D. and J. Mestres (2010). "In Silico Receptorome Screening of Antipsychotic Drugs." Molecular Informatics **29**(6-7): 543-551.

Wang, F., D. Liu, H. Wang, C. Luo, M. Zheng, H. Liu, W. Zhu, X. Luo, J. Zhang and H. Jiang (2011). "Computational screening for active compounds targeting protein sequences: methodology and experimental validation." J Chem Inf Model **51**(11): 2821-2828.

Wang, R., X. Fang, Y. Lu, C. Y. Yang and S. Wang (2005). "The PDBbind database: methodologies and updates." Journal of Medicinal Chemistry **48**(12): 4111-4119.

Weill, N. and D. Rognan (2009). "Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands." J Chem Inf Model **49**(4): 1049-1062.

Weill, N. and D. Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites." J Chem Inf Model **50**(1): 123-135.

Weill, N., C. Valencia, S. Gioria, P. Villa, M. Hibert and D. Rognan (2011). "Identification of Nonpeptide Oxytocin Receptor Ligands by Receptor-Ligand Fingerprint Similarity Search." Molecular Informatics **30**(6-7): 521-526.

Wendt, B., U. Uhrig and F. Bos (2011). "Capturing Structure-Activity Relationships from Chemogenomic Spaces." J Chem Inf Model **51**(4): 843-851.

Willett, P., J. M. Barnard and G. M. Downs (1998). "Chemical similarity searching." J Chem Inf Comput Sci **38**(6): 983-996.

Wolber, G. and T. Langer (2005). "LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters." J Chem Inf Model **45**(1): 160-169.

Wolber, G., T. Seidel, F. Bendix and T. Langer (2008). "Molecule-pharmacophore superpositioning and pattern matching in computational drug design." Drug Discov Today **13**(1-2): 23-29.

Xie, L. and P. E. Bourne (2009). "A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery." Bioinformatics **25**(12): i305-312.

Yang, L., H. Luo, J. Chen, Q. Xing and L. He (2009). "SePreSA: a server for the prediction of populations susceptible to serious adverse drug reactions implementing the methodology of a chemical-protein interactome." Nucleic Acids Research **37**(Web Server issue): W406-412.

Yera, E. R., A. E. Cleves and A. N. Jain (2011). "Chemical structural novelty: on-targets and off-targets." Journal of Medicinal Chemistry **54**(19): 6771-6785.

Zhou, T., L. Commodore, W. S. Huang, Y. Wang, T. K. Sawyer, W. C. Shakespeare, T. Clackson, X. Zhu and D. C. Dalgarno (2010). "Structural analysis of DFG-in and DFG-out dual Src-Abl inhibitors sharing a common vinyl purine template." Chem Biol Drug Des **75**(1): 18-28.

Chapitre 2 :

sc-PDB : Base de données cristallographiques pour l'identification et la distinction des différents sites *druggables* dans les protéines

Ce chapitre a fait l'objet d'une publication :

sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins

Jamel Meslamani, Didier Rognan and Esther Kellenberger

Bioinformatics, **2011**, 27, 1324–1326

1. Contexte

La base de données Protein DataBank (PDB) (Berman *et al.* 2007), répertorie environ 83 000 structures cristallographiques dont 59 900 contiennent un ligand co-cristallisé. La base PDB constitue ainsi une source considérable de structures bioactives pour les chercheurs. Ses entrées sont enregistrées dans des fichiers pdb qui contiennent les coordonnées des atomes déterminées par cristallographie. Seulement, les fichiers pdb ne sont pas adaptés aux méthodes de criblages virtuels basées sur la structure car ils sont dédiés au stockage et à la description de protéines mais pas à une description moléculaire détaillée. C'est la raison pour laquelle en 2004, le laboratoire a essayé de mettre en place une base de données de sites de liaisons extraits de la base PDB et qui contient des fichiers adaptés aux méthodes de criblage virtuel. La base de données des sites sc-PDB (Kellenberger *et al.* 2006) a donc vu le jour.

La version 2011 actuelle regroupe 9 877 entrées incluant 3 034 protéines différentes et 5 339 ligands différents. La base fournit une annotation fonctionnelle des protéines et une description chimique des ligands ainsi que les interactions au sein du complexe. Les sites de liaisons sont définis par les acides aminés qui sont distants de moins de 6.5 Å du centre de masse défini par les atomes lourds du ligand. Des fichiers de type mol2 sont disponibles pour chaque ligand, site de liaison et protéine de chaque complexe constituant ainsi une base de données idéale pour du criblage virtuel et surtout pour les méthodes basées sur la structure.

Il faut noter que deux entrées sc-PDB d'une même protéine ne partagent pas forcément le même site de liaison. Les protéines kinases par exemple ont un site catalytique et un site allostérique.

Nous savons qu'il est intéressant d'identifier le site de liaison impliqué lors d'un profilage sur plusieurs cibles. C'est pourquoi nous avons essayé de répertorier les différents sites pour toutes les protéines d'une manière automatique. Dès lors, l'utilisateur pourra choisir un site de liaison spécifique à la fois pour effectuer un criblage virtuel ou un profilage biologique. A l'aide de cette procédure, il est désormais possible de ne sélectionner qu'un représentant de chaque site lors de criblage afin de réduire les temps de calcul.

2. Introduction

The Protein Data Bank (PDB) is the main public resource of biologically active 3D structures available to study the interactions that govern ligand binding to protein (Berman *et al.* 2007).To assist structure-based approaches in drug design, we have parsed the PDB to identify binding sites suitable for the docking of a *druglike* ligand and so have created a database named sc-PDB. The protein selection is based on the molecular weight, buried surface area and chemical structure of ligands as well as the volume of corresponding cavities (Kellenberger *et al.* 2006; Kellenberger *et al.* 2008). Since its creation in 2004, the database is updated annually, with regular improvements. Notably, the curation of ligand chemical structures (2005), the optimization of ligand-bound coordinates (2005) and the systematic description of the ligand binding mode (2006) give significant added values to the structural information contained in the database. The sc-PDB is annotated at a functional level and a sc-PDB target name is assigned to each entry. However, two sc-PDB entries with the same sc-PDB *target name* do not necessarily describe an identical binding site. For example, there are 37 copies of the tyrosine-kinase scr in the sc-PDB (Fig. 1A); in 24 entries, the binding site is the ATP binding site of the kinase domain (Site 1), while in other entries, the binding site is located in the SH2 domain and accommodates ligands of variable size (Sites 2 and 3).

The present application aims at the distinction of the sc-PDB binding sites for a particular protein. A hierarchical classification was established based on the geometrical and physico-chemical diversity of binding sites. All the binding sites found similar for a given protein were structurally aligned to yield a new set of coordinates for the ligand, site and protein files.

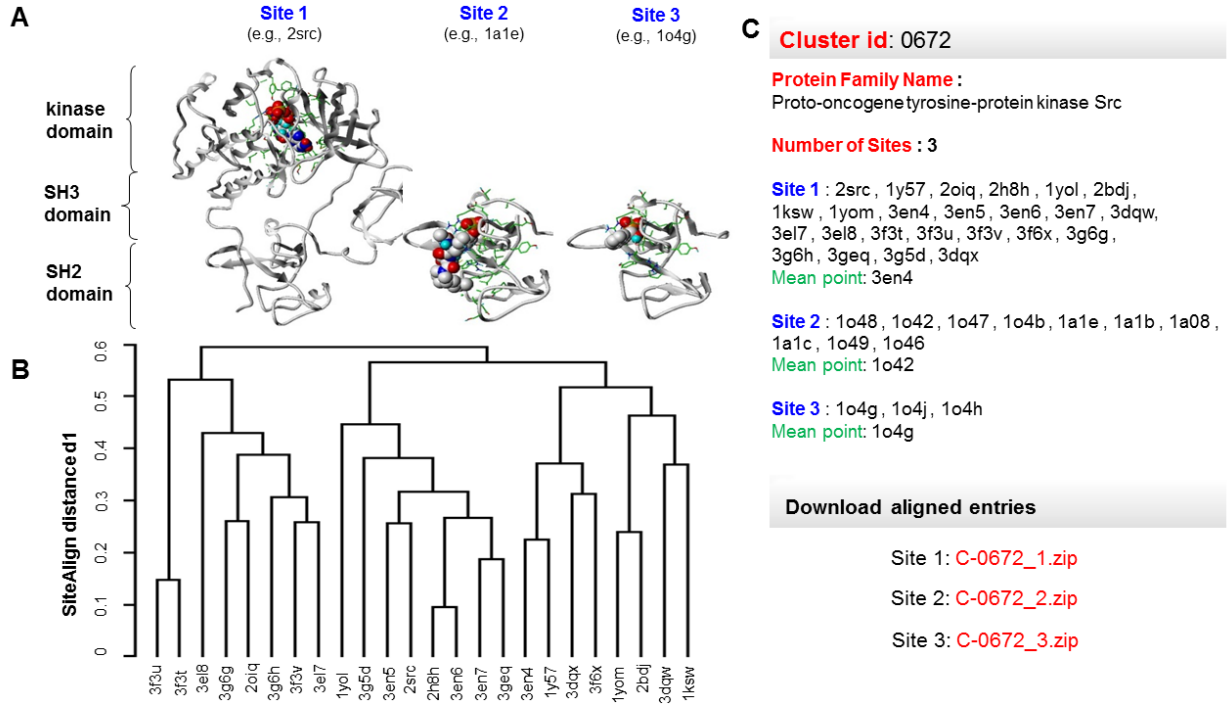


Figure 1: The proto-oncogene tyrosine kinase scr family. (**A**) 3D representation of the three binding sites. The protein chains are represented using ribbons, the bound ligands using balls and the binding sites residues using capped sticks. (**B**) Hierarchical classification of sc-PDB entries corresponding to Site 1. The dendrogram is available in sc-PDB web pages. (**C**) Web report of the classification.

3. Classification of binding sites by local structural similarity

The sc-PDB data are organized into a hierarchical classification scheme. The first level of the classification is the protein itself, as defined by the sc-PDB target name, which combines biological information retrieved from UniprotKB (UniProt 2010) and PDB archives. Typically, the Uniprot recommended name was considered if the PDB file includes the appropriate cross reference (a word matching check validates the consistency between PDB and Uniprot names). In the case of polyproteins (e.g. HIV gag-pol) or multifunctional proteins, the sc-PDB target name was chosen according to the domain function (e.g. EC number). Further simplifications of sc-PDB names remove tags for cellular locations and maturation states. Lastly, in the absence of Uniprot reference in the PDB entry, the sc-PDB name was directly extracted from the PDB file and manually curated for a better uniformity within the database. The current 8 166 sc-PDB (v.2010) entries represent 1 168 protein families and 1 470 singletons.

The second level of the classification distinguishes binding sites which significantly differs in size, shape or location in the protein. The third level reports homogeneous classes of structurally similar binding sites. The second and third levels of classification were obtained by clustering all members of a protein family (first-level classification) according to local structural similarity between sites. The all-against all comparison of sites was performed using the 3D alignment program SiteAlign (Schalon *et al.* 2008). SiteAlign searches for the best superposition of a target site onto a query site (the largest one) by optimizing a global similarity measure which estimates the agreement between the topological and physico-chemical attributes of site-specific 1D fingerprints. The algorithm was successfully applied to identify a novel off-target for some protein kinase inhibitors (Defranchi *et al.* 2010). Extensive validation of the program enabled the definition of an alignment score threshold value (SA_{score}>20) for distinguishing similar from dissimilar binding sites (Schalon *et al.* 2008). The SA_{score} quantifies all the binding sites discrepancies which result from changes in the sequence (e.g. between orthologous proteins, between wild-type and mutant or chimeric proteins), in the structure (different side chain positioning, motion in the backbone) and in the number of residues in the binding site (which is directly related to the size and the position of the bound ligand). In the current work, the SA_{scores} of the all-against-all comparison were stored in a distance matrix, which was converted into a Boolean matrix [A(i, j)=1 if $SA_{score} \geq 20$, else A(i, j)=0]. The Boolean matrix can be viewed as the adjacent matrix of an undirected graph, where each node represents a site, and an edge is defined between two nodes if the corresponding sites are similar. The Boolean matrix allowed the identification of all graph components (i.e. the maximal connected subgraphs), thereby grouping the sites of a protein family into one or more clusters. Lastly, a hierarchical clustering within clusters of similar sites was generated using the complete linkage method and distance equal to $[1-\ln (SA_{score})/\ln(SA_{scoremax})]$.

All sc-PDB entries belonging to the same cluster were structurally aligned to the mean point entry, whose average square distance to all other sites is the smallest in the cluster. The rotation and translation matrix applied to ligand, site and protein coordinates of the target entry were obtained from the global structural alignment of the target and reference proteins using the combinatorial extension (CE) program (Shindyalov *et al.* 1998). If the proteins contain several peptidic chains, all-against-all chain comparisons were performed, and the transformation which

minimizes the distance between the binding site centres was chosen. CE was here preferred over SiteAlign because it finds better alignments for a closely related structural ensemble. In contrast to SiteAlign, CE does not take into account changes in the amino acid type or changes in the rotameric state of residues. CE superimposition of protein structures thus allows the user to directly perceive the contribution of protein flexibility to the SA_{score}. In the example shown in Figure 1, the 24 copies of the kinase domain (here labeled as Site 1) are grouped in three main branches, which are not directly indicative of the binding site flexibility, the ligand chemotype or the ligand binding mode.

The absence of correlation between SA_{score} and deviation in the backbone coordinates of aligned sites rules out the interpretation of the dendrogram in terms of protein flexibility (the mean RMSD computed for the alpha carbon atoms of the residues defining the smallest site is 0.5±0.2Å). Annotation of the sc-PDB files rather suggests that the dendrogram reflects variations in sequence length (the number of residues in site ranges from 34 to 54) and nature (avian and human proteins, wild-type and mutant proteins).

4. A freely available source of 3D-aligned structures of ligand-bound protein binding sites

The clustering achieved on the 1 168 protein families of the 2010 sc-PDB release produced 783 singletons, 307 classes with two distinct sites, 60 classes with three distinct sites, 10 classes with four distinct sites, 4 classes with five distinct sites, 3 classes with six distinct sites and 1 class with eight sites. The sc-PDB classification is freely available via the database webserver (http://bioinfo-pharma.u-strasbg.fr/scPDB), by selecting the 'Clusters of Binding Site' page.

The query form allows search by target name, by number of distinct sites within a protein class or by PDB ID. The result page returns all classes matching the request, provides the user with the detailed content of each class (Fig. 1B and C) and enables the download of MOL2 files of structurally aligned entries. Alternatively, all sc-PDB binding sites similar to a particular binding site entry may be retrieved and ranked by decreasing similarity to the query.

The sc-PDB classification should foster drug design applications for two main reasons:

(i) it avoids comparing ligands that do not share the same binding site

(ii) it gives clues to evaluate the influence of ligand binding on binding site diversity for applications in structure-based methods (e.g. docking, site detection and comparison).

5. Statistiques révisées

La version actuelle de la sc-PDB v.2011 inclut 9 877 entrées avec 1 912 classes de sites de liaisons différents occupés par plus d'une entrée et 1 689 singletons. Lors de la création de classes de sites de liaison pour cette dernière version, nous avons utilisé le programme Shaper (Desaphy *et al.* 2012) pour comparer les sites. Shaper est un programme de comparaison de site de liaisons qui se base sur une comparaison de points de cavités annotés par des propriétés pharmacophoriques. La procédure de classification des sites reste inchangée et identique à celle utilisée avec le programme SiteAlign. Nous avons changé de programme d'alignement de site car Shaper était beaucoup plus rapide.

Nous remarquons à partir de la figure 2 que 70% des entrées sc-PDB n'ont qu'un seul site de liaison, 25% en possèdent deux et seulement 4% et 1% en possèdent à la fois 3 et 4 sites.

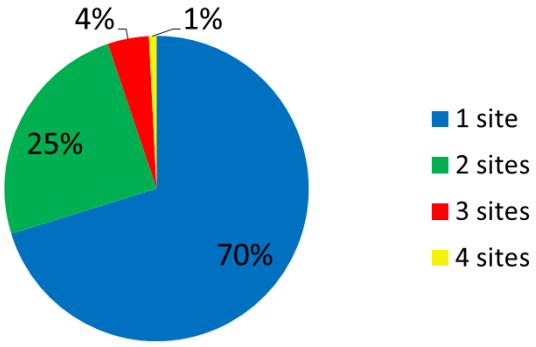


Figure 2 : Distribution du nombre de sites identifiés pour toutes les protéines de la sc-PDB 2011.

D'après la distribution du nombre de classes protéiques KEGG (Tanabe *et al.* 2012) pour les protéines qui possèdent un, deux, trois et quatre sites de liaisons différents (Figure 3), on s'aperçoit que ce sont les enzymes qui disposent principalement du plus grand nombre de site de

liaisons. Evidemment, ces statistiques sont un peu biaisées car les enzymes sont surreprésentés dans la base avec environ 65% des entrées.

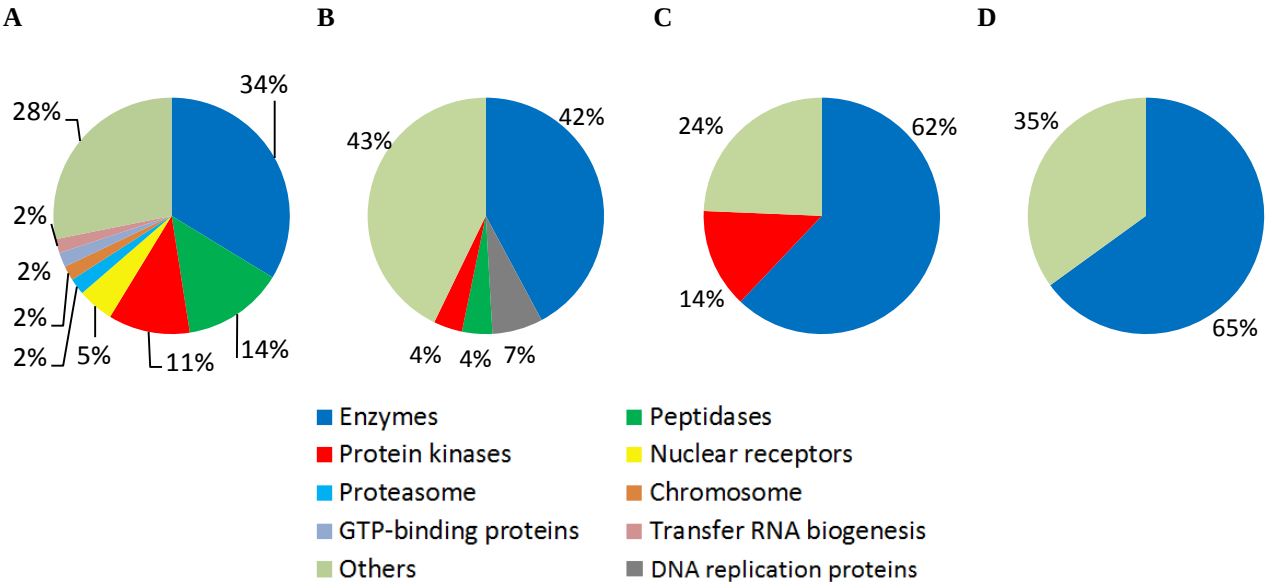


Figure 3: Pourcentages des classes protéiques KEGG pour **(A)** les proteines possédant 1 site de liaison. **(B)** les protéines possédant 2 sites de liaisons. **(C)** les protéines possédant 3 sites de liaisons. **(D)** les protéines possédant 4 sites de liaisons.

6. Applications et perspectives

Les sites de la sc-PDB étant groupés à l'aide de leur similarité au sein de chaque protéine, un représentant de chaque classe de site de liaison peut être ainsi utilisé pour un criblage virtuel ou du profilage à l'aide de sites de liaisons. Notre procédure permettra d'analyser les interactions avec les ligands pour chaque classe de site de liaison au sein de chaque protéine, afin de comprendre et distinguer celles qui disposent de plusieurs modes d'interactions au sein du même site.

La base sc-PDB est de plus en plus utilisée par la communauté scientifique. Le nombre d'utilisateurs est en constante augmentation avec plus de 53% de nouvelles visites cette année (Figure 4, source : google analytics).

La base a été appliquée à différentes thématique de recherche à savoir :

(i) l'arrimage moléculaire en validant un algorithme d'arrimage GlamDock (Tietze *et al.* 2007)

(ii) la prédiction de résidus d'interactions au sein d'un complexe protéine-ligand (Barillari *et al.* 2008)

(iii) la prédiction de sites de liaisons (Kasahara *et al.* 2010; Volkamer *et al.* 2010)

(iv) la validation de descripteurs de sites de liaisons (Schalon *et al.* 2008; Weill *et al.* 2010)

(v) le profilage biologique pour l'identification de nouvelles cibles à l'aide des sites de liaisons (Defranchi *et al.* 2010)

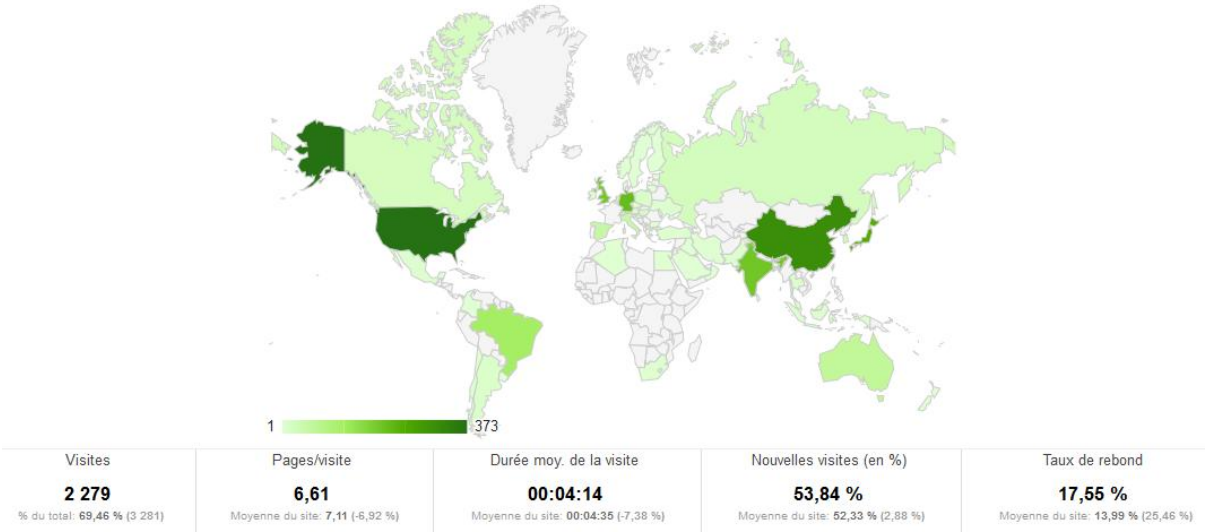


Figure 4: Répartition des utilisateurs de la sc-PDB pour la période du 30 Juillet 2011 au 30 Juillet 2012. Les données relatives à la France ont été écartées car le site est souvent utilisé dans notre laboratoire. Source : google analytics.

Le site web est actualisé annuellement avec le rajout d'une fonctionnalité à chaque mise à jour. A tire d'exemple, la distinction des sites au sein des protéines a été introduite pour la version 2009, l'année suivante, l'annotation des entrées a été revisitée et une recherche par nature et nombre d'interactions (Marcou *et al.* 2007) entre ligands et protéines a été introduit. Et pour la version 2011 actuelle, une recherche d'entrées par tailles de leurs sites est désormais disponible. D'autre amélioration et fonctionnalités sont prévus tel que la sélection des entrées qui présentent des interactions similaires, ainsi qu'une exploration des similarités des ligands, sites et interactions des entrées à travers des réseaux de graphes interconnectés.

7. Références

Barillari, C., G. Marcou and D. Rognan (2008). "Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores." J Chem Inf Model **48**(7): 1396-1410.

Berman, H., K. Henrick, H. Nakamura and J. L. Markley (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." Nucleic Acids Res **35**(Database issue): D301-303.

Defranchi, E., C. Schalon, M. Messa, F. Onofri, F. Benfenati and D. Rognan (2010). "Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements." PLoS One **5**(8): e12214.

Desaphy, J., K. Azdimousa, E. Kellenberger and D. Rognan (2012). "Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes." J Chem Inf Model.

Kasahara, K., K. Kinoshita and T. Takagi (2010). "Ligand-binding site prediction of proteins based on known fragment-fragment interactions." Bioinformatics **26**(12): 1493-1499.

Kellenberger, E., N. Foata and D. Rognan (2008). "Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems." J Chem Inf Model **48**(5): 1014-1025.

Kellenberger, E., P. Muller, C. Schalon, G. Bret, N. Foata and D. Rognan (2006). "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank." J Chem Inf Model **46**(2): 717-727.

Marcou, G. and D. Rognan (2007). "Optimizing fragment and scaffold docking by use of molecular interaction fingerprints." J Chem Inf Model **47**(1): 195-207.

Schalon, C., J. S. Surgand, E. Kellenberger and D. Rognan (2008). "A simple and fuzzy method to align and compare druggable ligand-binding sites." Proteins **71**(4): 1755-1778.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Eng **11**(9): 739-747.

Tanabe, M. and M. Kanehisa (2012). "Using the KEGG Database Resource." Curr Protoc Bioinformatics **Chapter 1**: Unit1 12.

Tietze, S. and J. Apostolakis (2007). "GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes." J Chem Inf Model **47**(4): 1657-1672.

UniProt, C. (2010). "The Universal Protein Resource (UniProt) in 2010." Nucleic Acids Res **38**(Database issue): D142-148.

Volkamer, A., A. Griewel, T. Grombacher and M. Rarey (2010). "Analyzing the topology of active sites: on the prediction of pockets and subpockets." J Chem Inf Model **50**(11): 2041-2052.

Weill, N. and D. Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites." J Chem Inf Model **50**(1): 123-135.

Chapitre 3 :

Enrichir les prédictions de modèles chémogénomiques à l'aide d'un descripteur de cavité 3D

Ce chapitre a fait l'objet d'une publication :

Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel.

Jamel Meslamani and Didier Rognan.

Journal of Chemical Information and Modelling, **2011**, 51, 1593–1603

1. Contexte

Les approches chémogénomiques computationnelles permettent la prédiction d'associations protéine-ligand en entrainant les algorithmes de machines d'apprentissages sur des données expérimentales connues dans le but de distinguer les bonnes associations protéine-ligand des mauvaises.

Dans un modèle chémogénomique, un complexe protéine-ligand est représenté à l'aide des descripteurs de son ligand et des descripteurs de sa protéine.

Les modèles chémogénomique développés à l'aide des machines d'apprentissage, prédisent les associations protéine-ligand sous forme d'une réponse binaire (modèles de classification : association possible ou pas) ou sous forme de prédiction des affinités (modèles de régression : valeurs de constante d'inhibition Ki par exemple).

Nous nous sommes intéressés dans cette première étude à élaborer des modèles chémogénomiques de classification SVM en utilisant les données disponibles sur des complexes cristallographiques issue de la base de données sc-PDB (Meslamani *et al.* 2011).

Nous avons choisi d'utiliser cette base de données cristallographiques, car nous voulions tirer parti d'un nouveau descripteur de cavité 3-D qui avait été développé au laboratoire (Weill *et al.* 2010).

Au commencement de ma thèse en 2009, les modèles chémogénomique étaient construits à partir de descripteurs de ligands et de descripteurs de séquences 2-D pour les protéines (van Westen *et al.* 2011). Nous avons ainsi décidé d'utiliser le descripteur de cavité 3-D FuzCav (Weill *et al.* 2010) développé dans notre laboratoire pour examiner l'apport d'un descripteur tridimensionnel de cavité par rapport à un descripteur 2-D de séquence protéique. Dès lors nous avons comparé les performances des prédictions des modèles générés à partir des différents descripteurs de protéines.

Les modèles générés et validés par un ensemble de données externe permettront la prédiction d'associations protéine-ligand couvrant ainsi un large spectre de protéines cristallisés. Mais ces modèles vont également servir à établir un profil biologique (défini par les protéines incluses dans les modèles) de bases de données.

Grâce à ces modèles, nous avons pu profiler deux chimiothèques différentes. La première contient 1 130 molécules synthétisées dans le groupe du Dr. Jean Jacques Bourguignon du Laboratoire d'Innovation Thérapeutique UMR 7200 (Illkirch), et la deuxième consiste en la chimiothèque nationale essentielle du CNRS (http://chimiotheque-nationale.enscm.fr) contenant 640 molécules.

2. Introduction

The ever-increasing availability of target-ligand binding data (Overington 2009; Wang *et al.* 2009) has significantly modified the scope of computational methods for analyzing and predicting structure activity relationships. Classical quantitative structure activity relationships (QSAR) focusing on a particular series of bioactive compounds binding to a single target are unable to extrapolate to different although neighboring chemical and target spaces.

Computational chemogenomics (Rognan 2007) (also referred to proteochemometrics (van Westen *et al.* 2011)) circumvents known limitations of pure ligand-based QSAR approaches to predict the binding of several ligands to several targets. Instead of considering only ligand properties, both ligand and target descriptors (Vert *et al.* 2008; Wassermann *et al.* 2009; van Westen *et al.* 2011) are used to predict target-ligand associations (binding affinities (Bock *et al.* 2005; Strombergsson *et al.* 2008) ,complex formation (Weill *et al.* 2009)).

Since similar receptors are supposed to bind to similar ligands (Klabunde 2007) predicting interactions in a target-ligand interaction matrix can be inferred from known data on similar ligands and/or similar targets. Most chemogenomic approaches have been limited to well-defined target spaces (receptor subtypes or enzyme isoforms (van Westen *et al.* 2011)) but extension to larger bioactivity spaces (G protein-coupled receptors (Klabunde 2007; Jacob *et al.* 2008; Wang *et al.* 2009) , kinases (Lapins *et al.* 2010), DrugBank (Nagamine *et al.* 2009) ,PubChem (Ning *et al.* 2009), Target Data Bank (Strombergsson *et al.* 2008), BindingDB (Strömbergsson *et al.* 2010)) have been reported.

To set up a chemogenomic model, three components are necessary. First, one should carefully choose a data set of experimental data describing target-ligand interactions (three-dimensional structures (Berman *et al.* 2000), binding affinities (Roth *et al.* 2004; Wang *et al.* 2004; Block *et al.* 2006; Liu *et al.* 2007; Overington 2009), or simple annotations (Jagarlapudi *et al.* 2009)) and remove target-ligand redundancy. Second, descriptors should be selected for describing ligands and targets (Wang *et al.* 2004; Block *et al.* 2006; van Westen *et al.* 2011). Many descriptors (1-D, 2-D, 3-D) have been reported in the literature (van Westen *et al.* 2011) but no general consensus has been reached to date. It can just be pointed out that 3-D descriptors for either ligands (Mahe *et al.* 2006) or targets (Wassermann *et al.* 2009; van Westen *et al.* 2011) did not seem to outperform simpler attributes. Last, a computational method should be chosen for

modeling the data. Many methods (Block *et al.* 2006) (partial least-squares, support vector machines, naïve Bayesian classifier, decision trees, random forest, neural networks, rough set modeling) have been used to model and predict chemogenomic data, again with no general consensus on what should be the best approach.

Chemogenomic models have clearly been shown to outperform pure ligand-based models in independent validation studies (Geppert *et al.* 2009; Weill *et al.* 2009). This enhancement has been proposed to be mainly due to ligand nearest neighbor effects (Geppert *et al.* 2009) (high similarity between a target-annotated ligand and a target-orphan compound), although one may argue that the corresponding study was limited to a few related targets.

The present study aims at precisely estimating the benefit (or disadvantage) of using an accurate 3-D pocket descriptor in chemogenomic modeling. Up to now, 3-D target descriptors (e.g., fixed-length pharmacophore fingerprint) have mostly been depending on a prior sequence alignment of a consensus set of cavity-lining residues (Jacob *et al.* 2008; Weill *et al.* 2009) or a full structural alignment of targets (Wassermann *et al.* 2009). As far as targets are highly related, the corresponding alignment is straightforward and the resulting similarity score will be relevant.

For targets sharing sequence and fold-independent pocket similarities, there is, however, a high risk of underestimating target similarity from a suboptimal alignment (Weill *et al.* 2010). Due to the constraints imposed by a 3-D target descriptor (restriction of target space to high-resolution X-ray or NMR target structures) and the paucity of accurate alignment-independent target (binding site) similarity measures (Yeturu *et al.* 2008; Weill *et al.* 2010) , true target-ligand binding site descriptors have not been used in chemogenomic modeling.

We herewith report the usage of a real 3-D cavity descriptor FuzCav (Weill *et al.* 2010) recently reported by our group for measuring local binding site similarities among unrelated targets. Briefly, the FuzCav descriptor is a vector of 4 834 integers reporting counts of all possible pharmacophoric feature (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, aromatic, hydrophobic) triplets from binding site-lining residues. Since cavity descriptions and comparisons are fast and independent of a prior 3-D structural alignment, the FuzCav descriptor is perfectly suited to generate a novel target kernel focusing on the most precise information required for describing ligand binding. To be confident in the binding site definition, the current study is limited to high resolution target-ligand complexes extracted from the sc-PDB data set (Kellenberger *et al.* 2006).

Using support vector machine (SVM) classifiers, separate kernels for measuring pairwise ligand similarities and target similarities have been investigated. Two conclusions could be drawn from the present report: (i) chemogenomic models clearly outperform simpler ligand-based models, notably when trained on a limited number of ligands (<40), and (ii) a true 3-D cavity descriptor slightly enhances the accuracy of SVM models when compared to a sequence-based target descriptor.

2. Computational Methods

2.1. Definitions

A ***target*** is defined as a macromolecule (protein, nucleic acid) to which a small molecular weight compound (***ligand***) binds to and for which an X-ray structure is available in the Protein Data Bank (PDB). Each target is assigned a unique name (***target name***) according to the UniProt nomenclature (UniProt 2010).

For each target-ligand complex, a ***binding site*** is defined as any residue (amino acid, cofactor, ion) close to the ligand (see precise definition in a previous report (Kellenberger *et al.* 2006)). Target-ligand complexes for which the ligand and the binding site are estimated to be druggable are stored in a subset of the PDB, named sc-PDB (Kellenberger *et al.* 2006). sc-PDB targets are clustered according to the pharmacophoric properties of their ligand-binding sites to yield clusters. A particular sc-PDB target may thus exhibit different ligand-binding sites located in the same cluster (if sites are similar) or in different clusters (if sites are dissimilar, e.g. catalytic and allosteric sites).

Alternatively, two different targets sharing similar binding sites (e.g., ATP-binding site of protein kinases) may be grouped in the same cluster. In the sc-PDB database, there is no redundancy at the target-ligand complex level, which means that two copies of the same target-ligand complex (obtained at different resolutions, for example) cannot coexist (the complex with the lowest resolution is kept as the single copy).

2.2. Training dataset

Target-ligand interaction data were gathered from the sc-PDB data set of druggable target-ligand X-ray structures (Kellenberger *et al.* 2006). In the sc-PDB archive, the ligand is considered from a purely pharmacological point of view (detergents, ions, and molecules devoid of any known biological activity are discarded) and explicitly defines the binding site as any surrounding residue (protein amino acid, cofactor, ion).

For each entry, atomic coordinates of the target, the binding site, and the atom type-curated ligand are stored. The in-house developed FuzCav algorithm (Weill *et al.* 2010) was then utilized to systematically measure the pairwise similarity of all 7 078 sc-PDB binding sites (release v 2009) to yield a full similarity matrix that was converted to a distance matrix (distance = 1-similarity score) and clustered by hierarchical clustering with the average linkage method using a FuzCav similarity threshold of 0.16 as a stopping criterion for cluster agglomeration. Only clusters annotated by more than 10 unique ligands were kept to finally yield 87 clusters describing 2 882 sc-PDB complexes (581 different targets and 2 605 different ligands).

For each cluster, a "mean ligand i" was identified as the one that is the most similar to all ligands j in the cluster as follows: $mean_i = \sum_i \sum_j d^2(i,j)$ with d=(1-Tc) and Tc=the similarity computed by the Tanimoto coefficient on SciTegic ECFP_4 circular 2-D fingerprints (Rogers *et al.* 2010). The 2 882 target-ligand PDB complexes were used as positive instances (true complexes) in the SVM classifications.

Negative instances (false complexes) were generated by randomly pairing decoy ligands with binding sites of the above-described 581 targets. Decoy ligands were selected from the in-house Bioinfo-DB database of commercially available *druglike* compounds (BioinfoDB 2011).

For each binding site cluster, a ratio of true to false complexes was fixed (20% true complexes, 80% false complexes). Decoy ligands were selected to span similar physicochemical property ranges (molecular weight and log P) than active ligands (the ensemble of ligands in a defined cluster) but to be different enough from the "mean ligand".

Any decoy whose Tanimoto coefficient (ECFP_4 fingerprint) was less than 0.15 was selected and considered dissimilar to the "mean ligand". A final random selection of remaining decoys provided the desired ratio of active ligands (true complexes) to inactive ligands (false complexes).

2.3. External Validation Set

Additional ligands of the 581 training set targets were retrieved from six bioactivity databases: ChEMBL_02 (Overington 2009) , PDSP Ki (Roth *et al.* 2000), Accelrys MDDR 2009.2 (Accelrys 2011), DrugBank 2.5 (Wishart *et al.* 2008), BindingDB (Liu *et al.* 2007) and STITCH 2.0 (Kuhn *et al.* 2010).

First, target names in bioactivity databases were compared to that of the sc-PDB target name and kept if identical or very similar. Second, all external ligands of the selected targets were retrieved as either SD files or SMILES strings. For each compound, a maximum of 400 conformers were generated using the default settings of Omega (OpenEye 2010). All compounds were ionized with Filter (OpenEye 2010) and their 2-D structures standardized with the ChemAxon standardizer utility (ChemAxon 2010). External compounds were then compared to the training sc-PDB ligands with ROCS (OpenEye 2010) and kept if similar to at least one training compound (Colorscore ≥ 0.5 and Comboscore ≥ 1.2).

After checking for duplicates and redundancy with sc-PDB ligands, 14 117 additional ligands for 531 out of the initial 581 targets (60 out of 87 clusters) could be finally retrieved.

A total of 328 308 positive instances were created as follows: an external ligand L_e was paired to target T_i of ligand L_i if L_i and L_e were found to be similar enough by ROCS (Colorscore ≥ 0.5 and Comboscore ≥ 1.2). In addition, L_e was also paired to target T_j if T_j shares with T_i an identical target name and cluster number. Negative instances were generated by randomly pairing, for each cluster, drug-like decoys to targets until the number of positive instances is reached. Decoys were retrieved as previously described for setting up the training set and selected only when different from training set decoys.

2.4. Ligand Descriptors

Ligands were represented by a hashed ECFP_4 extended-connectivity fingerprint (Rogers *et al.* 2010). The fingerprint was hashed to a 1024 bit string using a specified hash function in PipelinePilot 7.5 (Accelrys 2010) as described by Hert et al (Hert *et al.* 2004).

2.5. Target Descriptors

Three descriptors of increasing complexity were used for the targets: (i) the Uniprot name (UniProt 2010) (ii) the SPECTRUM sequence-based descriptor (Leslie *et al.* 2002) registering counts for each of the 20^3 possible tripeptides when browsing the amino acid sequence from the N-terminus to the C-terminus, (iii) a 3-D structural descriptor of the binding site as implemented in FuzCav (Weill *et al.* 2010). Up to six pharmacophoric features (hydrogen bond acceptor/donor, positive/negative ionizable, aromatic, aliphatic) were mapped to the C_α carbon atom of each sc-PDB binding site residue. Counts of all possible pharmacophoric triplets within specific distance ranges ([0-4.8 Å], [4.8-7.2 Å], [7.2-9.5 Å], [9.5-11.9 Å], [11.9-14.3 Å]) between C_α atoms were stored in a FuzCav fingerprint of 4 834 integers.

2.6. Support Vector Machines (SVM) Classification and Kernels

All models were generated using the SVM^{light} package (SVMlight 2010). The similarity between two target-ligand complexes $\langle c_i, c_j \rangle$ was measured as previously described (Jacob *et al.* 2008; Wassermann *et al.* 2009) from a target-ligand kernel $K(c_i,c_j)$ as the product of two separate kernels for the target pair $K_{target}(T_i,T_j)$ and the ligand pair $K_{ligand}(l_i,l_j)$:

$$\langle c_i, c_j \rangle = K(c_i, c_j) = K_{ligand}(l_i, l_j) \times K_{target}(t_i, t_j)$$

The Tanimoto kernel was used to calculate the similarity between pairs of ligands represented by their hashed ECFP_4 fingerprints.

$$KL = K_{ligand}(l_i, l_j) = \frac{\langle l_i,l_j \rangle}{\langle l_i,l_i \rangle + \langle l_j,l_j \rangle - \langle l_i,l_j \rangle}$$

Three kernels based on the above-described three target descriptors were utilized to measure pairwise target similarities as follows:

The Uniprot Name Kernel: The Uniprot name kernel was used to differentiate targets according to their names:

$$KT1 = K_{target}(t_i, t_j) = 1 \text{ if } t_i = t_j$$

$$KT1 = K_{target}(t_i, t_j) = 0 \text{ if } t_i \neq t_j$$

This kernel implies that only ligand information is used for each target separately in the learning process. From here on, using this kernel will be referred as a control ligand-based approach.

The Spectrum Kernel: It is a sequence similarity kernel used for target classification (Leslie *et al.* 2002). Each target is represented by a 20^3 dimensional vector counting occurrences of all possible tripeptides in the sequence. The Spectrum kernel between two targets is then computed as

$$\mathrm{KT2} = \mathrm{K}_{target}(\mathrm{t_i}, \mathrm{t_j}) = \frac{\langle \mathrm{t_i}, \mathrm{t_j} \rangle}{\sqrt{\langle \mathrm{t_i}, \mathrm{t_i} \rangle} \times \sqrt{\langle \mathrm{t_j}, \mathrm{t_j} \rangle}}$$

The FuzCav Similarity Kernel: Targets are represented by their FuzCav descriptors, and the similarity between two cavities is measured by a standard Tanimoto coefficient as follows:

$$\mathrm{KT3} = \mathrm{K}_{\mathrm{target}}(\mathrm{t_i}, \mathrm{t_j}) = \frac{\langle \mathrm{t_i}, \mathrm{t_j} \rangle}{\langle \mathrm{t_i}, \mathrm{t_i} \rangle + \langle \mathrm{t_j}, \mathrm{t_j} \rangle - \langle \mathrm{t_i}, \mathrm{t_j} \rangle}$$

Table 1 shows a summary of all descriptors and kernels used in this study.

Object	Descriptor	Kernel	Index
Ligand	ECFP_4 fingerprints	Tanimoto	KL
Target	Uniprot name	Uniprot	KT1
	Tripeptide occurrence	Spectrum	KT2
	FuzCav fingerprint	Tanimoto	KT3
Target-Ligand		Tanimoto ⊗ Uniprot	KTL1
Complex		Tanimoto ⊗ Spectrum	KTL2
		Tanimoto ⊗ FuzCav	KTL3

Table 1: Descriptors and Kernels

A 5-fold cross-validation procedure was used to split each of the 87 clusters five times into a training (four-fifths of the data set) and a test set (one-fifth of the data set) and analyze the predictivity of SVM models on the remaining test sets, using the best trade-off C value optimized for each model.

2.7. Model Evaluation

Statistical parameters for evaluating the different SVM models were the recall, precision, specificity, and F-measure:

Recall = TP/(TP+FN)

Precision = TP/(TP+FP)

Specificity = TN/(TN+FP)

F-measure = 2(Recall*Precision)/(Recall+Precision).

Where TP = true positives, FP = false positives, TN = true negatives, FN = false negatives.

3. Results and Discussion

3.1. Composition and Analysis of the Training Data Set

The current study specifically aims at unambiguously evaluating the benefit of using true 3D binding site information in chemogenomic classification models of target-ligand binary associations. Since the target kernel used in this study focuses on ligand binding site similarity and local models were derived for target clusters of similar binding sites, the present analysis is constrained to a reduced set of 87 clusters covering 581 unique targets for which (i) a high-resolution X-ray structure of a ligand-bound state is available and (ii) at least 10 ligands have been cocrystallized with similar binding sites.

There is no real consensus about the minimal number of ligands required to describe an activity class. In one of the most exhaustive studies published to date (Keiser *et al.* 2007) a threshold of five different ligands was used for druglike compounds. Since PDB ligands are notoriously less diverse than druglike compounds, we therefore chose a slightly higher value of 10 compounds.

Most of the 87 target clusters are paired with less than 20 compounds (Figure 1A) and only four clusters (see a complete description of cluster contents in Supporting Information Table 1) are populated by more than 100 different compounds. The corresponding binding sites and target families have either been heavily investigated (e.g., cluster 30, HIV-1 protease catalytic site;

cluster 57, serine/threonine target kinase ATP-binding site; cluster 50, trypsin-related catalytic site) or exhibit ligand promiscuity (cluster 5, nucleotide-binding sites).

The diversity of each training ligand set, estimated as the mean pairwise intermolecular dissimilarity (Turner *et al.* 1997) is remarkably high (Figure 1B). Sixty-two out of the 86 training sets show a diversity above 0.7. Since target space has been discretized by binding site diversity, the number of unique targets in each cluster varies considerably from a single target (26 out of 87 clusters) to a maximum of 157 (cluster 5).

A very large majority of clusters regroup less than five different targets (Figure 1C). The annotation of the target training set according to the Enzyme Commission (EC) number (http://www.chem.qmul.ac.uk/iubmb/) reveals no major change with respect to the sc-PDB data set, suggesting that our selection of target-ligand complexes has not introduced any major bias toward a particular target space (Figure 1D).

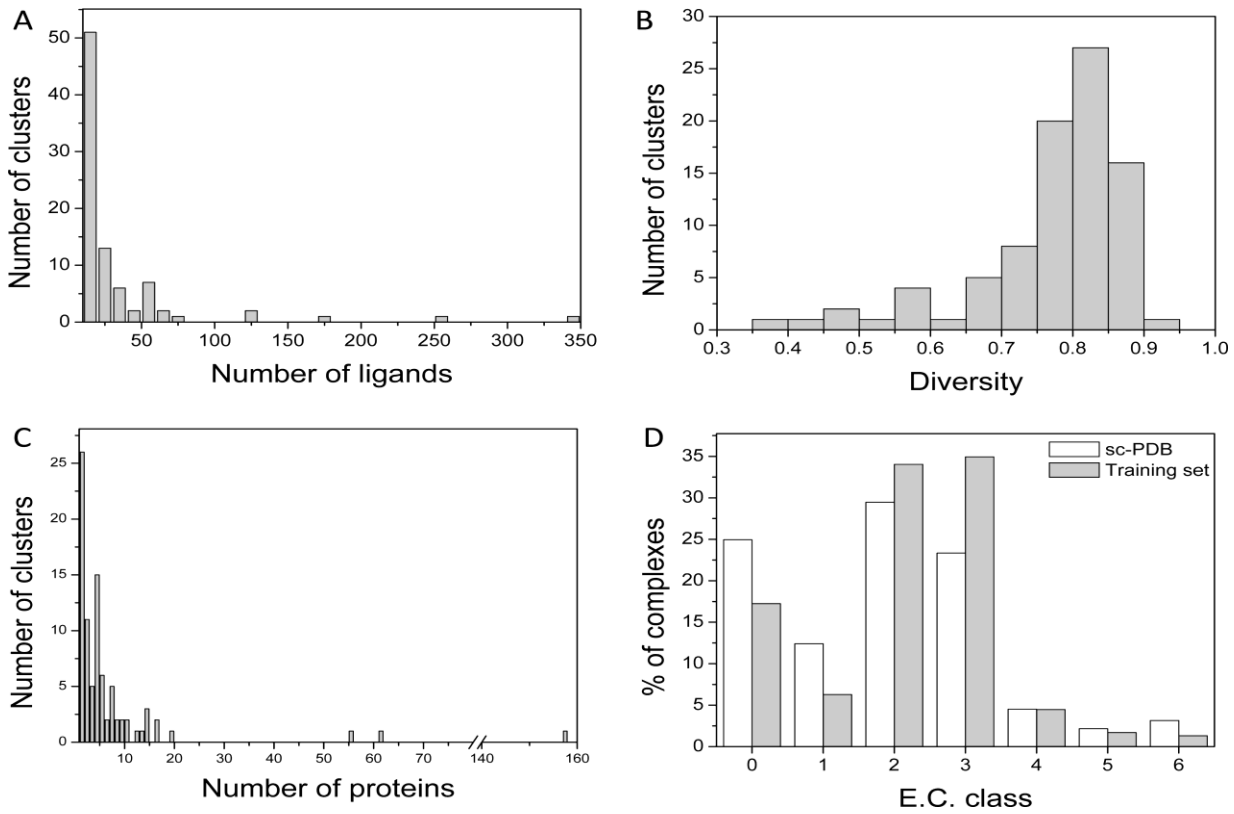


Figure 1: Diversity analysis of the target ligand training set. **(A)** Distribution of the number of training set ligands for 87 binding site clusters. **(B)** Distribution of the chemical diversity of training set ligands for 87 binding site clusters. **(C)** Distribution of the number of training set targets for 87 binding site clusters. **(D)** Functional annotation of training set targets by Enzyme Commission (EC) number: 0, no EC number; 1, oxidoreductase; 2, transferase; 3, hydrolase; 4, lyase; 5, isomerase; 6, ligase.

3.2. Cross-Validation Results

Cross-validation of SVM models on the training data set clearly indicates a superior performance of chemogenomic with respect to ligand-based models. Prediction of target-ligand pairing (2 882 true pairs; 11 528 decoys) was realized using 87 local SVM models and a 5-fold cross-validation protocol. Each local model addresses targets whose binding sites are grouped in the same cluster, their corresponding PDB ligands, and the related decoys.

The accuracy of all models was estimated from the F-measure parameter (see Computational Methods), which presents the advantage to take both recall and precision into account.

Varying the target kernel on this data set unambiguously demonstrates that the best models were obtained using a 3D binding set kernel (KTL3), then with a target sequence-based kernel (KTL2), and last with a purely ligand-based approach (Figure 2).

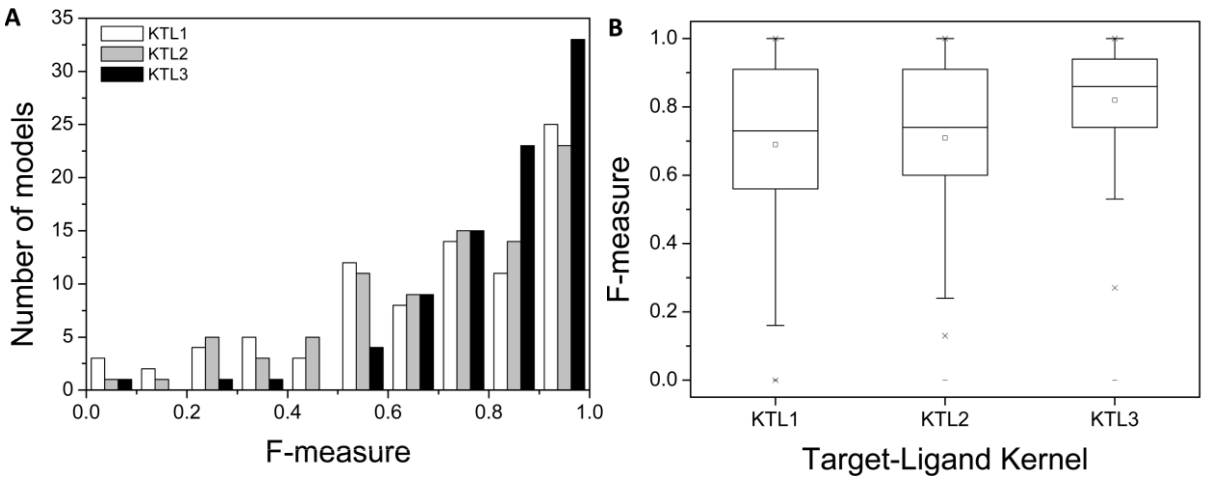


Figure 2: Accuracy, estimated by the F-measure of 87 local SVM models for predicting target-ligand pairing (2 882 true pairs and 11 528 false pairs) using a pure ligand-based (KTL1) and two chemogenomic approaches based on a target sequence (KTL2) and target 3D structure (KTL3) kernel. **(A)** Distribution of the F-measure. **(B)** Box-and-whisker plot of F-measure distributions for the three target-ligand kernels. The box delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

Considering an F-measure value above 0.5 as acceptable, only 3 out of 87 models (models 45, 548, and 645; see complete results in Supporting Information Table 2) are unsatisfactory (Figure

3). The low F-measure value observed for these three models are mostly attributable to low recall values, the specificity being still always above 90% (Figure 3).

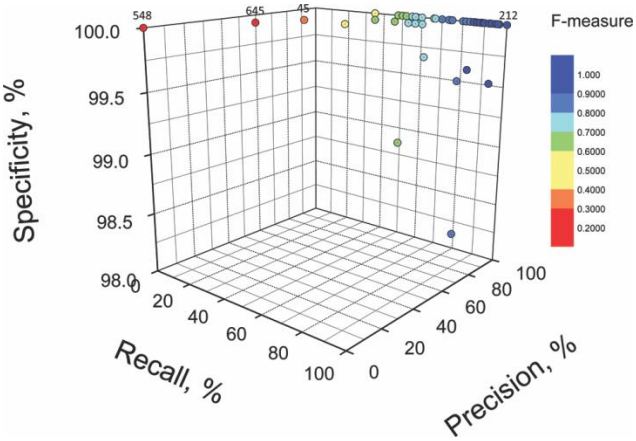


Figure 3: Recall, specificity, and precision of 87 SVM local models using the KTL3 target-ligand kernel. Data are color-coded according to the F-measure value. Binding site clusters discussed in the text are labeled according to their number.

Inspecting the training ligands and cavities of these difficult clusters (e.g.,cluster 45) reveals much poorer pairwise ligand and pairwise cavity similarities with respect to clusters (e.g., cluster 212), yielding nearly perfect SVM classification models (Figure 4).

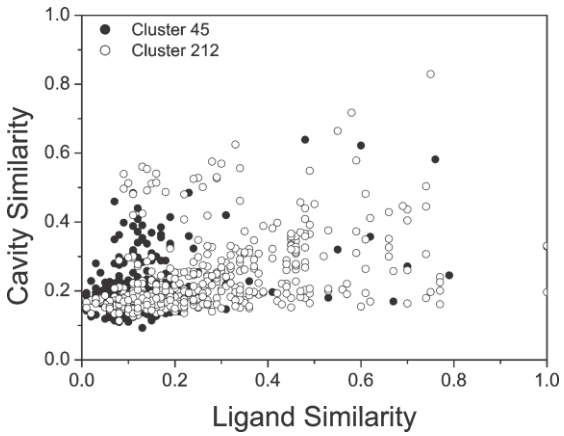


Figure 4: Pairwise ligand and cavity similarities for target-ligand complexes yielding poor (cluster 45, F-measure = 0.48) and perfect (cluster 212, F-measure = 1.0) SVM classification models. Cluster 45 comprises seven targets and 28 ligands, whereas cluster 212 is composed of five targets and 24 ligands (see Supporting Information Table 1). Similarities are computed according to the Tanimoto coefficient on hashed ECFP_4 ligand fingerprints and FuzCav cavity fingerprints.

Failure of SVM models to recall true target-ligand complexes belonging to these three clusters may thus be attributable first to an incomplete coverage of the corresponding target-ligand space by existing PDB complexes and then to the promiscuity of the corresponding binding sites toward different chemotypes.

We next carefully inspected cross-validation results for the three target-ligand kernels and all 87 clusters, notably the impact of using target information either in the form of amino acid composition (KLT2 kernel) or binding site 3D structure (KTL3 kernel).
Plotting the difference of the F-measure for each local model clearly shows a benefit of combining ligand and target descriptors for almost all models. Interestingly, the benefit tends to be more important for models trained with a limited number of positive instances, whatever the kernel used (Figure 5).

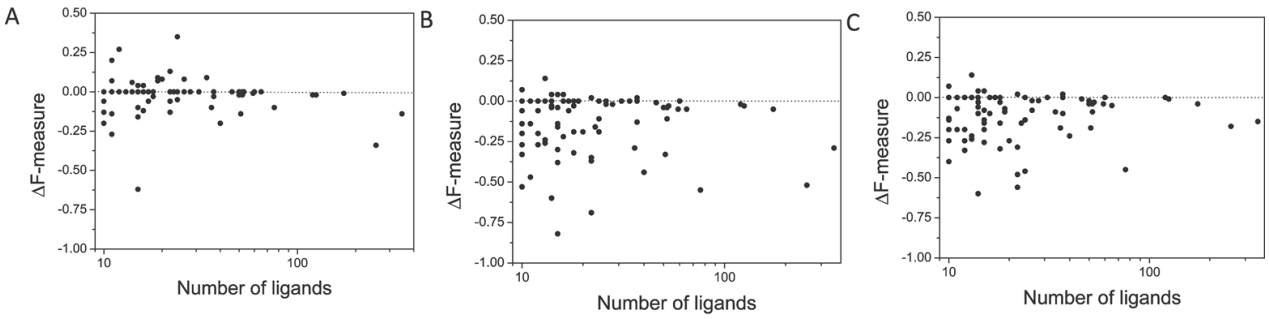


Figure 5: Difference in the F-measure value (ΔF-measure) of three chemogenomic models (KTL1, KTL2, KTL3) in classifying 2 882 true target-ligand PDB complexes and 11 528 target-ligand decoys: **(A)** F(KTL1)-F(KTL2), **(B)** F(KTL1)-F(KTL3), **(C)** F(KTL2)-F(KTL3). Values are plotted according to the number of true ligands for each binding site cluster.

Likewise, chemogenomic models also profit from the number of targets taken into account (Figure 6). Single target models are not suitable for the kind of approach developed herein, a benefit being seen when considering more than four or five targets (Figure 6).

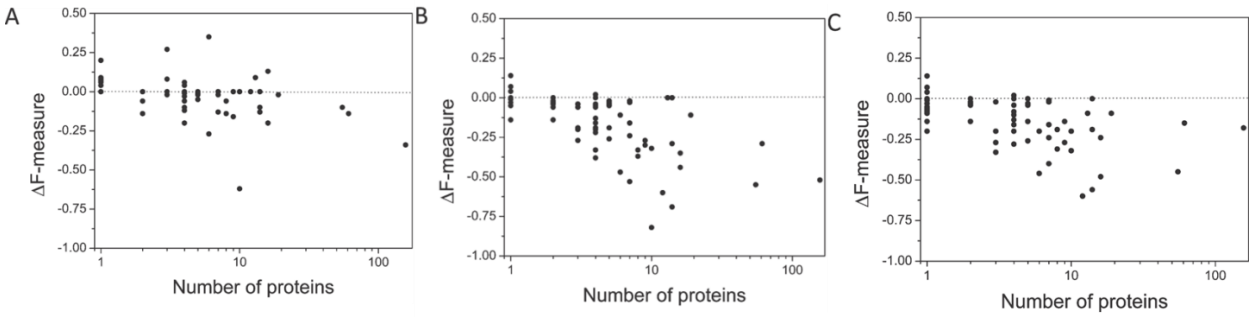


Figure 6: Difference in the F-measure value (ΔF-measure) of three chemogenomic models (KTL1, KTL2, KTL3) in classifying 2 882 true target-ligand PDB complexes and 9 128 target-ligand decoys: **(A)** F(KTL1)-F(KTL2), **(B)** F(KTL1)-F(KTL3), **(C)** F(KTL2)-F(KTL3). Values are plotted according to the number of different targets for each binding site cluster.

Conversely to previous studies (Jacob *et al.* 2008; Wassermann *et al.* 2009) that did not noticed any advantage of using target 3-D information with respect to much simpler descriptors, the present report demonstrates a significant superiority of the target kernel (KT3) measuring binding site 3D similarities with respect to the sequence-based Spectrum KT2 kernel (Figures 5 and 6). The observed discrepancy to previous results may be explained by three important factors. First, target space has been here discretized by a novel approach considering binding site 3-D similarity, irrespective of the target name and family. Second, it is the first time to the best of our knowledge that a true binding site 3-D descriptor FuzCav (Weill *et al.* 2010) which has shown its ability to discriminate true similar from true dissimilar cavities, has ever been applied to chemogenomic modeling. Third, the target-ligand data set has been restricted to target-ligand complexes of known X-ray structures in which a pharmacological ligand binds to a druggable cavity (Kellenberger *et al.* 2006). Assuming that similar receptors bind similar ligands (Klabunde 2007), it is therefore logical that focusing on the target binding sites at the most precise level provides a true advantage in predicting target-ligand associations.

Of course, we acknowledge that the applicability domain of the current approach is limited to a tiny target-ligand space and may exclude important targets (e.g., membrane receptors) for which precise 3-D structural information is missing. The main advantage of using local models is that results may be examined for different target-ligand subspaces to infer possible guidelines for best practice chemogenomic modeling. Of course, all positive and negative instances may be pooled into a single SVM model. A 10-fold cross-validation procedure was applied to the entire data set and recall, selectivity, specificity, and F-measure of the global model were determined as

previously described for the local model. Overall statistics were in favor of a global modeling procedure, whatever the target kernel used (Table 2).

Model	Local		Global
	Mean	Median	
KTL1	0.69	0.73	0.75
KTL2	0.7	0.74	0.90
KTL3	0.82	0.86	0.91

Table 2: Performance of a Global and Local Models (F-measure) on the Cross-Validated Training Set

Notably, the sequence-based target model (KTL2) appears as good as the structure-based kernel (KTL3) in a global SVM modeling. We, however, acknowledge that cross-validated models described here may exhibit overestimated statistics due to the decoys selection protocol. Selecting decoys is indeed a tricky process for which many routes may be followed. The selection based on ECFP_4 similarity to a mean ligand was just done to be sure that actives and decoys do not share the same scaffolds.

What is important is that the models are extensively challenged by external test sets and that selectivity, specificity, and precision remain acceptable. We therefore designed the largest possible external ligand set to challenge the cross-validated models.

3.3. Large-Scale Prediction of Target-Ligand Associations from an External Data Set Validation Results

The best models derived after cross-validation may not be the most predictive when applied to an external data set (Weill *et al.* 2009). Likewise, the superiority of the KTL3 kernel noticed in the cross-validation study may disappear when applied to an external test set. We therefore decided to validate the previously reported models, using exactly the same three target-ligand kernels, on the largest possible external data set.

For each of the 581 sc-PDB targets, six external ligand sources (ChEMBL (Overington 2009), PDSP Ki (Ning *et al.* 2009), MDDR (Accelrys 2011), DrugBank (Wishart *et al.* 2008),

BindingDB (Liu *et al.* 2007) and STITCH 2.0 (Kuhn *et al.* 2010)) were browsed to retrieve 14 117 compounds for 531 targets of the training set.

Since, target space is organized by binding site and not target name, we had to verify that both training and external ligands are likely to share the same binding site by computing their pharmacophore-annotated shape similarity and only to retain pairs for which the similarity, as estimated by ROCS, is above a certain threshold (Colorscore ≥ 0.5 and Comboscore ≥ 1.2).

For each of the 60 remaining target clusters, an external test set of ligands could be defined exhibiting in most of the cases more than 10 compounds and an acceptable chemical diversity (Figure 7).

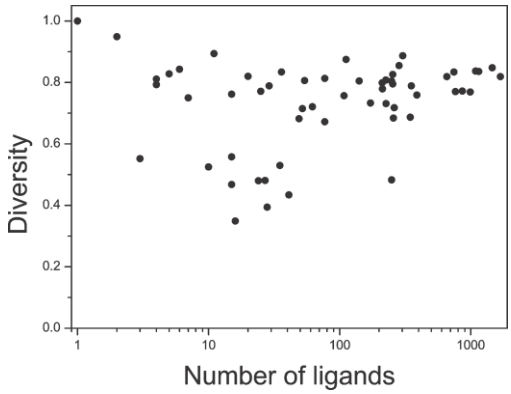


Figure 7: Diversity of the 60 external test sets, estimated from the mean pairwise intermolecular dissimilarity (Turner *et al.* 1997) of the corresponding ligand ECFP_4 fingerprints, as a function of their size (number of ligands).

In the current external validation, the ratio of positive to negative instances was fixed to 1 for all clusters. Modification of this ratio by varying the number of target-ligand decoys did not affect the obtained results. Target-ligand pairing was predicted in two possible modes, by PDB entry and by target binding site.

Predicting target-ligand binding on a PDB entry basis corresponds to answer the following question: "To which PDB entry (e.g., 1dm2) is this compound predicted to bind to?"

For many entries, however, multiple copies of the same target binding site are available. For example, cluster 57 regroups 347 PDB entries out of which 99 correspond the ATP binding site of the cell division protein kinase 2 (CDK2). Predicting the binding of a true CDK2 inhibitor from the test set may fail for some of the 99 entries but succeed for some others. Fusing the results for each target binding site thus enables one to escape from particular binding site singularities (e.g., site-directed mutagenesis, induced fit, amino acid omission in the PDB entry).

In the second prediction mode, a target-ligand pairing was then predicted for each target binding site; in other words, to answer the question, "To which binding site (e.g., CDK2 ATP-binding site) is this compound predicted to bind to?"

In order to directly compare predictions from local models with that from the global one, the 60 external test sets were iteratively used for predicting target-ligand pairing with the global model.

Four main observations could be drawn by analyzing the prediction statistics (Figures 8 and 9 and Table 3) :

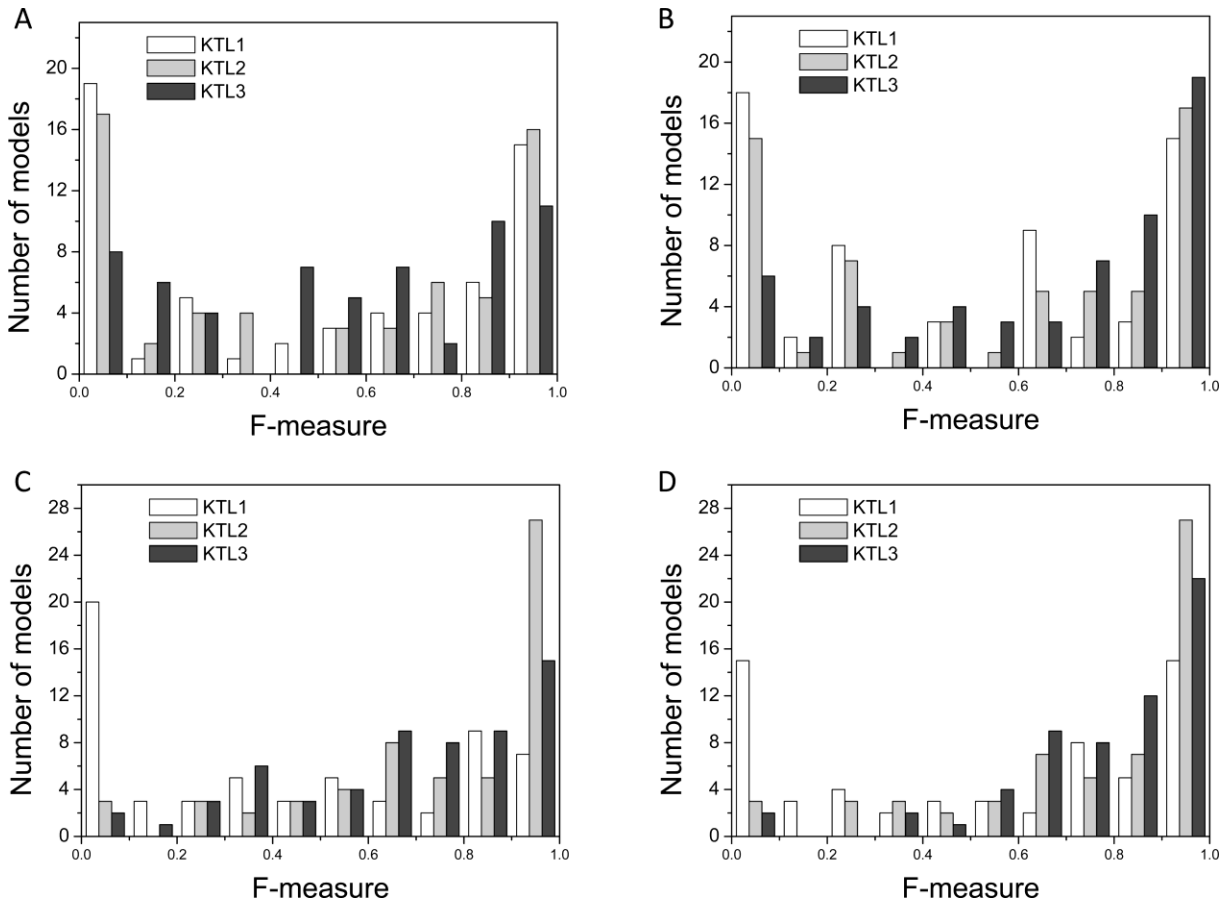


Figure 8: Distribution of the F-measure values for three SVM target-ligand kernels (KTL1, KTL2, KTL3) predicting target-ligand binary association (PDB entry prediction mode) for an external test set (14 114 ligands, 531 targets, 328 308 true complexes seeded with the same number of false target-ligand decoys): **(A)** local models, prediction by PDB entry; **(B)** local models, prediction by target binding site; **(C)** global model, prediction by PDB entry; **(D)** global model, prediction by target binding site.

	Local		Global	
Model	Mean	Median	Mean	Median
KTL1	0.499	0.575	0.419	0.385
KTL2	0.508	0.595	0.741	0.840
KTL3	0.548	0.605	0.675	0.725
KTL1	0.472	0.460	0.521	0.680
KTL2	0.538	0.655	0.751	0.860
KTL3	0.659	0.770	0.779	0.865

Table 3: Performance of a Global and Local Models (F-measure) on the External Test Set

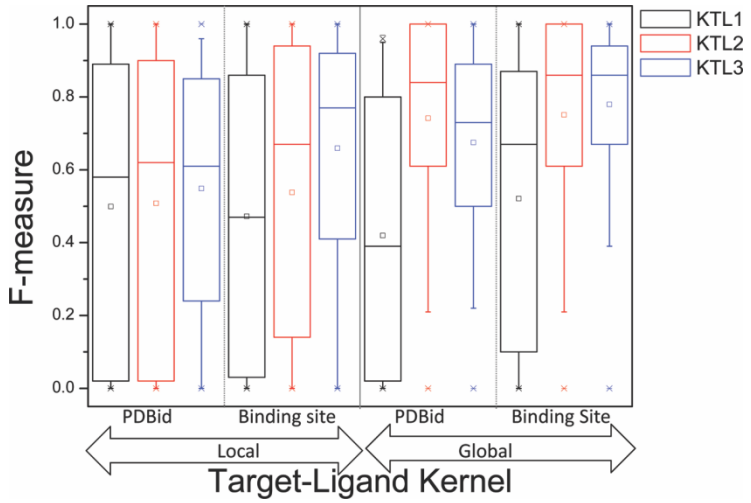


Figure 9: Box-and-whisker plot of F-measure distributions for three SVM target-ligand kernels (KTL1, KTL2, KTL3) predicting target-ligand binary association (PDB entry prediction mode) for an external test set (14 114 ligands, 531 targets, 328 308 true complexes seeded with the same number of false target-ligand decoys). The box delimits the 25th and 75th percentiles, and the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

(i) Chemogenomic models (KTL2, KTL3) almost always outperform the ligand-based model (KTL1 kernel), as evidenced by the higher proportion of models with F-measure higher than 0.5, the narrower distribution, and the higher mean and median values.

(ii) Among chemogenomic models, using a structure-based kernel (KTL3) leads to slightly better predictions than a sequence-based kernel (KTL2).

(iii) Global models perform better than local models.

(iv) As expected from the data fusion, predicting target-ligand pairing on a target binding site basis is preferred to a PDB entry basis, whatever the modeling procedure (local or global).

Out of the 12 prediction modes (see complete results in Supporting Information Tables 3-6), the best predictions are obtained using a global model with a structure-based kernel (KTL3) and predicting pairing on a target binding site basis. Since precision and specificity are excellent (99%) for all 60 external test sets (see complete results in Supporting Information Table 6), the predictive property of the global model is therefore almost dependent on the recall capability (71%). Altogether, an F-measure value higher than 0.5 was obtained for 42 out of the 60 external test sets.

Examining the recall values against the diversity and the size (number of external ligands) of the external test sets did not reveal clear trends. Considering the unprecedented diversity and size of the external test set (>14 000 ligands and 531 targets), the performance of this chemogenomic model is remarkable and suggests the use of cavity 3-D kernels whenever possible to predict target-ligand pairing. Using a sequence-based target kernel in a global model also provides very good statistics, but with a wider distribution of F-measure values (Figure 9) and slightly poorer models (F-measure < 0.5) with respect to the structure-based kernel (Figure 8).

The accuracy of the target-sequence kernel is however quite promising, notably for predicting ligand pairing to targets of unknown 3-D structure, and therefore considerably extends the applicability domain of chemogenomic QSAR modeling. Altogether, we therefore recommend different possible strategies with respect to the question that has to be answered.

If one wants to exactly know to which protein a ligand may bind, it is better to use a global model (KTL3 kernel), which gives a probability of target-ligand association for every target of our data set.

However, if the issue is to know to which kind of binding site (e.g., ATP-biding site of Ser/Thr protein kinases, catalytic site of serine endopeptidases) a ligand (or a focused library) may bind to, it is better to use local models focusing on well-defined target spaces.

Moreover, the present study permits one to design good chemogenomic modeling practices with respect to the existing knowledge on targets and their ligands. Taking into account target information (sequence or structure) makes sense only when the number of known ligands for this peculiar target is sufficiently low (roughly below 40-50).

When this is the case, using structural information about the ligands binding site clearly provides an advantage with respect to simpler sequence information in predicting novel target-ligand associations.

Since these conclusions are similarly drawn by models derived from the training set and more importantly from the external test set, we therefore propose a pragmatic in-silico target profiling strategy taking advantage of three possible situations.

- In the first one (3-D structure of the target is available and less than 50 ligands are known), we propose to combine a ligand descriptor and a 3-D binding site descriptor in separate kernels.
- In scenario 2 (3-D structure of the target is not available and less than 50 ligands are known), we propose to combine a ligand descriptor and a target sequence descriptor in separate kernels.
- Last, in the case where more than 50 ligands for a particular target are known, we propose to use a simple ligand similarity kernel and a SVM classification model.

These proposals are based on the statistics (mainly the F-measure) derived from models applied to the external data set.

Like any model, the conclusions are of course partly dependent on the input data and therefore the decoy ligand selection and known actives. We, however, believe that the general trends indicated in the present study are data-set-independent, notably because of the very large external test set used to challenge the current SVM models.

4. Conclusion

Chemogenomic (or proteochemometric) QSAR modeling methods are taking an increasing importance in predicting, at a very high throughput, the binding of numerous ligands to numerous targets.

A key advantage of chemogenomic modeling with respect to ligand centric methods is the applicability to orphan targets or at least to targets for which ligand information is sparse.
By looking at known data on the neighboring target-ligand space, novel target-ligand relationships may be inferred. Various approaches for predicting target-ligand binary association or binding affinities have been proposed and shown to be remarkably efficient and predictive. Whatever the method, a target-ligand space to which it is applicable must be defined. Defining this space will strongly influence both the methods and descriptors used to describe targets and their ligands.
An exhaustive definition (e.g., any biologically relevant target with more than five ligands) implies a relatively fuzzy and rough definition of target space, usually at the amino acid sequence level, since precise information on binding site location and target-ligand interactions are missing. Conversely, a restricted definition as the one used in the current study (any biologically relevant target cocrystallized with drug-like compounds) considerably limits the applicability range of the method, but it enables a fine modeling of target cavity attributes responsible for ligand binding.
In the current report, we unambiguously demonstrate that target binding site descriptors, when available, enhance the performance of chemogenomic methods in predicting target-ligand binary associations, with respect to simpler sequence-based target attributes and pure ligand-based modeling.
The proposed method, despite its accuracy, still suffers from two drawbacks:

(i) it is only applicable to 531 targets (mostly enzymes) of known high-resolution X-ray structure.

(ii) it just predicts the likehood of target-ligand binding but not a binding affinity nor functional effects (agonist vs antagonist, competitive vs noncompetitive inhibition). Extending

the approach to a much larger target-ligand space (e.g.,membrane receptor and their ligands) requires either changing target descriptors (e.g., sequence-based attributes), although we know this change may be slightly detrimental to the model accuacy, or following a more pragmatic target-based approach utilizing various models for different target-ligand spaces (e.g., 3-D binding site kernels for PDB targets and sequence-based kernels for other targets). One may even imagine varying the property to predict according to known data (binding affinities or target-ligand association) for particular target-ligand subspaces. In many instances (e.g., biogenic amine G protein-coupled receptor ligands), so many data are available that chemogenomic methods are not required or even not suitable, as evidenced by the present report, for accurate predictions. Along with the ever increasing amount of public target-ligand binding data, we believe that target class specific methods, descriptors, and property predictions will enhance the applicability of ligand profiling methods to a large array of biologically relevant targets and propose usable computational preclinical safety profiles in early drug discovery phases.

5. Commentaires et applications dans le cadre d'un profilage

Le modèle global généré dans cette étude est capable d'établir des prédictions sur 581 cibles différentes. Dès lors, il constitue un bon outil pour du profilage sur ces cibles.

Ce modèle a servi à deux profilages, le premier sur une chimiothèque de 1 130 molécules du groupe du Dr. Jean Jacques Bourguignon du Laboratoire d'Innovation Thérapeutique UMR 7200 (Illkirch), et le deuxième sur la CNE.

Malheureusement, aucune validation expérimentale n'a pu être réalisée au moment de la rédaction du manuscrit. Pour les résultats de la première chimiothèque, un projet d'analyse est en cours mais d'autres priorités ont fait que la validation a été retardée.

Concernant les données de profilage de la CNE, les différents laboratoires partenaires n'ont pas souhaité partager les résultats avec toute la communauté scientifique. Les résultats ont été communiqués à chaque laboratoire mais nous ne savons pas si les propriétaires des molécules ont essayé de faire valider nos prédictions.

Bien entendu, il a fallu définir un domaine d'applicabilité avant d'effectuer le criblage sur le modèle global. Il n'y a malheureusement aucune étude qui se penche sur ce problème dans le cadre des modèles chémogénomique.

Nous avons proposé un domaine d'applicabilité basé sur les proches voisins des molécules (approche kNN) inspiré de l'étude de Shen et al. (Shen *et al.* 2002)

5.1. Domaine d'applicabilité

Nous avons choisi de définir un domaine d'applicabilité pour chaque cible incluse dans le modèle. Ce domaine d'applicabilité déterminera si la molécule à profiler appartient au non au domaine d'applicabilité.

Le domaine est défini pour chaque cible et à partir de la distribution de similarité de ses ligands. Pour chaque molécule i de l'ensemble des molécules d'entrainement N de la cible A, ses *k* plus proches voisins (*k* = 3 choisi pour cette étude) sont déterminés et une distance D est calculée selon :

$$D = \sum_{i=1}^{N}\sum_{j=1}^{k} d_{i,j}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \langle d \rangle)^2}$$

d_i est la distance pour chaque molécule i avec ses *k* plus proches voisins.

$\langle d \rangle$ est la moyenne des distances des molécules d'entrainement N avec leurs *k* plus proches voisins.

d = 1 – Tanimoto (ECFP_4)

Pour chaque molécule *x* à profiler, ses *k* plus proches voisins à partir de l'ensemble des ligands N de la cible A sont déterminées et une distance D_{TS} est calculée :

$$D_{TS} = \frac{1}{k}\sum_{i=1}^{k} d(x,i) \text{ où i est un ligand de la cible A.}$$

Si la distance $D_{TS} \leq D + Z \times \sigma$ (Z est un paramètre empirique égal à 0.5) alors la molécule *x* est incluse dans le domaine d'applicabilité de la cible A.

Cette approche est très employée et très efficace car elle permet d'écarter facilement les molécules différentes des ligands connus. L'inconvénient de cette approche est qu'elle est sensible à un éventuel effet de bord où une molécule va être incluse dans le domaine bien qu'elle ne le devrait pas. (Figure 10)

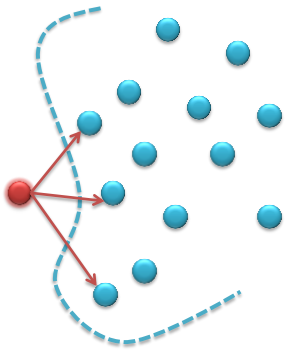


Figure 10 : Effet de bord possible avec le domaine d'applicabilité *k*NN. La molécule en rouge devrait être considérée comme hors du domaine.

5.2. Détermination de profil biologique pour les molécules

Lorsqu'on profile une chimiothèque sur une série de cibles, il est intéressant de voir quelles molécules possèdent des profils biologiques similaires. Ceci peut être un bon moyen pour évaluer la diversité de la chimiothèque afin de l'améliorer et même anticiper l'association à d'autres cibles. Pour ce faire, nous avons généré une empreinte d'affinité qui se compose d'une chaine de bits.

Dans un premier temps, une empreinte de 581 bits a été créé où chaque bit correspond à une association possible entre la molécule profilée et une des 581 cibles disponibles.

Dans un deuxième temps, une empreinte de 645 bits a été générée où chaque bit correspond à une association possible entre la molécule profilée et un site de liaison unique.

Un coefficient de Tanimoto permet d'évaluer la similarité entre deux empreintes dans les deux cas. Pour permettre la sélection de molécules de profil biologique similaire un seuil empirique de 0.7 a été choisi.

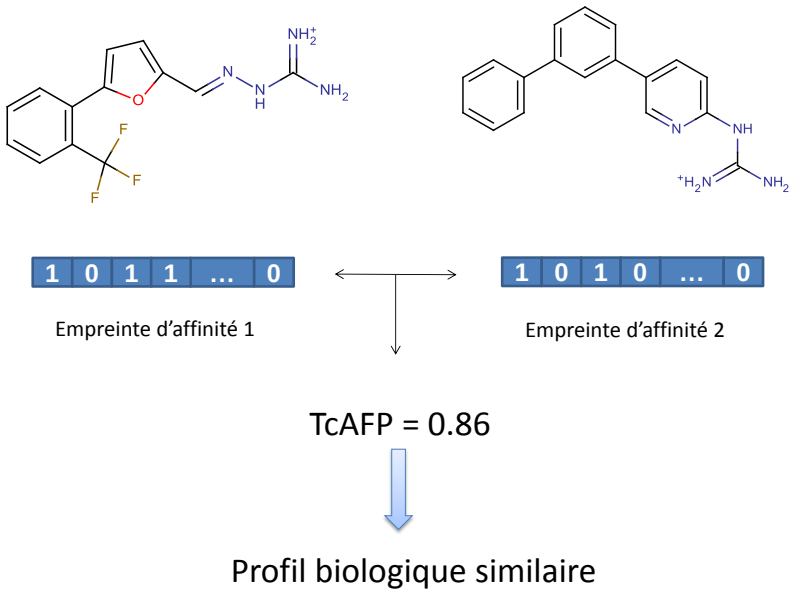


Figure 11 : Exemple de deux molécules possédant un profil biologique similaire (Tanimoto = 0.86 sur les empreintes d'affinités protéique) mais des faibles similarités structurales Tanimoto MACCS = 0.43 et Tanimoto ECFP_4 = 0.19

5.3. Interface web dynamique pour la présentation des résultats de profilage

Nous avons développé une interface web pour présenter les résultats de profilage aux équipes de chimistes qui ont contribué en fournissant leurs molécules. Les pages web sont un moyen très facile pour mettre en évidence des résultats. Ainsi, nos collaborateurs chimistes peuvent explorer leurs résultats d'une manière très simple. Dans le cadre du profilage des molécules de la CNE, un identifiant et un mot de passe ont été générés pour chaque équipe pour des soucis de confidentialité.

Les résultats de profilage ont été insérés dans une base *mysql* v.5.0.95 (http://www.mysql.com) et des pages web dynamiques en *JavaServer Pages* (http://fr.wikipedia.org/wiki/JavaServer_Pages) ont été élaborées pour permettre l'affichage des données.

Nous allons montrer le contenu de cette interface pour l'exemple du profilage de la CNE. L'interface est disponible à cette adresse : http://cheminfo.u-strasbg.fr:8080/CNE .

Voici un exemple de résultats affiché par cette interface:

- Page d'accueil contenant les possibilités de navigation : affichage des résultats par molécule, par protéine, affichage de molécules de profils biologique similaires, statistiques sur les protéines potentielles pour chaque molécule (Figure 13).
- Résultats par nom de protéine : affichage de toutes les molécules potentielles (Figure 14).
- Résultats par molécule : affichage de toutes les protéines potentielles (Figure 15) ; des molécules qui ont un profil biologique similaire (Figure 16) ; le nombre de protéines classés par familles (Figure 17).

Figure 12 : Accueil de l'interface pour consulter les résultats. Le formulaire doit être rempli avec l'identifiant et le mot de passe pour assurer la connexion. Adresse : http://cheminfo.u-strasbg.fr:8080/CNE

Figure 13 : Page d'accueil pour notre laboratoire qui permet d'explorer les résultats par les cibles prédites lors du profilage, par les molécules profilées, les molécules qui ont des profils biologique similaires et finalement des statistiques sur la classification des cibles prédites.

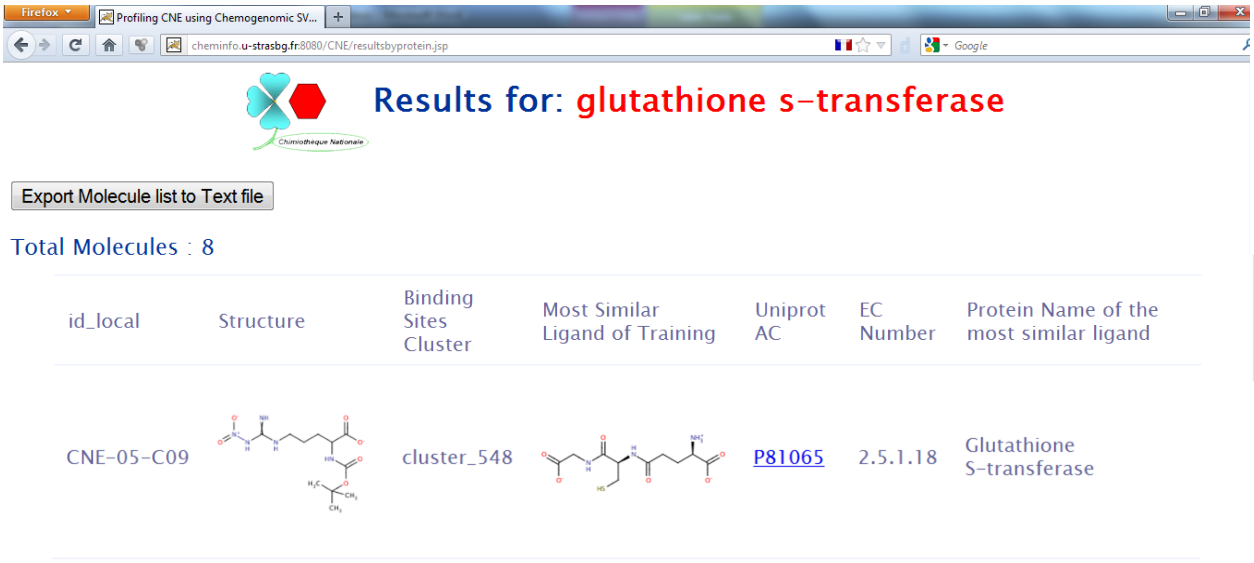


Figure 14 : Page des résultats des ligands potentiels de la cible "Glutathione S-transferase". Les structures des molécules profilées sont affichées ainsi que le ligand du jeu d'entrainement le plus similaire (Tanimoto ECFP_4) et également un lien vers l'annotation de la protéine dans la base Uniprot.

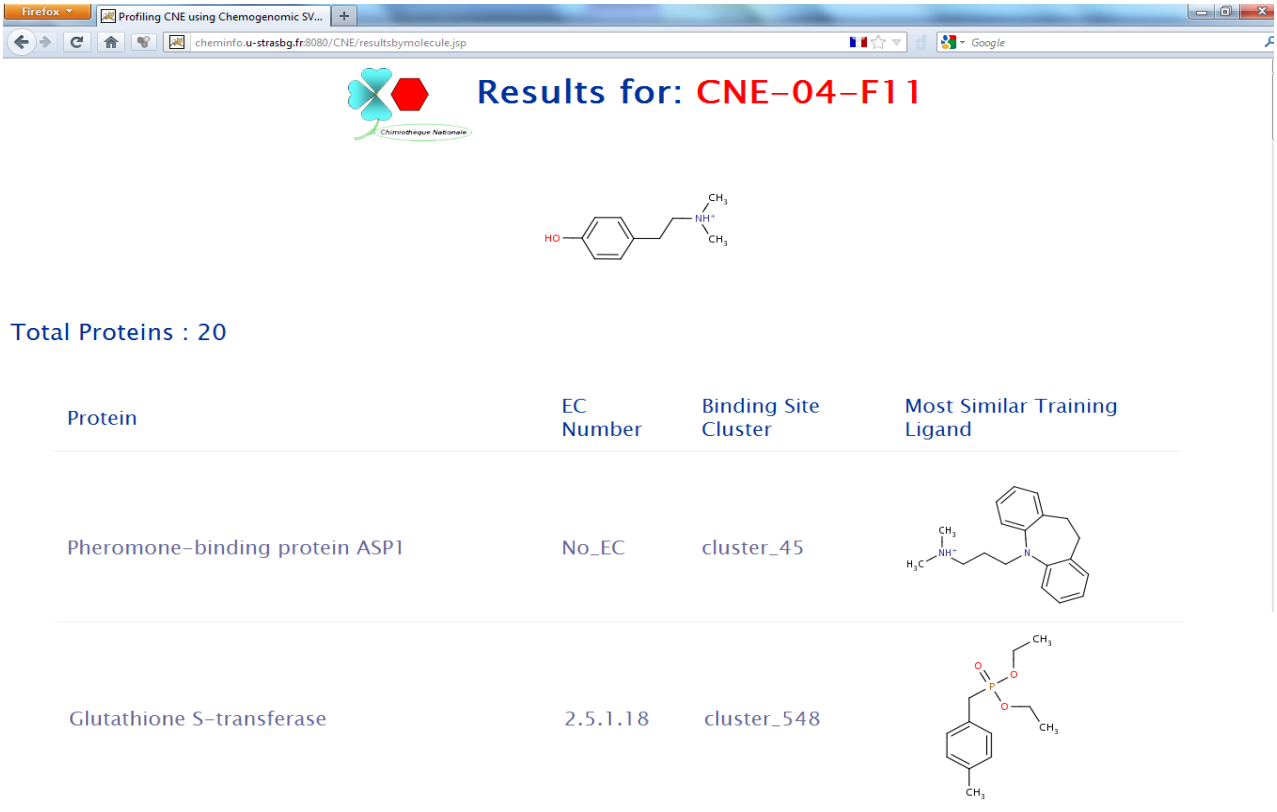


Figure 15 : Page de résultats des cibles potentielles pour la molécule "CNE-04-F11". Un tableau contenant les informations sur la protéine ainsi que son ligand est affiché.

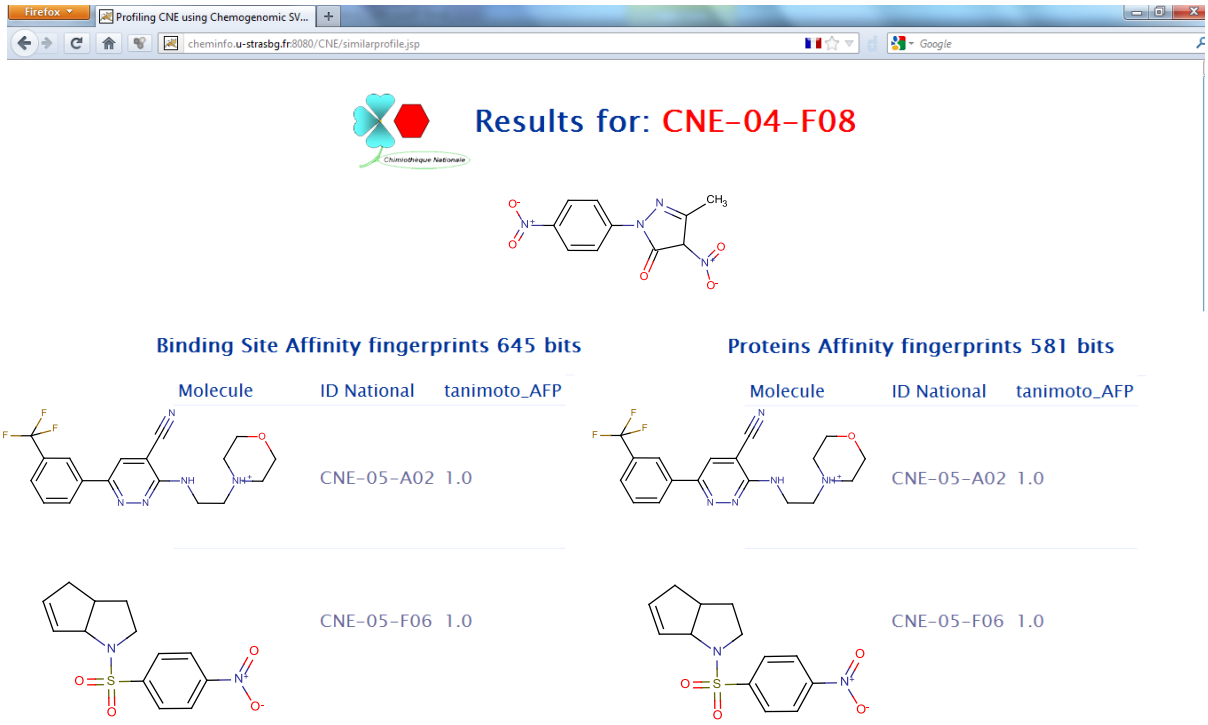
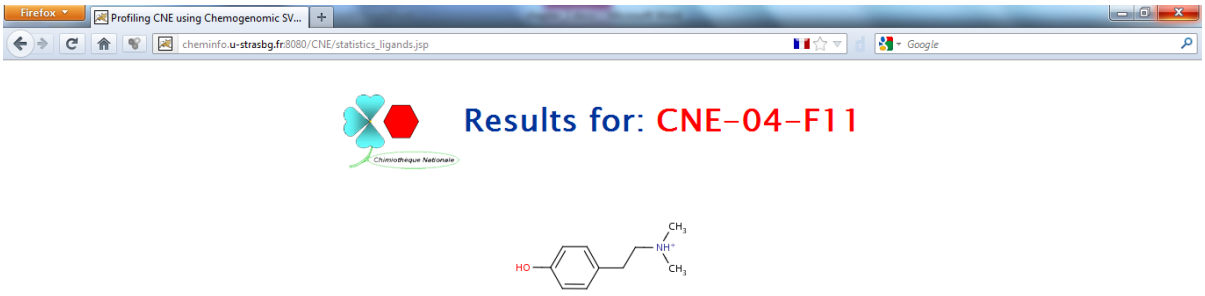


Figure 16 : Page de résultats affichant les molécules de profil biologique similaire selon les deux empreintes d'affinités, celle basée sur l'association avec les sites de liaisons et celle basée sur l'association avec les protéines.



molecule	nuclear receptors	GPCR	Ion channels	Enzymes	No Enzymes	Oxodoreductases	Transferases	Hydrolases	Isomerases	Ligases	Lyases
CNE-04-F11	0	0	0	14	6	3	4	4	2	0	1

Figure 17 : Page de résultats des cibles potentielles pour la molécule "CNE-04-F11" classé par familles protéiques.

6. Références

Accelrys (2010). "Pipeline Pilot, version 7.5; Accelrys Software Inc.: San Diego, CA."

Accelrys (2011). "http://accelrys.com/products/databases/bioactivity/mddr.html (accessed July 12, 2012).".

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

BioinfoDB. (2011). "http://bioinfo-pharma.u-strasbg.fr/bioinfo (accessed July 12, 2012).".

Block, P., C. A. Sotriffer, I. Dramburg and G. Klebe (2006). "AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB." Nucleic Acids Research **34**(Database issue): D522-526.

Bock, J. R. and D. A. Gough (2005). "Virtual screen for ligands of orphan G protein-coupled receptors." J Chem Inf Model **45**(5): 1402-1414.

ChemAxon (2010). "Standardizer, version 5.5.0.1; ChemAxon Kft.: Budapest, Hungary.".

Geppert, H., J. Humrich, D. Stumpfe, T. Gartner and J. Bajorath (2009). "Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors." J Chem Inf Model **49**(4): 767-779.

Hert, J., P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer (2004). "Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures." Org Biomol Chem **2**(22): 3256-3266.

Jacob, L., B. Hoffmann, V. Stoven and J. P. Vert (2008). "Virtual screening of GPCRs: an in silico chemogenomics approach." Bmc Bioinformatics **9**: 363.

Jagarlapudi, S. A. and K. V. Kishan (2009). "Database systems for knowledge-based discovery." Methods Mol Biol **575**: 159-172.

Keiser, M. J., B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet (2007). "Relating protein pharmacology by ligand chemistry." Nature Biotechnology **25**(2): 197-206.

Kellenberger, E., P. Muller, C. Schalon, G. Bret, N. Foata and D. Rognan (2006). "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank." J Chem Inf Model **46**(2): 717-727.

Klabunde, T. (2007). "Chemogenomic approaches to drug discovery: similar receptors bind similar ligands." Br J Pharmacol **152**(1): 5-7.

Kuhn, M., D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer and P. Bork (2010). "STITCH 2: an interaction network database for small molecules and proteins." Nucleic Acids Research **38**(Database issue): D552-556.

Lapins, M. and J. E. Wikberg (2010). "Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques." Bmc Bioinformatics **11**: 339.

Leslie, C., E. Eskin and W. S. Noble (2002). "The spectrum kernel: a string kernel for SVM protein classification." Pac Symp Biocomput: 564-575.

Liu, T., Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic Acids Research **35**(Database issue): D198-201.

Mahe, P., L. Ralaivola, V. Stoven and J. P. Vert (2006). "The pharmacophore kernel for virtual screening with support vector machines." J Chem Inf Model **46**(5): 2003-2014.

Meslamani, J., D. Rognan and E. Kellenberger (2011). "sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins." Bioinformatics **27**(9): 1324-1326.

Nagamine, N., T. Shirakawa, Y. Minato, K. Torii, H. Kobayashi, M. Imoto and Y. Sakakibara (2009). "Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening." Plos Computational Biology **5**(6): e1000397.

Ning, X., H. Rangwala and G. Karypis (2009). "Multi-assay-based structure-activity relationship models: improving structure-activity relationship models by incorporating activity information from related targets." J Chem Inf Model **49**(11): 2444-2456.

OpenEye (2010). "Filter, version 2.1.1; OpenEye Scientific Software: Santa Fe, NM.".

OpenEye (2010). "Omega, version 2.4.3; OpenEye Scientific Software: Santa Fe, NM."

OpenEye (2010). "ROCS, version 3.0.0; OpenEye Scientific Software: Santa Fe, NM.".

Overington, J. (2009). "ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr." J Comput Aided Mol Des **23**(4): 195-198.

Rogers, D. and M. Hahn (2010). "Extended-connectivity fingerprints." J Chem Inf Model **50**(5): 742-754.

Rognan, D. (2007). "Chemogenomic approaches to rational drug design." Br J Pharmacol **152**(1): 38-52.

Roth, B. L., E. Lopez, S. Patel and W. K. Kroeze (2000). "The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches?" Neuroscientist **6**(4): 252-262.

Roth, B. L., D. J. Sheffler and W. K. Kroeze (2004). "Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia." Nat Rev Drug Discov **3**(4): 353-359.

Shen, M., A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn and A. Tropsha (2002). "Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods." Journal of Medicinal Chemistry **45**(13): 2811-2823.

Strombergsson, H., P. Daniluk, A. Kryshtafovych, K. Fidelis, J. E. Wikberg, G. J. Kleywegt and T. R. Hvidsten (2008). "Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space." J Chem Inf Model **48**(11): 2278-2288.

Strömbergsson, H., M. Lapins, G. J. Kleywegt and J. E. S. Wikberg (2010). "Towards Proteome–Wide Interaction Models Using the Proteochemometrics Approach." Molecular Informatics **29**(6-7): 499-508.

SVMlight (2010). "SVMlight, version 6.02, http://svmlight.joachims.org/ (accessed July 12, 2012).".

Turner, D. B., S. M. Tyrrell and P. Willett (1997). "Rapid Quantification of Molecular Diversity for Selective Database Acquisition." J Chem Inf Comput Sci **37**(1): 18-22.

UniProt, C. (2010). "The Universal Protein Resource (UniProt) in 2010." Nucleic Acids Research **38**(Database issue): D142-148.

van Westen, G. J. P., J. K. Wegner, A. P. Ijzerman, H. W. T. van Vlijmen and A. Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets." Medchemcomm **2**(1): 16-30.

Vert, J. P. and L. Jacob (2008). "Machine learning for in silico virtual screening and chemical genomics: new strategies." Comb Chem High Throughput Screen **11**(8): 677-685.

Wang, R., X. Fang, Y. Lu and S. Wang (2004). "The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures." Journal of Medicinal Chemistry **47**(12): 2977-2980.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant (2009). "PubChem: a public information system for analyzing bioactivities of small molecules." Nucleic Acids Research **37**(Web Server issue): W623-633.

Wassermann, A. M., H. Geppert and J. Bajorath (2009). "Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects." J Chem Inf Model **49**(10): 2155-2167.

Weill, N. and D. Rognan (2009). "Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands." J Chem Inf Model **49**(4): 1049-1062.

Weill, N. and D. Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites." J Chem Inf Model **50**(1): 123-135.

Wishart, D. S., C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali (2008). "DrugBank: a knowledgebase for drugs, drug actions and drug targets." Nucleic Acids Research **36**(Database issue): D901-906.

Yeturu, K. and N. Chandra (2008). "PocketMatch: a new algorithm to compare binding sites in protein structures." Bmc Bioinformatics **9**: 543.

3.8. Annexes

Supplementary Table 1: Protein-ligand binding site annotation

http://pubs.acs.org/doi/suppl/10.1021/ci200166t/suppl_file/ci200166t_si_001.pdf

Supplementary Table 2: Statistics of the cross-validated local models on the training set (prediction by PDB entry)

	KTL1				KTL2				KTL3			
Cluster	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**
1	95	100	100	0.97	80	100	100	0.88	95	100	100	0.97
5	23.53	98.57	99.9	0.38	56.08	99.31	99.9	0.72	83.92	98.21	99.61	0.9
11	58	100	100	0.73	60	100	100	0.75	74	100	100	0.84
12	15	100	100	0.26	30	100	100	0.46	55	100	100	0.7
13	30	60	100	0.4	50	80	100	0.6	70	80	100	0.73
14	47.62	100	100	0.63	38.09	100	100	0.54	47.62	100	100	0.63
16	48.57	100	100	0.64	60	100	100	0.74	88.57	100	100	0.93
32	0	0	100	0	46.67	100	100	0.62	73.33	100	100	0.82
39	89.17	99.2	99.79	0.94	91.67	100	100	0.96	92.5	100	100	0.96
43	60	100	100	0.74	60	100	100	0.74	60	100	100	0.72
44	73.33	100	100	0.82	80	100	100	0.88	80	100	100	0.88
45	25	60	100	0.35	25	60	100	0.35	28.33	60	100	0.37
50	86.4	100	100	0.93	90.4	100	100	0.95	92.8	100	100	0.96
57	42.61	97.98	99.78	0.59	58.84	97.2	99.57	0.73	83.77	92.16	98.2	0.88
58	25	80	100	0.37	30	80	100	0.43	70	80	100	0.74
61	10	20	100	0.13	30	60	100	0.4	50	80	100	0.6
68	96.67	100	100	0.98	96.67	100	100	0.98	96.67	100	100	0.98
71	100	100	100	1	100	100	100	1	100	100	100	1
75	100	100	100	1	100	100	100	1	100	100	100	1
81	10	40	100	0.16	20	60	100	0.29	75	100	100	0.85
88	50	80	100	0.6	60	100	100	0.74	60	100	100	0.74
90	40	100	100	0.55	54	100	100	0.69	80	97.14	99.51	0.88
92	12	100	100	0.21	18.67	100	100	0.31	62.67	98.18	99.67	0.76

95	100	100	100	1	90	100	100	0.93	100	100	100	1
97	84.49	100	100	0.91	85.65	100	100	0.92	91.98	100	100	0.96
98	30	60	100	0.4	30	60	100	0.4	60	80	100	0.67
105	82	100	100	0.89	84	100	100	0.91	88	100	100	0.93
111	53.33	60	100	0.56	53.33	60	100	0.56	66.67	80	100	0.7
114	45	80	100	0.56	45	80	100	0.56	80	100	100	0.88
122	50	80	100	0.6	50	80	100	0.6	70	100	100	0.8
126	96	100	100	0.98	96	100	100	0.98	96	100	100	0.98
134	25	80	100	0.37	15	60	100	0.24	60	100	100	0.72
139	66	100	100	0.79	66	100	100	0.79	66	100	100	0.79
144	84	100	100	0.91	84	100	100	0.91	88	100	100	0.93
145	100	100	100	1	100	100	100	1	80	100	100	0.86
147	80	100	100	0.88	73.33	100	100	0.82	86.67	100	100	0.92
148	13.33	40	100	0.2	13.33	40	100	0.2	73.33	100	100	0.8
149	85	100	100	0.91	85	100	100	0.91	90	100	100	0.94
173	60	100	100	0.72	60	100	100	0.72	93.33	100	100	0.96
174	50	100	100	0.67	50	100	100	0.67	50	100	100	0.67
201	73.34	100	100	0.84	66.67	100	100	0.8	80	100	100	0.88
204	57.14	100	100	0.72	62.86	100	100	0.75	74.28	100	100	0.85
212	70	100	100	0.81	78	100	100	0.86	100	100	100	1
225	80	100	100	0.89	80	100	100	0.89	86	100	100	0.93
230	40	80	100	0.52	40	80	100	0.52	66.67	100	100	0.78
235	60	80	100	0.67	60	80	100	0.67	60	80	100	0.67
236	60	80	100	0.69	65	80	100	0.72	80	100	100	0.88
241	73.85	100	100	0.84	73.85	100	100	0.84	81.54	100	100	0.89
253	83.03	100	100	0.9	84.7	100	100	0.91	89.7	100	100	0.95
254	88.69	98	99.52	0.93	88.69	98	99.52	0.93	94.14	98.33	99.52	0.96

259	100	100	100	1	100	100	100	1	90	100	100	0.93
261	70	80	100	0.73	70	80	100	0.73	100	100	100	1
277	50	80	100	0.6	50	80	100	0.6	60	100	100	0.74
280	60	100	100	0.74	40	60	100	0.47	70	100	100	0.8
321	100	100	100	1	100	100	100	1	93.33	100	100	0.96
336	40	80	100	0.5	40	80	100	0.5	60	80	100	0.66
341	73.34	100	100	0.84	73.34	100	100	0.84	73.34	100	100	0.84
361	60	100	100	0.74	60	100	100	0.74	57.14	100	100	0.72
375	100	100	100	1	100	100	100	1	100	100	100	1
383	0	0	100	0	10	20	100	0.13	50	60	100	0.53
391	80	100	100	0.87	90	100	100	0.93	90	100	100	0.93
394	58	100	100	0.73	26	80	100	0.38	74	100	100	0.84
403	100	100	100	1	100	100	100	1	100	100	100	1
409	66	100	100	0.78	66	100	100	0.78	70	100	100	0.82
415	68	100	100	0.78	68	100	100	0.78	89.33	100	100	0.94
416	85	100	100	0.92	75	100	100	0.85	85	100	100	0.92
422	40	80	100	0.52	53.33	100	100	0.68	73.33	100	100	0.82
472	100	100	100	1	100	100	100	1	100	100	100	1
476	80	100	100	0.86	80	100	100	0.86	83.33	100	100	0.89
490	51.11	96.67	99.44	0.65	51.11	96.67	99.44	0.65	53.33	93.81	98.9	0.66
518	80	80	100	0.8	80	80	100	0.8	73.33	80	100	0.76
548	0	0	100	0	0	0	100	0	0	0	100	0
570	66.67	80	100	0.72	66.67	80	100	0.72	66.67	80	100	0.72
573	100	100	100	1	70	100	100	0.8	100	100	100	1
634	90	100	100	0.93	90	100	100	0.93	90	100	100	0.93
642	60	100	100	0.74	70	100	100	0.8	70	100	100	0.8
645	20	40	100	0.27	20	40	100	0.27	20	40	100	0.27

650	100	100	100	1	100	100	100	1	100	100	100	1
687	56.67	100	100	0.71	50	100	100	0.65	56.67	100	100	0.71
704	50	80	100	0.6	50	80	100	0.6	50	80	100	0.6
739	50	100	100	0.67	50	100	100	0.67	80	100	100	0.87
754	86.67	100	100	0.9	80	100	100	0.86	80	100	100	0.86
779	33.33	60	100	0.4	40	80	100	0.5	66.67	100	100	0.78
805	40	100	100	0.56	53.33	100	100	0.68	66.67	100	100	0.78
821	52	100	100	0.67	44	100	100	0.59	52	100	100	0.67
912	20	80	100	0.32	15	60	100	0.24	35	100	100	0.51
931	86.67	100	100	0.92	86.67	100	100	0.92	86.67	100	100	0.92

Supplementary Table 3: Statistics of the local models on the test set (prediction by PDB entry)

	KTL1				KTL2				KTL3			
Cluster	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**
277	72.17	99.87	99.91	0.84	72.08	99.87	99.91	0.84	73.4	99.87	99.91	0.85
687	100	100	100	1	98.87	100	100	0.99	28.59	100	100	0.44
476	94.55	100	100	0.97	93.82	100	100	0.97	72.44	100	100	0.84
754	95.37	100	100	0.98	79.53	100	100	0.89	81.12	100	100	0.9
1	95.79	99.79	99.8	0.98	55.74	99.92	99.95	0.72	70.68	99.93	99.94	0.83
416	98.39	99.97	99.97	0.99	61.71	99.94	99.97	0.76	48.63	100	100	0.65
821	16.17	99.96	99.99	0.28	9.65	99.94	99.99	0.18	15.47	99.94	99.99	0.27
472	100	100	100	1	100	100	100	1	71.19	100	100	0.83
490	0	0	100	0	0	0	100	0	0	0	100	0
241	87.37	99.98	99.98	0.93	85.5	99.99	99.99	0.92	54.1	99.92	99.96	0.7
642	0	0	100	0	63.64	100	100	0.78	81.82	100	100	0.9

13	48.44	100	100	0.65	48.27	100	100	0.65	44.98	100	100	0.62
122	29.07	100	100	0.45	21.14	100	100	0.35	25.32	100	100	0.4
391	100	100	100	1	91.67	100	100	0.96	75	100	100	0.86
88	90	100	100	0.95	90	100	100	0.95	88.89	100	100	0.94
61	1.39	100	100	0.03	0.56	100	100	0.01	8.64	100	100	0.16
280	0	0	100	0	0	0	100	0	0	0	100	0
98	16.67	100	100	0.29	0.28	100	100	0.01	5.93	100	100	0.11
739	60	100	100	0.75	55	100	100	0.71	36.67	100	100	0.54
230	15	100	100	0.26	11.79	100	100	0.21	7.5	100	100	0.14
570	0	0	100	0	0	0	100	0	10.42	100	100	0.19
148	13.33	100	100	0.24	0	0	100	0	13.33	100	100	0.24
548	0.64	100	100	0.01	0.64	100	100	0.01	0.64	100	100	0.01
422	1.06	100	100	0.02	1.06	100	100	0.02	13.83	100	100	0.24
201	53.08	99.95	99.97	0.69	22.42	99.89	99.97	0.37	32.07	99.96	99.98	0.49
336	92.59	100	100	0.96	92.59	100	100	0.96	92.59	100	100	0.96
32	2.63	100	100	0.05	8.95	100	100	0.16	32.11	100	100	0.49
805	0	0	100	0	0	0	100	0	0	0	100	0
149	0	0	100	0	0	0	100	0	100	100	100	1
236	48.64	100	100	0.65	20.59	100	100	0.34	40.46	100	100	0.58
114	54.44	100	100	0.71	54.44	100	100	0.71	43.87	100	100	0.61
912	24.45	100	100	0.39	24.45	100	100	0.39	24.79	99.66	99.92	0.4
134	0	0	100	0	0	0	100	0	9.87	100	100	0.18
58	40.95	100	100	0.58	44.97	100	100	0.62	38.06	100	100	0.55
43	38	99.78	99.92	0.55	39.87	99.84	99.93	0.57	25.75	99.78	99.91	0.41
81	66.67	100	100	0.8	66.67	100	100	0.8	58.33	100	100	0.74

415	14.71	100	100	0.26	13.87	100	100	0.24	14.29	100	100	0.25
126	100	100	100	1	100	100	100	1	91.3	100	100	0.95
394	62.62	100	100	0.77	35.92	100	100	0.53	48.32	100	100	0.65
212	39.59	100	100	0.57	15.61	100	100	0.27	39.18	100	100	0.56
139	63.45	99.41	99.63	0.77	58.53	99.41	99.65	0.74	49.34	99.43	99.63	0.66
144	86.52	100	100	0.93	86.93	100	100	0.93	68.08	100	100	0.81
45	1.79	100	100	0.04	1.3	100	100	0.03	1.42	97.22	99.93	0.03
14	0.11	100	100	0	0.11	100	100	0	0.62	100	100	0.01
16	69.85	99.91	99.94	0.82	81.29	99.92	99.94	0.9	80.25	99.85	99.86	0.89
204	54.07	99.9	99.95	0.7	77.35	99.84	99.87	0.87	68.6	99.75	99.79	0.81
12	7.32	100	100	0.14	94.08	100	100	0.97	42.86	100	100	0.6
225	0	0	100	0	0	0	100	0	54.29	100	100	0.7
105	93.88	100	100	0.97	95.92	100	100	0.98	92.24	100	100	0.96
90	0.09	100	100	0	16.7	100	100	0.29	85.84	99.79	99.82	0.92
409	0	0	100	0	0	0	100	0	0	0	100	0
11	0	0	100	0	0.08	100	100	0	7.23	100	100	0.13
254	85.13	99.83	99.86	0.92	88.5	99.83	99.85	0.94	78.63	99.78	99.79	0.88
253	0	0	100	0	100	100	100	1	100	100	100	1
68	100	99.59	99.58	1	100	99.59	99.58	1	92.92	99.55	99.58	0.96
92	0.13	94.12	99.99	0	0.13	94.12	99.99	0	1.48	92.65	99.86	0.03
50	80.84	100	100	0.89	83.89	99.98	99.98	0.91	80.88	99.99	99.99	0.89
97	77.87	99.91	99.93	0.88	77.77	99.9	99.92	0.87	86.02	99.87	99.88	0.92
5	74.26	100	100	0.85	37.89	100	100	0.55	31.81	99.32	99.71	0.48
57	28.54	99.33	99.81	0.44	46.57	99.26	99.65	0.63	62.15	98.74	99.03	0.76

Supplementary Table 4: Statistics of the local models on the test set (prediction by target binding site)

	KTL1				KTL2				KTL3			
Cluster	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**
277	72.17	100	100	0.84	72.17	100	100	0.84	76.42	100	100	0.87
687	100	100	100	1	100	100	100	1	100	100	100	1
476	94.55	100	100	0.97	94.55	100	100	0.97	82.1	100	100	0.9
754	95.37	100	100	0.98	95.37	100	100	0.98	88.89	100	100	0.94
1	95.79	100	100	0.98	93.1	100	100	0.96	77.39	100	100	0.87
416	98.39	100	100	0.99	99.2	100	100	1	74.7	100	100	0.86
821	16.17	100	100	0.28	17.22	100	100	0.29	38.52	100	100	0.56
472	100	100	100	1	100	100	100	1	100	100	100	1
490	0	0	100	0	0	0	100	0	0	0	100	0
241	87.37	100	100	0.93	87.89	100	100	0.94	94.59	99.73	99.74	0.97
642	0	0	100	0	25	100	100	0.4	75	100	100	0.86
13	48.44	100	100	0.65	49.33	100	100	0.66	66.67	100	100	0.8
122	29.07	100	100	0.45	29.07	100	100	0.45	48.84	100	100	0.66
391	100	100	100	1	100	100	100	1	100	100	100	1
88	90	100	100	0.95	90	100	100	0.95	90	100	100	0.95
61	14.29	100	100	0.25	14.29	100	100	0.25	28.57	100	100	0.44
280	0	0	100	0	0	0	100	0	0	0	100	0
98	16.67	100	100	0.29	16.67	100	100	0.29	16.67	100	100	0.29
739	60	100	100	0.75	60	100	100	0.75	60	100	100	0.75
230	15	100	100	0.26	15	100	100	0.26	20	100	100	0.33
570	0	0	100	0	0	0	100	0	25	100	100	0.4

148	13.33	100	100	0.24	0	0	100	0	13.33	100	100	0.24
548	1.76	100	100	0.03	1.76	100	100	0.03	1.76	100	100	0.03
422	1.92	100	100	0.04	1.92	100	100	0.04	23.08	100	100	0.38
201	53.08	100	100	0.69	54.03	100	100	0.7	60.19	100	100	0.75
336	92.59	100	100	0.96	92.59	100	100	0.96	92.59	100	100	0.96
32	1.3	100	100	0.03	11.69	100	100	0.21	38.96	100	100	0.56
805	0	0	100	0	0	0	100	0	0	0	100	0
149	0	0	100	0	0	0	100	0	100	100	100	1
236	43.55	100	100	0.61	43.55	100	100	0.61	70.97	100	100	0.83
114	53.54	100	100	0.7	53.54	100	100	0.7	62.83	100	100	0.77
912	11.35	100	100	0.2	11.35	100	100	0.2	14.18	100	100	0.25
134	0	0	100	0	0	0	100	0	11.11	100	100	0.2
58	31.1	100	100	0.47	35.83	100	100	0.53	59.45	100	100	0.75
43	12.86	100	100	0.23	25.44	100	100	0.41	36.64	99.5	99.82	0.54
81	50	100	100	0.67	50	100	100	0.67	75	100	100	0.86
415	16	100	100	0.28	24	100	100	0.39	32	100	100	0.48
126	100	100	100	1	100	100	100	1	92.86	100	100	0.96
394	9.69	100	100	0.18	16.62	100	100	0.29	58.12	100	100	0.74
212	44.83	100	100	0.62	62.07	100	100	0.77	62.07	100	100	0.77
139	63.45	100	100	0.78	64.26	100	100	0.78	74.3	99.46	99.6	0.85
144	83.47	100	100	0.91	84.14	100	100	0.91	82.12	100	100	0.9
45	2.68	100	100	0.05	2.68	100	100	0.05	7.14	100	100	0.13
14	0.39	100	100	0.01	0.39	100	100	0.01	3.94	100	100	0.08
16	51.73	100	100	0.68	70.23	100	100	0.83	86.42	100	100	0.93
204	46.12	99.56	99.8	0.63	78.71	99.62	99.7	0.88	77.19	99.48	99.6	0.87

12	7.32	100	100	0.14	95.12	100	100	0.97	43.9	100	100	0.61
225	0	0	100	0	0	0	100	0	54.29	100	100	0.7
105	93.88	100	100	0.97	95.92	100	100	0.98	97.96	100	100	0.99
90	0.12	100	100	0	7.42	100	100	0.14	84.8	99.73	99.77	0.92
409	0	0	100	0	0	0	100	0	0	0	100	0
11	0	0	100	0	1.85	100	100	0.04	25.93	100	100	0.41
254	83.74	99.86	99.88	0.91	88.84	99.87	99.88	0.94	84.33	99.86	99.88	0.91
253	0	0	100	0	100	100	100	1	100	100	100	1
68	100	100	100	1	100	100	100	1	100	100	100	1
92	0.66	100	100	0.01	0.66	100	100	0.01	9.57	100	100	0.17
50	69.03	100	100	0.82	79.55	100	100	0.89	89.77	100	100	0.95
97	76.26	99.8	99.85	0.86	77.17	99.8	99.85	0.87	96.65	100	100	0.98
5	44.16	100	100	0.61	62.34	100	100	0.77	62.34	100	100	0.77
57	28.41	99.76	99.93	0.44	54.98	99.13	99.52	0.71	78.59	98.71	98.97	0.88

Supplementary Table 5: Statistics of the global model on the test sets (prediction by PDB entry)

	KTL1				KTL2				KTL3			
Cluster	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**
105	91.2	100	100	0.95	1.87	100	100	0.04	31.84	71.94	87.21	0.44
114	40.83	100	100	0.58	19.01	100	100	0.32	33.8	81.11	89.2	0.48
11	2.33	100	100	0.05	8.86	100	100	0.16	31.03	74.58	80.79	0.44
122	18.87	100	100	0.32	0	0	99.92	0	19.08	96.27	98.84	0.32
126	91.3	100	100	0.95	91.3	100	100	0.95	91.3	99.83	99.83	0.95
12	3.14	100	100	0.06	42.86	100	100	0.6	49.13	92.16	90.24	0.64

134	0.08	100	100	0	16.89	100	100	0.29	30.27	86.6	76.27	0.45
139	48.71	99.37	99.6	0.65	3.78	99.3	99.97	0.07	33.98	88.09	94.01	0.49
13	38.4	100	100	0.55	1.78	100	100	0.03	42.04	98.95	99.44	0.59
144	66.2	100	100	0.8	17.33	99.92	99.98	0.3	62.72	93.46	94.82	0.75
148	13.33	100	100	0.24	0	0	100	0	0	0	90	0
149	0	0	100	0	100	100	100	1	100	100	100	1
14	3.2	100	100	0.06	2.75	100	100	0.05	22.5	78.02	89.31	0.35
16	60.8	99.9	99.93	0.76	53.78	99.77	99.86	0.7	12.49	65.06	92.18	0.21
1	72.86	99.75	99.76	0.84	56.75	99.12	99.33	0.72	58.5	97.03	97.64	0.73
201	22.42	99.94	99.98	0.37	44.32	99.88	99.92	0.61	75.34	76.26	65.29	0.76
204	42.67	99.9	99.95	0.6	0.83	99.12	99.99	0.02	40.62	83.99	90.5	0.55
212	33.16	100	100	0.5	46.33	100	100	0.63	67.65	85.66	83.31	0.76
225	0	0	100	0	2.86	100	100	0.06	0	0	97.14	0
230	8.57	100	100	0.16	7.14	95.24	99.5	0.13	42.5	79.87	85	0.55
236	17.09	100	100	0.29	15.63	99.07	99.83	0.27	38.75	93.43	96.92	0.55
241	87.37	99.99	99.99	0.93	0	0	99.97	0	48.62	93.81	96.79	0.64
253	0	0	100	0	0	0	100	0	16.67	50	83.33	0.25
254	71.96	99.85	99.88	0.84	7.73	100	100	0.14	73.95	76.3	72.75	0.75
277	72.17	99.87	99.91	0.84	3.58	100	100	0.07	38.87	97.17	98.87	0.56
280	0	0	100	0	5	100	100	0.1	50	100	100	0.67
32	0.53	100	100	0.01	37.89	100	100	0.55	24.21	69.7	80.2	0.36
336	92.59	100	100	0.96	3.7	100	100	0.07	54.81	72.55	79.26	0.62
391	75	100	100	0.86	75	100	100	0.86	75	100	100	0.86
394	30.46	100	100	0.47	12.86	100	100	0.23	49.63	89.23	89.64	0.64
409	0	0	100	0	0	0	100	0	0	0	100	0

415	10.08	100	100	0.18	18.91	100	100	0.32	15.13	78.26	93.87	0.25
416	57.76	99.96	99.97	0.73	49.8	99.91	99.94	0.66	42.39	95.28	97.17	0.59
422	1.06	100	100	0.02	8.51	100	100	0.16	27.66	83.87	92.42	0.42
43	24.77	99.85	99.94	0.4	1.07	95	99.91	0.02	26.32	85.16	92.75	0.4
45	1.1	100	100	0.02	6.09	100	100	0.11	27.93	84.42	91.19	0.42
472	70.62	100	100	0.83	63.28	100	100	0.78	85.88	98.7	98.81	0.92
476	77.87	100	100	0.88	5.95	100	100	0.11	56.9	99.64	99.75	0.72
490	0	0	100	0	44.23	100	100	0.61	44.23	93.88	96.74	0.6
50	75.23	99.98	99.99	0.86	11.77	99.83	99.98	0.21	58.1	88.02	91.54	0.7
548	2.41	100	100	0.05	14.99	99.16	99.83	0.26	26.56	76	89.02	0.39
570	0	0	100	0	0	0	100	0	29.17	66.67	78.12	0.41
57	23.23	99.25	99.78	0.38	2.87	99.69	99.99	0.06	31.79	79.42	89.86	0.45
58	40.95	100	100	0.58	0.52	100	100	0.01	12.31	90.48	98.48	0.22
5	22.34	100	100	0.37	17.14	99.58	99.9	0.29	45.7	83.82	88.38	0.59
61	0.84	100	100	0.02	18.94	100	100	0.32	15.32	100	100	0.27
642	0	0	100	0	90.91	100	100	0.95	100	100	100	1
687	27.91	100	100	0.44	92.09	99.88	99.88	0.96	91.3	99.26	99.29	0.95
68	94.69	99.56	99.58	0.97	11.67	100	100	0.21	79.9	95.04	95.77	0.87
739	45	100	100	0.62	0	0	100	0	18.33	68.75	88.89	0.29
754	84.77	100	100	0.92	63.37	100	100	0.78	78.4	98.2	98.38	0.87
805	0	0	100	0	0	0	100	0	0	0	66.67	0
81	50	100	100	0.67	83.33	100	100	0.91	83.33	90.91	91.67	0.87
821	12.63	100	100	0.22	0.32	93.04	99.97	0.01	15.6	82.56	96.32	0.26
88	90	100	100	0.95	33.33	100	100	0.5	90	97.59	97.78	0.94
90	0.09	100	100	0	5.63	96.88	99.82	0.11	36.21	95.91	98.46	0.53

912	24.45	100	100	0.39	0.08	100	100	0	44.65	67.96	78.95	0.54
92	10.19	99.92	99.99	0.18	1.56	93.87	99.88	0.03	35.59	68.2	80.52	0.47
97	75.76	99.91	99.93	0.86	11.68	99.65	99.95	0.21	54.7	94.12	96.49	0.69
98	0.28	100	100	0.01	9.32	100	100	0.17	28.81	91.07	84.85	0.44

Supplementary Table 6: Statistics of the global model on the test sets (prediction by target binding site)

	KTL1				KTL2				KTL3			
Cluster	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**	**Recall**	**Precision**	**Specificity**	**Fmeasure**
105	97.96	100	100	0.99	4.08	100	100	0.08	100	92.45	91.84	0.96
114	58.85	100	100	0.74	39.82	100	100	0.57	81.42	86.79	87.61	0.84
11	7.41	100	100	0.14	27.78	100	100	0.43	87.04	87.04	87.04	0.87
122	38.37	100	100	0.55	0	0	100	0	56.4	100	100	0.72
126	92.86	100	100	0.96	92.86	100	100	0.96	92.86	100	100	0.96
12	7.32	100	100	0.14	43.9	100	100	0.61	70.73	82.86	85.37	0.76
134	2.78	100	100	0.05	13.89	100	100	0.24	61.11	68.75	72.22	0.65
139	71.08	99.44	99.6	0.83	9.64	100	100	0.18	96.39	94.12	93.98	0.95
13	57.33	100	100	0.73	5.33	100	100	0.1	70.67	100	100	0.83
144	81.18	100	100	0.9	33.2	99.6	99.87	0.5	92.34	94.63	94.76	0.93
148	13.33	100	100	0.24	0	0	100	0	0	0	86.67	0
149	0	0	100	0	100	100	100	1	100	100	100	1
14	5.51	100	100	0.1	8.27	100	100	0.15	66.93	86.29	89.37	0.75
16	72.83	100	100	0.84	70.81	99.59	99.71	0.83	46.82	83.94	91.04	0.6
1	77.78	99.51	99.62	0.87	73.95	97.47	98.08	0.84	85.44	95.71	96.17	0.9
201	57.35	100	100	0.73	66.82	100	100	0.8	100	72.76	62.56	0.84

204	62.76	100	100	0.77	2.02	100	100	0.04	91.42	91.33	91.32	0.91
212	62.07	100	100	0.77	65.52	100	100	0.79	100	82.86	79.31	0.91
225	0	0	100	0	2.86	100	100	0.06	0	0	97.14	0
230	15	100	100	0.26	20	100	100	0.33	100	80	75	0.89
236	59.68	100	100	0.75	40.32	100	100	0.57	100	100	100	1
241	87.37	100	100	0.93	0	0	99.74	0	100	96.52	96.39	0.98
253	0	0	100	0	0	0	100	0	16.67	50	83.33	0.25
254	82.31	99.86	99.88	0.9	23.92	100	100	0.39	99.94	80.54	75.85	0.89
277	72.17	100	100	0.84	4.72	100	100	0.09	56.13	95.2	97.17	0.71
280	0	0	100	0	9.09	100	100	0.17	27.27	100	100	0.43
32	1.3	100	100	0.03	42.86	100	100	0.6	38.96	68.18	81.82	0.5
336	92.59	100	100	0.96	3.7	100	100	0.07	100	84.38	81.48	0.92
391	100	100	100	1	100	100	100	1	100	100	100	1
394	39.01	100	100	0.56	18.46	100	100	0.31	63.09	85.46	89.27	0.73
409	0	0	100	0	0	0	100	0	0	0	100	0
415	32	100	100	0.48	36	100	100	0.53	60	88.24	92	0.71
416	74.7	100	100	0.86	69.88	100	100	0.82	88.76	95.67	95.98	0.92
422	1.92	100	100	0.04	13.46	100	100	0.24	32.69	80.95	92.31	0.47
43	30.58	100	100	0.47	3.31	97.3	99.91	0.06	52.53	88.54	93.2	0.66
45	3.57	100	100	0.07	17.86	100	100	0.3	84.82	91.35	91.96	0.88
472	100	100	100	1	100	100	100	1	100	100	100	1
476	82.1	100	100	0.9	13.23	100	100	0.23	82.1	100	100	0.9
490	0	0	100	0	50	100	100	0.67	50	100	100	0.67
50	84.66	100	100	0.92	20.17	100	100	0.34	94.89	92.52	92.33	0.94
548	2.82	100	100	0.05	16.2	100	100	0.28	32.39	75.41	89.44	0.45

570	0	0	100	0	0	0	100	0	68.75	73.33	75	0.71
57	41.8	99.51	99.79	0.59	7.07	100	100	0.13	89.64	90.51	90.6	0.9
58	57.48	100	100	0.73	0.79	100	100	0.02	28.74	93.59	98.03	0.44
5	53.25	100	100	0.69	32.47	100	100	0.49	84.42	89.04	89.61	0.87
61	28.57	100	100	0.44	28.57	100	100	0.44	71.43	100	100	0.83
642	0	0	100	0	75	100	100	0.86	100	100	100	1
687	100	100	100	1	100	100	100	1	100	100	100	1
68	100	100	100	1	40	100	100	0.57	100	100	100	1
739	60	100	100	0.75	0	0	100	0	100	83.33	80	0.91
754	88.89	100	100	0.94	84.26	100	100	0.91	90.74	98	98.15	0.94
805	0	0	100	0	0	0	100	0	0	0	0	0
81	50	100	100	0.67	75	100	100	0.86	100	100	100	1
821	23.65	100	100	0.38	1.39	100	100	0.03	93.57	96.42	96.52	0.95
88	90	100	100	0.95	40	100	100	0.57	100	100	100	1
90	0.12	100	100	0	7.19	96.88	99.77	0.13	43.27	96.63	98.49	0.6
912	11.35	100	100	0.2	0.71	100	100	0.01	97.16	85.09	82.98	0.91
92	18.48	100	100	0.31	6.6	100	100	0.12	73.93	77.24	78.22	0.76
97	91.93	100	100	0.96	22.68	100	100	0.37	98.78	96.43	96.35	0.98
98	16.67	100	100	0.29	33.33	100	100	0.5	100	75	66.67	0.86

Chapitre 4 :

Evaluation des pharmacophores d'interactions protéine-ligand à des fins de profilage

Ce chapitre a fait l'objet d'une publication :

Protein-Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling

Jamel Meslamani, Jiabo Li,[ϕ] Jon Sutter,[ϕ] Adrian Stevens,[†] Hugues-Olivier Bertrand,[ϫ] and Didier Rognan.

Journal of Chemical Information and Modelling, **2012**, 52, 943−955

ϕ Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, California 92121, United States
† Accelrys Ltd., 334 Cambridge Science Park, Cambridge CB4 OWN, England
ϫ Accelrys SARL, Parc Club Orsay Université, 20 Rue Jean Rostand, 91898 Orsay Cédex, France

1. Contexte

Les approches de criblage virtuel utilisant les pharmacophore d'interactions protéine-ligand (Leach *et al.* 2011) ont récemment eu beaucoup d'engouement notamment grâce à l'accroissement des structures cristallographiques et des différents algorithmes pour leur élucidation.

Leur efficience dans le cadre d'un profilage sur un grand nombre de cibles n'a cependant pas été mise à l'épreuve. Ceci peut s'expliquer par le fait qu'il est nécessaire d'effectuer plusieurs traitements des fichiers pdb des structures cristallographiques afin de pouvoir directement dériver un pharmacophore.

Avec nos collaborateurs d'Accelrys, nous avons implémenté et testé une méthode pour générer d'une manière automatique les pharmacophores d'interactions pour la construction d'une base de données de pharmacophores qui servira à tester l'approche dans le cadre d'un profilage.

Plusieurs approches existent pour pratiquer du profilage, il était donc intéressant de comparer les performances des pharmacophores avec les méthodes employées d'une manière classique à savoir les méthodes basées sur les ligands et l'arrimage moléculaire.

Nous validerons, dans ce chapitre, le concept de profilage à l'aide des pharmacophores et nous discuterons de validation expérimentale de quelques cibles secondaires identifiées grâce à cette méthode.

2. Introduction

Knowledge on protein−ligand binding data (affinity, structure) is increasing at an amazing pace thanks to public initiatives to homogenize data archival and mining (Ekins *et al.* 2010; Wang, Y. *et al.* 2011).

On the target side, the Protein Data Bank (Berman *et al.* 2000) stores 78 000 three-dimensional (3-D) structures of proteins and protein−ligand complexes out of which about 10 000 relate to druggable proteins and their ligands (Meslamani *et al.* 2011).

On the ligand side, ChEMBL (Gaulton *et al.* 2011) is a repository of more than five million bioactivity data gathered from literature and addressing one million ligands and 8 700 molecular targets.

Computational chemists have rapidly developed so-called chemogenomic methods (Rognan 2007) to mine this vast matrix of experimental data in order to predict novel interactions.

Whereas many virtual screening methods (Schneider 2010) (similarity search, pharmacophore mapping, protein−ligand docking) have proven useful to predict novel ligands for a single target, profiling a single ligand against a set of heterogeneous targets has long been neglected.

Scientific and economic pressure to design drugs with controlled selectivity profiles (Hopkins *et al.* 2006; Morphy 2010) as well as the recent boost of drug repurposing (Ekins *et al.* 2011) led to the development of in-silico ligand profiling methods (Rognan 2010) aimed at :

(i) predicting potential targets (and thus a mechanism of action) for orphan bioactive ligands (Muller *et al.* 2006)

(ii) identifying off-targets responsible for side effects and adverse reactions (Yang *et al.* 2011)

(iii) and proposing novel targets for existing drugs (Keiser *et al.* 2009).

From a conceptual point of view, three groups of methods can be used to predict novel protein−ligand interactions. At the simplest level of theory is the concept that similar ligands bind to similar targets. Estimating the similarity between a ligand of interest and target-annotated compounds is thus an easy way to predict novel target-ligand associations (Yang *et al.* 2011) and even within certain limits binding affinities (Vidal *et al.* 2010). Interestingly, 2-D similarity methods have recently been shown to be effective for identifying main targets, whereas 3-D similarity methods were better suited for proposing off-targets (Yera *et al.* 2011).

Ligand-centric profiling methods are however restricted to targets for which sufficient ligand information is available. For example, the Similarity Ensemble Approach (SEA) developed by Keiser et al. only applies to 246 targets annotated by more than 100 ligands (Keiser *et al.* 2009; Keiser *et al.* 2009).

A second group of methods relies on the concept that similar ligands bind to similar binding sites. Binding site similarity either at the sequence (Surgand *et al.* 2006) or at the structure level (Xie *et al.* 2008) can thus be used as a means to pair an existing ligand (of known binding site) to a novel target as successfully evidenced by several independent reports (Martin *et al.* 2007; Kinnings *et al.* 2009; Defranchi *et al.* 2010). Again, the method has inherited limitations as it is restrained to the few targets for which a 3-D structure is available (structure-based approach) or to a target subfamily in order to avoid binding site-based misalignments (sequence based approach).

At the highest level of theory is the last group of approaches focusing on protein-ligand complexes that can be described either as simple 1-D fingerprints (van Westen *et al.* 2011), protein-ligand-derived pharmacophores (Wolber *et al.* 2005), or protein-ligand docking poses (Yang *et al.* 2009). Identification of novel targets, accounting for main or secondary effects, has been reported in numerous reverse docking studies (Muller *et al.* 2006; Yang *et al.* 2009; Durrant *et al.* 2010; Li *et al.* 2011; Yang *et al.* 2011), despite notorious deficiencies of empirical scoring functions to rank order target-ligand complexes by increasing binding free energies (Ferrara *et al.* 2004; Enyedy *et al.* 2008).

Chemogenomic (or proteochemometric) approaches correctly predicting novel binary associations also begin to appear in the literature (Wang, F. *et al.* 2011). Surprisingly, pharmacophores have been widely used in many areas of computer-aided drug design (Leach *et al.* 2011) but rarely in target fishing applications. The idea to screen protein-ligand-derived pharmacophores in order to identify potential targets of bioactive ligands was applied by Langer et al. in a series of retrospective screening experiments focusing on small protein-ligand matrices (Steindl *et al.* 2006; Markt *et al.* 2007; Steindl *et al.* 2007).

A noticeable hurdle to pharmacophore based ligand profiling is the automation of relevant pharmacophore perception and generation protocols, mainly due to the difficulty to correctly assign atom types and bond orders from raw PDB files (Wolber *et al.* 2005).

Up to now, only two pharmacophore collections are available. The Inte:Ligand's Pharmacophore Database (http://www.inteligand.com/pharmdb/) is a private-owned repertoire of 2 500 manually assembled pharmacophore models covering 300 clinically relevant pharmacological targets.

The PharmTargetDB includes over 7 000 receptor-based pharmacophore models from 1 500 protein-ligand structures and can be screened via the PharmMapper server (http://59.78.96.61/pharmmapper/) (Liu *et al.* 2010). Only the Inte:Ligand collection has been successfully screened to identify novel targets (acetylcholinesterase, human rhinovirus coat protein, and cannabinoid receptor type 2) for secondary metabolites from the medicinal plant Ruta graveolens (Rollinger *et al.* 2009).

There are two objective reasons explaining why pharmacophore- based target identification has not become yet a standard in-silico ligand profiling method:

(i) the abscence of an exhaustive collection of protein−ligand based and(or) ligand-based pharmacophore databases

(ii) the lack of clear benchmarks comparing the later approach to commonly used strategies (2-D and 3-D ligand similarity search, protein−ligand docking) in computational profiling.

We herewith present PharmaDB, the largest ever reported collection of structure based pharmacophore (68 056 entries) from 8 166 protein-ligand X-ray structures. A diverse set of 157 PDB ligands was profiled using 10 screening protocols on the entire pharmacophore collection, thus generating as many matrices of about 11 million data points. Pharmacophore mapping was compared to another 3-D structure-based method (docking) and to ligand-centric approaches (2-D and 3-D similarity search). In most cases, ligand-based profiling methods outperformed structure-based approaches in their ability to recover the true targets among top-scoring entries. Fine analysis of successes and failures for all methods suggests the design of a hybrid profiling method using the best possible approach as a function of ligand and binding site properties.

3. Methods

3.1. Generation of Receptor-Ligand Pharmacophores

The Receptor-Ligand Pharmacophore Generation (RLPG) Protocol in Discovery Studio 3.1 (Accelrys 2012) generates pharmacophore models directly from the receptor-ligand interactions as revealed in the 3-D structures. The RLPG protocol has some notable features: (i) it is fully automated and quickly converts receptor−ligand complexes into pharmacophore models, (ii) it uses adjustable constraints to determine the receptor−ligand interactions, and (iii) it creates all possible pharmacophore combinations, ranks the pharmacophores by decreasing selectivity score, and returns the top-ranked ones. The overall procedure is briefly described as follows. In the first step, pharmacophore features of the ligand are identified. Six standard pharmacophore features are considered: hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), positive ionizable (PI), negative ionizable (NI), hydrophobic (HYD), and ring aromatic (RA). In the second step, the algorithm prunes all features that do not match the protein-ligand interactions using adjustable topological rules (Sutter *et al.* 2011).

Hydrogen Bond Donors/Acceptors: Hydrogen bonds between the protein and ligand are identified, with a default distance of 3.0 Å between heavy atoms. If the enumerated HBA or HBD feature matches the hydrogen bond interaction between the receptor and ligand, it is retained. All others are removed.

Hydrophobic: Hydrophobic features on the ligand are retained if they are within 5.5 Å of the centroid of a hydrophobic residue (Ala, Cys, Ile, Leu, Met, Phe, and Val).

Positive and Negative Ionizable: If an opposite charge center is found on the protein side within 5.6 Å, the feature is retained. All others are removed.

Ring Aromatic: This feature is retained if an aromatic ring is found on the protein and is 2.5 Å away from the projection point of the RA on the ligand.

To construct pharmacophore models that are both sensitive and selective, all combinations of three to six features pharmacophores are enumerated and ranked by decreasing selectivity. Only the top 10 models are selected. There are two options for adding steric constraints to the pharmacophores: shape constraints or excluded volumes. The ligand is used as a template when

creating a shape. When adding excluded volumes, an exclusion sphere is added for each neighboring residue. The size of the exclusion sphere is proportional to the number of neighboring protein atoms within a 4-5 Å distance range.

3.2. Genetic Function Approximation (GFA) Model for Estimating Pharmacophore Selectivity

The selectively of a pharmacophore model depends on the number of features, the feature types and their 3-D arrangement. The selectivity is proportional to the number of hits retrieved upon searching a diverse 3-D database. The more hits from the 3-D search the less selective the model. However, it is not practical to screen thousands of pharmacophore candidates using a 3-D database search. Therefore, a mathematical model was created to predict the number of hits rather than performing the search itself. The model was built using default settings of the GFA algorithm (Rogers *et al.* 1994) embedded in Discovery Studio. Some details for building the GFA model are given as follows.

Druglike Diverse Database: A Catalyst 3-D database of 5 390 druglike diverse ligands was generated in Discovery Studio (default settings of the Build 3D database protocol).

The drug-like data set consisted of 3 000 drug-like compounds randomly selected from the BioinfoDB database v11.1 (http://bioinfo-pharma.u-strasbg.fr/bioinfo, (Kellenberger *et al.* 2004)) and 2 390 selected from the CAPDiverse database in Discovery Studio.

Diverse Pharmacophore Models: A total of 1 544 pharmacophore models are generated from 200 non redundant sc-PDB protein-ligand complexes with 2-8 features. Each pharmacophore is used to search the Druglike Diverse database, and the logarithmic value of the number of hits is used for training the GFA model. For pharmacophores with 2-5, 6, 7, and 8 features, the logarithmic value for zero hits is approximated as ln(0.3), ln(0.1), ln(0.03), and ln(0.01), respectively.

Descriptors: Two types of descriptors are used to describe a pharmacophore model: (i) feature set descriptors and (ii) feature-feature distance descriptors. Ten descriptors (number of features, count of certain feature types) were used to specify the feature set of a pharmacophore. The

remaining 210 descriptors are related to the feature locations. For each pair of feature types, the feature-feature distance is put into a distance bin (1-10), with a bin size of 2.0 Å. The distance bin count is used as the descriptor value. For instance, descriptors Desc11-Desc20 are used for HBA-HBA distances. Desc11 is the number of HBA-HBA distances in the range of 0-2.0 Å. Desc12 is the number of HBA-HBA distances in the range of 2.0-4.0 Å, and so on. Distances greater than 20.0 Å are counted in the last bin, i.e., Desc20. Similar descriptors are defined for the other types of feature-feature distances. The descriptors and the corresponding feature-feature distance types are shown in Table 1 of the Supporting Information.

GFA Model: Ten GFA models were created using the pharmacophore descriptors to predict for each of the 1,544 diverse pharmacophores the logarithmic value of the number of obtained hits (GFAscore). The best GFA model contains six terms and exhibits an R^2 value of 0.881 (Figure 1 of the Supporting Information). Selectivity is derived from the GFA score using the following equation:

$$\text{Selectivity} = 11 - \text{GFAscore}$$

The constant of 11 ensures that the selectivity scores will be positive in nearly all cases.

3.3. The PharmaDB Pharmacophore Data Set

The RPLG algorithm was applied to 8 166 protein-ligand complexes from the sc-PDB database (release 2010,(Meslamani *et al.* 2011)). Proteins and ligands were downloaded from the sc-PDB Web site (http://bioinfo-pharma.u-strasbg.fr) in mol2 file format with formal charges and used as input files by the pharmacophore generation protocol.

Three to six features were required for each pharmacophore model, and up to 10 pharmacophores per complex were created. Default settings were used to assign pharmacophoric features on both the ligand and the receptor, as well as for detecting receptor-ligand interactions on the fly. No ligand shape information was explicitly defined, only receptor-based exclusions spheres were included in all models. A total of 7 687 out of the starting 8 166 sc-PDB complexes yielded at least one valid pharmacophore. Altogether, the PharmaDB collection totals 68 056 pharmacophores (chm files) from 2 556 different targets and 3 916 unique ligands.

3.4. The sc-PDB Diverse Ligand Set

All sc-PDB binding sites were clustered according to their pairwise similarity computed by the FuzCav method (Weill *et al.* 2010). Briefly, FuzCav converts 3-D atomic coordinates into a vector of 4 834 integers reporting counts of all possible pharmacophoric feature triplets (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, aromatic, hydrophobic) from binding site-lining residues. The full similarity matrix was converted into a distance matrix, and a hierarchical clustering (average linkage) was then applied in addition to a stopping criterion (Distance > 0.84) for the cluster agglomeration. A total of 1 416 binding sites clusters containing 4 228 different ligands were defined. This ligand set was filtered for druglikeness with Filter (OpenEye 2012) (see filtering rules in Table 2 of the Supporting Information) to yield a total of 939 unique druglike sc-PDB ligands. A single ligand was randomly chosen for each populated cluster at the condition that the corresponding targets were non-redundant. Finally a total of 182 sc-PDB ligands was obtained and supplemented by 18 ligands from the Astex Diverse Set (Hartshorn *et al.* 2007) not present in the sc-PDB.

To remove chemical similarity, all 200 remaining ligands were compared using ECFP_4 circular fingerprints (Rogers *et al.* 2010) and kept if dissimilar enough (Tanimoto coefficient < 0.7) from all other compounds of the set. This procedure led to a total of 157 unique ligands (sc-PDB Diverse Ligand Set), available in 2-D sd and 3-D mol2 file formats for download from our Web site (http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/scPDB_DiverseSet.zip).

3.5. Computational Ligand Profiling

The 157 ligands from the sc-PDB Diverse Set were profiled against the 2 556 targets of the PharmaDB collection (7,687 sc-PDB entries) using four different virtual screening methods.

2-D Similarity Search: The similarity of the query ligand to the starting 3 916 sc-PDB ligands was computed from ECFP_4 fingerprints in PipelinePilot (Accelrys 2012) and the corresponding targets ranked by decreasing Tanimoto coefficient. The highest similarity value was kept for every sc-PDB target.

3-D Similarity Search: A conformer database (3-D sd file format) was generated using default FAST settings of the Conformation Generator component in Pipeline Pilot from each query ligand (2-D sd file) and compared to all sc-PDB ligands (X-ray structure) with ROCS (OpenEye 2012). The corresponding targets were ranked by decreasing Comboscore of their cognate ligands. The highest value was kept for every sc-PDB target.

PharmaDB Pharmacophore Search: The above-described conformer database of query ligands was mapped to all PharmaDB pharmacophores using default settings of the "citest" executable in Discovery Studio. Only the best mapping was kept for each query ligand, and the maximum number of omitted features was set to −1 in both rigid and flexible fitting mode. For every query ligand, targets were ranked by decreasing *fitvalue*. A second score, the *adjusted fitvalue* was computed as follows.

$$\text{Adjusted Fitvalue} = (\text{Fitvalue} \times M)/T$$

where M is the number of mapped features, and T is the number of total features in the pharmacophore model. This score will provide a little correction if the conformational sampling of the input ligand was not sufficient to get an optimal mapping to pharmacophore features of the model.

Docking: The 157 query ligands were docked to the 7 687 sc-PDB binding sites using default settings of Surflex (v2.412; (Spitzer *et al.* 2012)) and Plants (v.1.2; (Korb *et al.* 2009)) programs. A maximum of 10 docking poses was saved for every ligand by each program according to their native scoring function (pKd for Surflex, ChemPLP for Plants). All poses were rescored using the FingerPrintLib program (Marcou *et al.* 2007) which converts protein-ligand coordinates into molecular interaction fingerprints. Two fingerprints were computed: the first one registers eight interactions (one bit/interaction) per binding site residue (hydrophobic, aromatic, H-bond, ionic, and metal complexion), and the second stores only information from polar interactions (H-bond, ionic, and metal complexion).

Similarity of both fingerprints to the X-ray sc-PDB complex was expressed by Tanimoto coefficients (Tc1, Tc2). Only poses with a Tc1 value higher than 0.6 and a Tc2 value higher than 0.2 were kept. Remaining poses were then ranked by decreasing docking score.

4. Results and discussion

4.1. PharmaDB Pharmacophore Collection

A collection of 68 056 pharmacophore models has been automatically generated from 7 687 out of the starting 8 166 sc-PDB complexes (conversion rate of 94%). Several reasons of failure could be identified (Table 1).

Failure	Number of cases
No feature mapping	3
Minimal feature−feature distance criteria not fulfilled	79
No pharmacophore maps the ligand	11
Less than 3 features	372
Invalid valence	14

Table 1: Failures in Processing sc-PDB Entries

In most of the cases, no pharmacophore model was outputted because of a limited number of pharmacophore features (less than 3). The second major reason was the close proximity of some features that did not respect a minimal interdistance threshold (1.0 Å) and were therefore removed. Last, very few entries contain either valence errors for the bound ligand or did not lead to any feature mapping and should consequently be removed from the next sc-PDB version. A total of 54% of the pharmacophores have the maximum requested number of six features, 18% have five features, 15% have four features, and 13% have three features (Figure 1).

Therefore, a very large majority of sc-PDB entries is described by the upper limit of 10 different pharmacophore models. As expected, the selectivity is dependent on the number of features, complex pharmacophores being more selective than simpler ones (Figure 1).

The PharmaDB pharmacophore collection describes the interaction of 3 916 unique ligands with 2 556 unique targets and is by far the largest repository of pharmacophores reported to date.

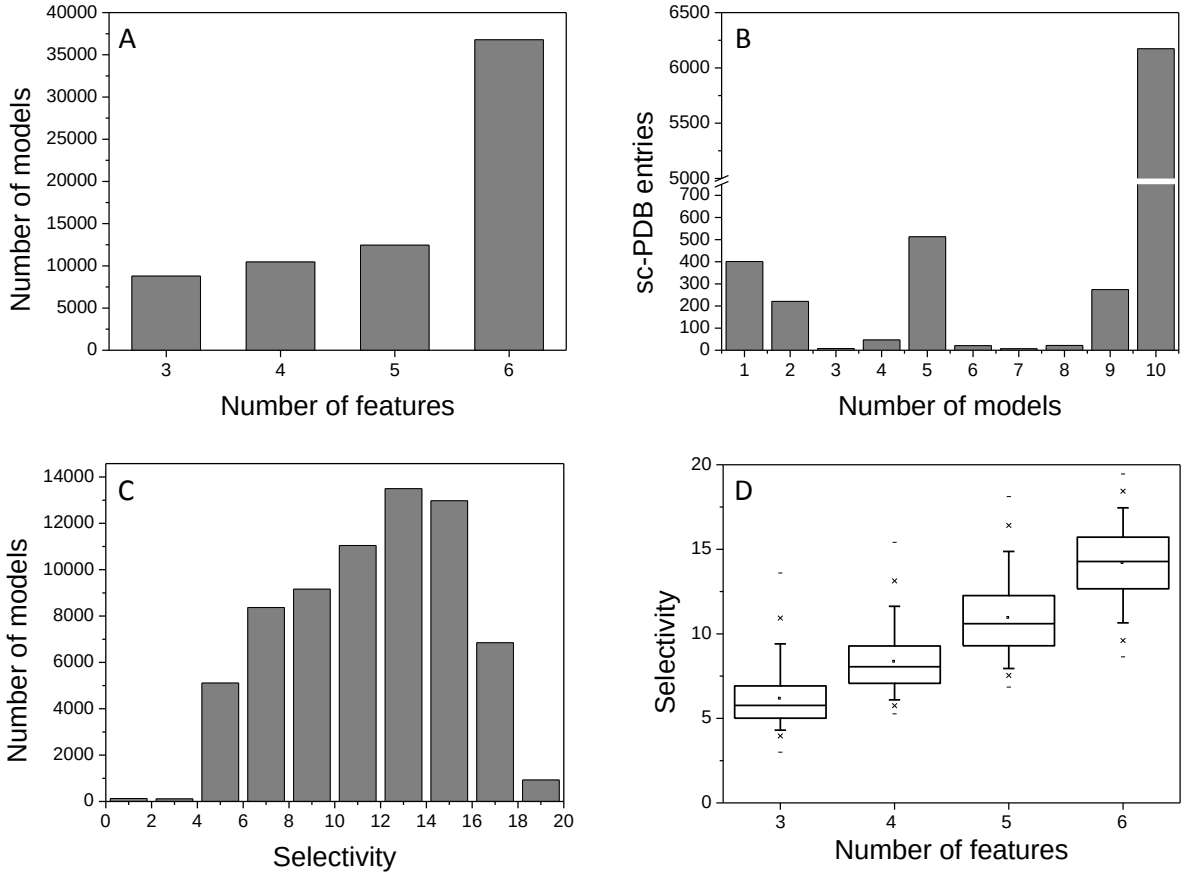


Figure 1: The PharmaDB collection of pharmacophores. **(A)** Distribution of the number of features by pharmacophore model. **(B)** Distribution of the number of pharmacophore models by sc-PDB entry. **(C)** Distribution of the selectivity value of pharmacophore models. **(D)** Box-and-whisker plot of selectivity value distributions according to the number of features in pharmacophore models. The box delimits the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

4.2. The sc-PDB Ligand Diverse Set

In order to generate the minimum amount of redundancy in the protein-ligand computational matrix, a diverse set of ligands was extracted from the sc-PDB on the basis of the 3-D diversity of the binding sites onto which they bind. The general idea was to select unique ligands binding to dissimilar cavities. Examination of standard molecular properties confirms that selected

ligands are druglike, not biased toward unintended property ranges, and chemically dissimilar (Figure 2 A-G).

Out of the 157 ligands of the sc-PDB Diverse Set, 130 (83%) have been cocrystallized with a single target in the sc-PDB, 19 have two different targets, six have three different targets, one ligand has four different targets, and one ligand has five different targets (Figure 2H; see full list of targets in Table 3 of the Supporting Information).

In total, 165 unique targets are addressed by the Diverse Set with a functional annotation, according to the Enzyme Commission number, similar to that of the full sc-PDB database (Figure 2I). For further comparing the merits of various profiling methods, the set was divided in two parts: Set 1 includes 29 ligands with targets present in multiple sc-PDB entries and Set 2 describes 128 compounds addressing a target present in a single sc-PDB entry. Because we decided to restrict chemical space to sc-PDB ligands (of known binding mode to their targets), Set 1 ligands can therefore be profiled with either ligand or structure-centric methods, whereas Set 2 ligands can only be profiled by structure-based approaches.

4.3. Comparison of Ligand-Based and Structure-Based Profiling Methods (Ligand Set 1)

Two ligand-based and two structure-based methods were used to profile the 128 ligands of the Diverse Set 1 against the 7 687 sc-PDB entries (Table 2). For the target-ligand-based pharmacophore approach, two scoring schemes were used according to the ligand-to-pharmacophore fitting procedure (rigid or flexible).

For both docking programs (Surflex, Plants), either the native docking score (pKd for Surflex, ChemPLP for Plants) or a combination of interaction fingerprint similarity and docking score (see Methods) was utilized to rank order sc-PDB targets.

Altogether, 10 profiling protocols were then used to yield as many protein-ligand interaction matrices (Table 2). In the current analysis, only protein-ligand complexes registered in the sc-PDB were considered as true positives, although still unknown cross-reaction of some ligands with sc-PDB targets are theoretically possible.

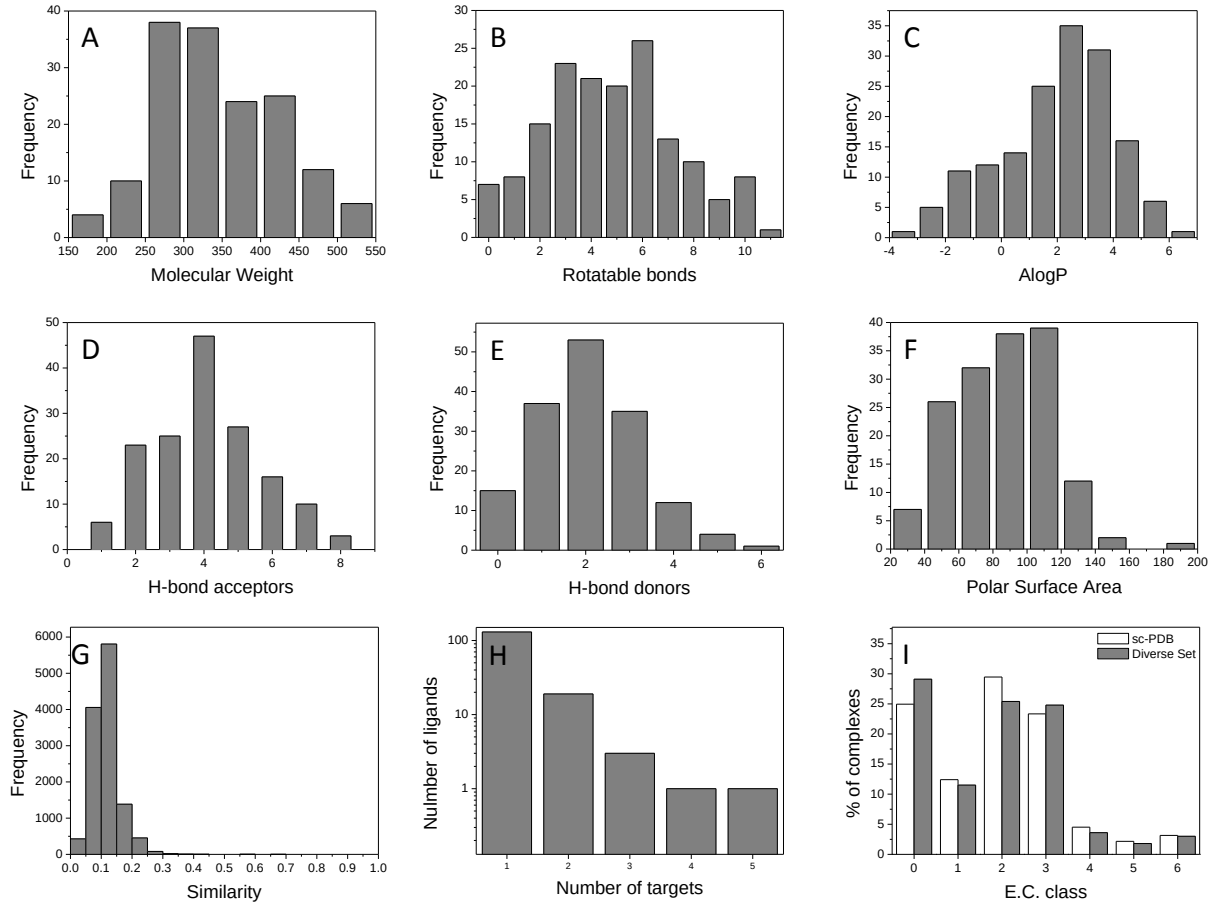


Figure 2: Properties of 157 druglike ligands of the sc-PDB Diverse Set. **(A)** Molecular weight distribution. **(B)** Number of rotatable bonds. **(C)** AlogP, computed logP. **(D)** Hydrogen-bond acceptor count; **(E)** Hydrogen-bond donor coun. **(F)** Polar surface area, Å2. **(G)** All-against-all ligand similarity expressed by a Tanimoto coefficient on ECFP4 fingerprints. **(H)** Distribution of sc-PDB targets among the Diverse Set. **(I)** Functional annotation of targets by Enzyme Commission (EC) number: 0, no EC number; 1, oxidoreductases; 2,transferases; 3, hydrolases; 4, lyases; 5, isomerases; and 6, ligases.

The first criteria to estimate the performance of each profiling method was to compute the rank of the true target for every ligand of the Diverse Set. A more qualitative analysis of the target fishing performance was realized by classifying the profiling in three categories: successful (rank of the true target ≤ 25), ambiguous ($25 < \text{rank} \leq 50$), and failed (rank > 50). In case a compound binds to more than one target (27/157 ligands), the highest-ranked target was considered. The threshold of 25 was chosen both on theoretical and practical considerations because it corresponds to the top 1% ranking targets and a target list of manageable size for experimental confirmation.

Protocol	Method	Scoring
Ligand-based		
ECFP_4	2-D similarity	Tanimoto coefficient
ROCS	3-D similarity	Comboscore
Structure-based		
Rigid1	Rigid fit to pharmacophore	FitValue
Rigid2	Rigid fit to pharmacophore	Adjusted FitValue
Flex1	Flexible fit to pharmacophore	Fitvalue
Flex2	Flexible fit to pharmacophore	Adjusted Fitvalue
Surflex1	Docking	pK_D
Surflex2	Docking	pK_D + IFP
Plants1	Docking	ChemPLP
Plants2	Docking	ChemPLP + IFP

Table 2: Computational Ligand Profiling Protocols

When all profiling methods can be compared (29 ligands of Set1), ligand-based methods clearly outperform structure-based approaches both quantitatively and qualitatively (Figures 3, 4; Table 4 of the Supporting Information). It should be stated at this point that sc-PDB entries cocrystallized with the ligand to profile were not considered in the analysis below.

A 2-D similarity search, the fastest method evaluated in the current study, was slightly better (median rank of the true target = 3, success rate = 76%) than 3-D pharmacophoric shape matching (median rank = 5, success rate = 72%). Although to be expected, this observation is appealing because only a very restricted ligand space (3 916 sc-PDB ligands) was considered, and a simplest nearest-neighbor approach was utilized to rank the corresponding targets. The propensity of structural biologists to cocrystallize multiple ligands of the same chemical series certainly provides a slight bias in our observation.

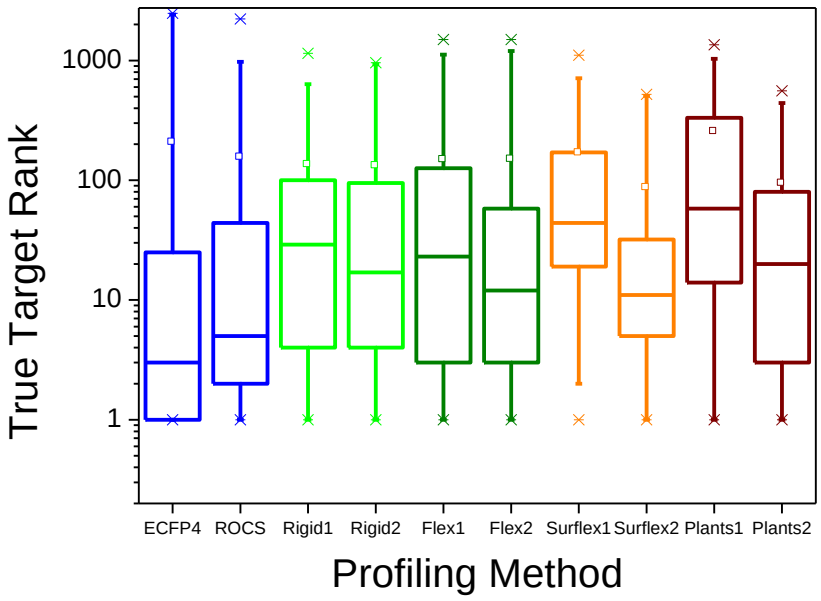


Figure 3: Comparative evaluation of 2 ligand-based and 8 structure-based protocols (see description in Table 2) for profiling 29 ligands (Set 1) against 2 556 unique sc-PDB targets. Box-and-whisker plot of the distribution of true target ranks. The box delimits the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

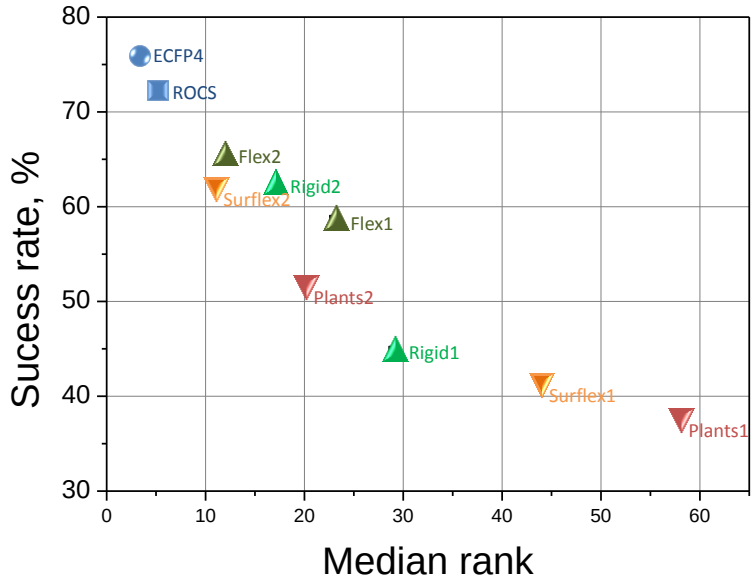


Figure 4: Profiling success rate of 29 ligands (Set 1) as a function of the true target median rank for 10 computational profiling protocols.

However, this bias would be largely counter-balanced by an extension of ligand space to larger bioactivity databases (e.g., CHEMBL (Gaulton *et al.* 2011)).

The two structure-based approaches (pharmacophore search, docking) were significantly less accurate than ligand-centric methods but could be improved by using customized scoring functions (Figure 4). Using an *adjusted fitvalue* improved the success rate of both rigid (+17%) and flexible (+7%) pharmacophore matches with respect to the native *fitvalue*.

A similar observation was done when docking poses, whatever the program (Surflex, Plants), were first post-processed by interaction fingerprint similarity to the native X-ray structure and then ranked by decreasing docking score (+21% and +14% increase in success rate for Surflex and Plants, respectively). Flexible pharmacophore match, although much more computer-demanding, was only marginally better (median rank = 12, 66% of success) than rigid fit (median rank = 17, 62% of success).

The current study confirms that unbiased docking is the worst performing method (44 and 20% of success for Surflex and Plants, respectively) due to the inaccuracy of standard scoring functions. We believe that this observation is independent of the docking tool as similar evidence were recently reported for the Glide program with obvious inter-protein scoring noises (e.g., some targets classes being systematically underestimated, others being systematically overestimated (Wang, W. *et al.* 2011)).

Fortunately, correcting the biases induced by the native scoring functions by a topological filter (removing poses leading to interaction fingerprints dissimilar to that obtained from known complexes) drastically enhance the performance of both docking tools (+21% and +14% for Surflex and Plants, respectively). Surflex was better suited than Plants for the current profiling set and achieved a performance comparable to that obtained with the best receptor-ligand pharmacophore matching protocols (Figure 4).

We could not identify any rule relating profiling accuracy to the binding affinity for the true target. For example, all methods correctly identified S-adenosyl-L-homocysteine hydrolase as a target of 3-deaza adenosine (HET code "AD3"), although its IC_{50} is reported to be close to 20 μM. Alternatively, no method was able to identify the true target of nanomolar ligands (e.g., I84 aldose reducatase inhibitor, P34 chomera toxin inhibitor; Table 4 of the Supporting Information).

4.4. Comparison of Structure-Based Profiling Methods (Ligand Set 2)

A deeper comparative analysis of structure-based pharmacophore and docking profiling methods could be drawn from the profiling results of 128 compounds from Set 2, cocrystallized with a single sc-PDB target. Obtained results were in general agreement with the above-reported data for Set 1 compounds, however with some variations (Figure 5):

(i) flexible fitting did not ameliorate the accuracy of pharmacophore matching with respect to the simpler rigid fitting procedure

(ii) use of an adjusted fitness value increases the success rate of the pharmacophore-based profiling (+13% for rigid fitting; +8% for flexible fitting)

(iii) unbiased docking with native scoring functions is not suited for ligand profiling

(iv) post-processing docking poses by interaction fingerprint similarity to the native X-ray structure dramatically enhances the success rate of docking-based profiling (+28% for Surflex, +26% for Plants).

For this data set, pharmacophore-based profiling was clearly superior to docking-based protocols. This statement should however be considered with caution because each compound was matched to a pharmacophore derived from its own interactions with the true target (self-matching), which was not the case for Set 1 compounds.

The performance of pharmacophore matching is therefore overestimated. Anyway, the very good median ranks indicate that the automatically derived pharmacophore queries are very specific. Surprisingly, rigid fitting was found to be slightly superior to flexible fit (Figure 5B), which might be due to the different ways ligand atoms may overlap excluded volumes in both fitting procedures.

4.5. Ligand-Dependent Performance of Profiling Methods

Analysis of the two profiling maps (Tables 4 and 5 of the Supporting Information) shows strong ligand dependencies for all profiling methods. Up to now, we have just analyzed the global performance of several profiling protocols, and current data clearly advise the usage of ligand-based 2-D similarity methods whenever feasible.

However, using a single profiling method for all ligands is not an optimal strategy. From here on, we will examine peculiar cases where a single protocol was successful, analyze the reasons for such behaviors, and propose a rationale for prioritizing the best possible method according to the protein-ligand context.

4.5.1. When To Use 2-D Similarity Search?

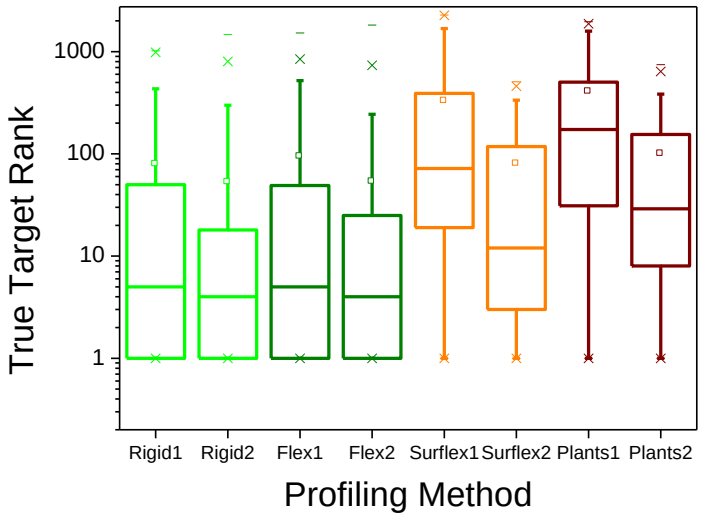
Out of the 29 profilings for which 2-D similarity search could be applied, only five of them failed to recover the true target among the first 50 scoring entries (Table 4 of the Supporting Information). Four of these failures (ET, P34, PVB, RNP) are simply due to an insufficient number of reference ligands (<10) for the cognate targets. The last ligand (I84) that could not be correctly profiled is a human aldose reductase (AR) inhibitor. Although 25 other AR inhibitors were present in the sc-PDB, none of them was similar enough to the query. AR is an enzyme whose 3-D structure is particularly flexible and offers multiple binding modes to chemically diverse inhibitors (Steuber *et al.* 2007). Such a behavior is however expected to be more the exception than the rule.

Thirteen out of the 14 proteins of Set 1 ligands, having more than 20 cocrystallized ligands, were indeed recovered among the top 1% scoring entries.

We therefore recommend the usage of ligand-based 2-D similarity methods whenever possible, which means when enough different ligands (> 20) can be used as references to annotate a target.

In addition to this general consideration, we describe here peculiar ligand and binding site properties favoring exclusively one of the four virtual profiling methods used in the current study. For example, profiling of the pantothenate kinase inhibitor PAU is only successful with a 2-D ligand similarity search (Table 4 of the Supporting Information). PAU is an acyclic small molecular-weight ligand (MW = 218) highly buried (96%) upon binding to its target (PDB entry 3af0, Figure 6A-B). Its closest sc-PDB ligand with respect to 2-D ECFP_4 fingerprints is its phosphate analogue PAZ, another pantothenate kinase inhibitor (PDB entry 3aez; Tc = 0.63; Figure 6C). The 2-D similarity is however not concomitant with 3-D similarity as estimated by ROCS overlay of flexible PAU to rigid PAZ because of the additional presence of an additional polar phosphate group in PAZ (Comboscore = 0.948; Figure 6D).

A



B

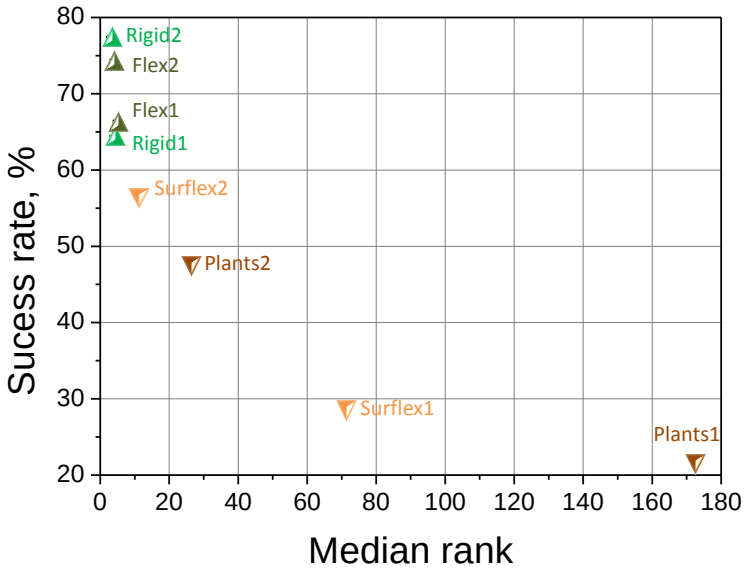


Figure 5: Comparative evaluation of 2 ligand-based and 8 structure-based protocols (see description in Table 2) for profiling 128 ligands (Set 2) of against 2 556 unique sc-PDB targets. **(A)** Box-and-whisker plot of the distribution of true target ranks. The box delimits the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. **(B)** Profiling success rate ligands as a function of the true target median rank for 10 computational profiling methods.

The corresponding true target is thus badly ranked (rank 293) in the ROCS-derived target list. The closest ligand to PAU in our 3-D descriptor space is another low molecular-weight inhibitor

(BTW, MW = 209) of a totally different target (carboxypeptidase A1, PDB entry 3i1u) that better matches in 3-D space (Comboscore = 1.413) although the pairwise 2-D similarity is low (Tc = 0.19; Figures 6 C,D).

Because of its low molecular weight and high polarity, PAU is found to fit many structure-based pharmacophores and to dock many binding sites, therefore explaining the failure of all structure-based protocols to correctly profile this compound. We therefore suggest profiling, whenever feasible, polar fragment-like compounds with 2-D similarity search methods.

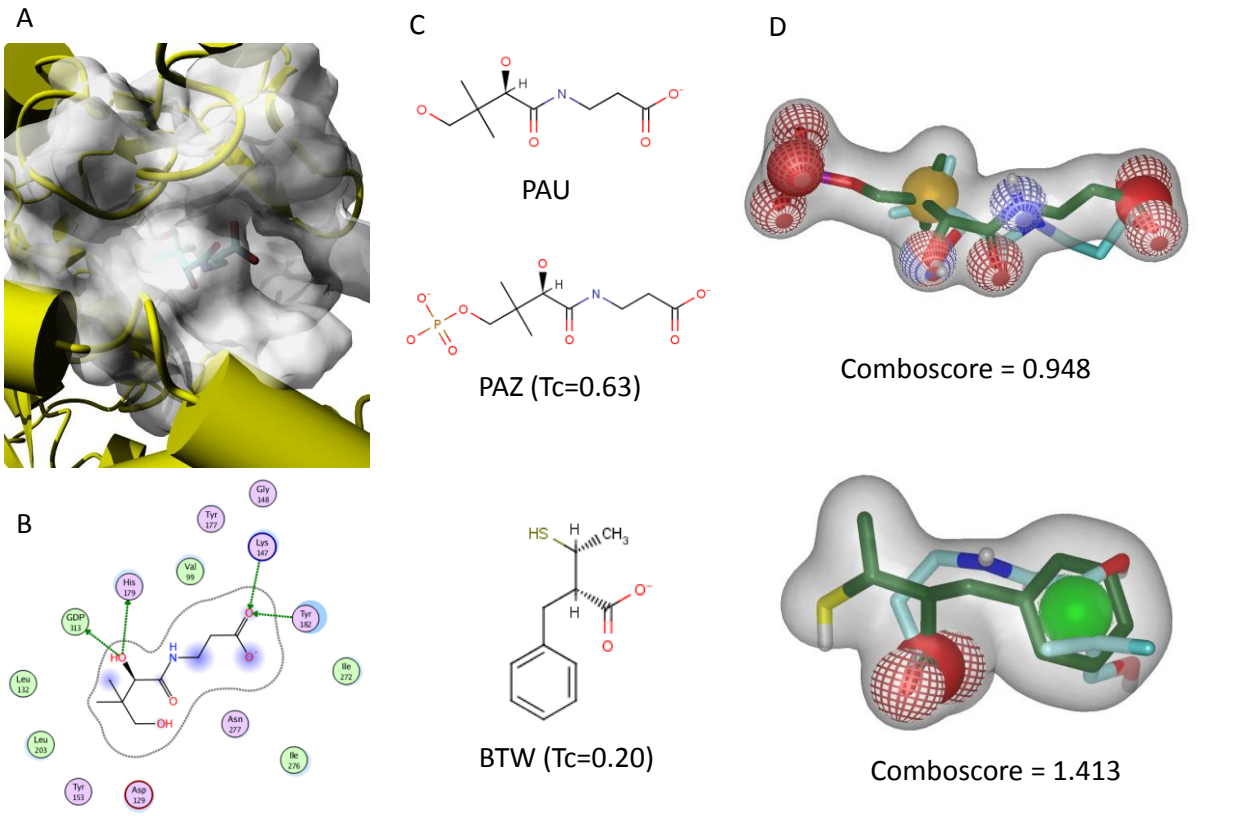


Figure 6: Profiling pantothenic acid (HET code = PAU). **(A)** X-ray structure of PAU (sticks) in complex with pantothenate kinase (white surface, pdb entry 3af0). **(B)** Schematic 2-D diagram of protein−ligand interactions. Apolar and polar binding site residues are circled in green and violet, respectively. Hydrogen-bonds are indicated by green arrows. Accessible ligand atoms are enclosed by a cyan dot. **(C)** Chemical structures of PAU, PAZ, and BTW ligands. 2-D similarity to PAU is indicated in brackets. **(D)** ROCS overlay of PAU (cyan carbon atoms) with PAZ and BTW (green carbon atoms). Oxygen, nitrogen, and sulfur atoms are colored in red, blue, and yellow, respectively. The shape of templates (PAZ, BTW) is displayed by a white surface. Pharmacophoric features are displayed by balls (H-bond acceptor, red mesh dot; H-bond donor, blue meshed dot; negative ionizable, red solid dot, hydrophobe, yellow dot; and ring, green dot).

4.5.2. When To Use 3-D Similarity Search?

In a single profiling example (ET, Table 4 of the Supporting Information), 3-D similarity search was the only method capable of recovering the true targets among the top 1% scoring entries. Ethidium (ET) is a polayaromatic compound (AlogP = 4.20) binding to the transcriptional regulatory protein Qacr mainly through deeply buried apolar and aromatic pi−pi interactions. A single hydrogen-bond to the binding site is observed (Figure 7A, B).

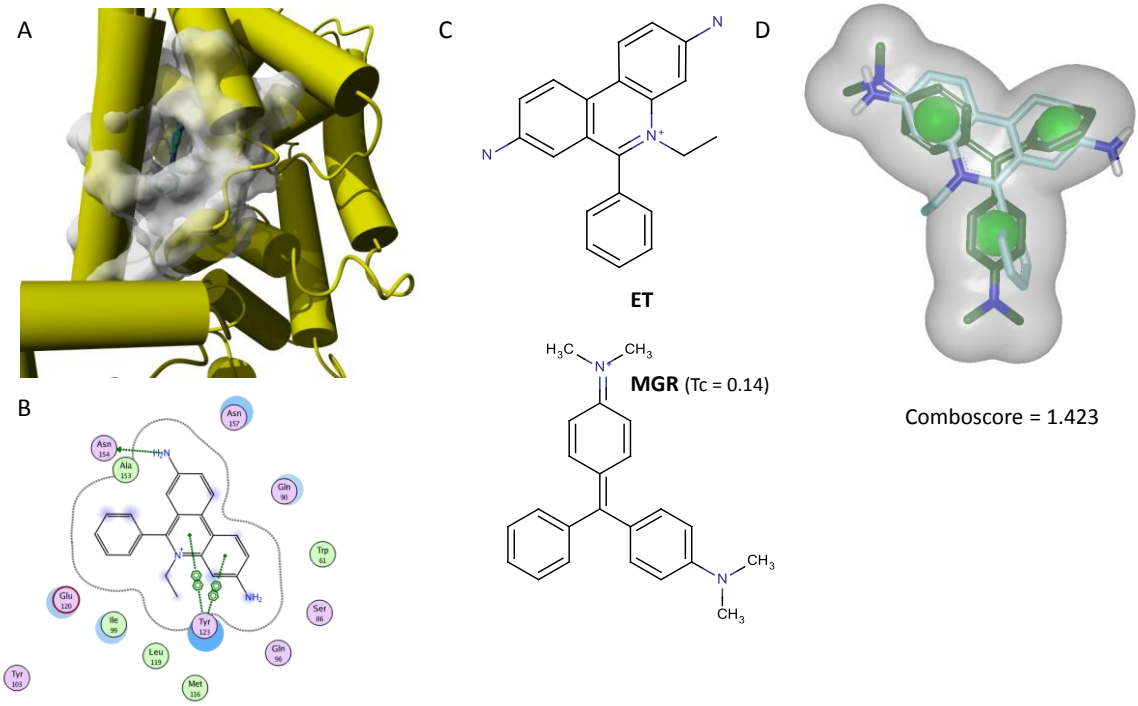


Figure 7 : Profiling Ethidium (HET code = ET). **(A)** X-ray structure of ET (sticks) in complex with the transcriptional regulator protein Qacr (white surface, pdb entry 3br3). **(B)** Schematic 2-D diagram of protein−ligand interactions. Apolar and polar binding site residues are circled in green and violet, respectively. Hydrogen-bonds are indicated by green arrows. Aromatic pi-pi interactions are displayed by dotted lines. Accessible ligand atoms are enclosed by a cyan dot. **(C)** Chemical structures of ET and MGR ligands. 2-D similarity of the closest ligand (MGR) to ET is indicated in brackets. **(D)** ROCS overlay of ET (cyan carbon atoms) with MGR (green carbon atoms). Oxygen and nitrogen atoms are colored in red and blue, respectively. The shape of MGR template is displayed by a white surface. Pharmacophoric features are displayed by balls (ring, green dot).

The 2-D similarity of ET to the six other Qacr sc-PDB ligands is low (Tc in the 0.09-0.14 range, Figure 7C), although 3-D ROCS similarity to one of these ligands (MGR) is high (Figure 7D). The shape and three ring aromatic features of MGR are well matched by ligand ET.

This ligand matches many other pharmacophores dominated by apolar and aromatic features and exhibiting low specificity values.
Docking is also not suited for profiling such compounds where interactions are not directional and therefore badly scored. We thus recommend 3-D similarity search for profiling hydrophobic compounds exhibiting few hydrogen-bond donors/acceptors.

4.6. Receptor-Ligand Pharmacophore versus Docking-Based Profiling.

Ligand mapping to a receptor-ligand-based pharmacophore can be considered as a variant of molecular docking in which the estimation of protein-ligand interactions is not quantified by a binding energy but a fitness to a topological description.
In order to delineate differences in both approaches, we identified ligand profiling cases where rigid pharmacophore search (Rigid2 protocol) was successful but both docking-based protocols (Surflex2, Plants2) failed, and vice versa. Scores used in this comparison were the adjusted fitvalue for the rigid pharmacophore method and the docking score preprocessed by interaction fingerprints for docking as mentioned above. For nine ligands, only pharmacophore search was successful, whereas eight cases could be reported in which only docking-based approaches were efficient in recovering the true target among the top 1% scoring targets. Examining the molecular properties of both ligand sets show some clear tendencies: ligands suited for docking-based profiling are more polar (higher number of hydrogen-bond donors, lower clogP, and higher polar surface area) than ligands suited for receptor-ligand-based profiling (Figure 8). Interestingly, the properties of the corresponding binding sites matched the above-reported ligands properties. Targets recovered by docking exhibit binding cavities that are more buried, polar, and smaller than those recovered by pharmacophore search (Figure 9). This observation corroborated most benchmarks, indicating that this method is highly sensitive to the directionality of protein-ligand interactions (e.g., hydrogen-bond count), cavity size, and polarity (Steuber *et al.* 2007). Receptor-ligand-based pharmacophore search, which can be considered as a constrained docking in which positional and pharmacophoric constraints have been automatically derived, does not suffer from this drawback.

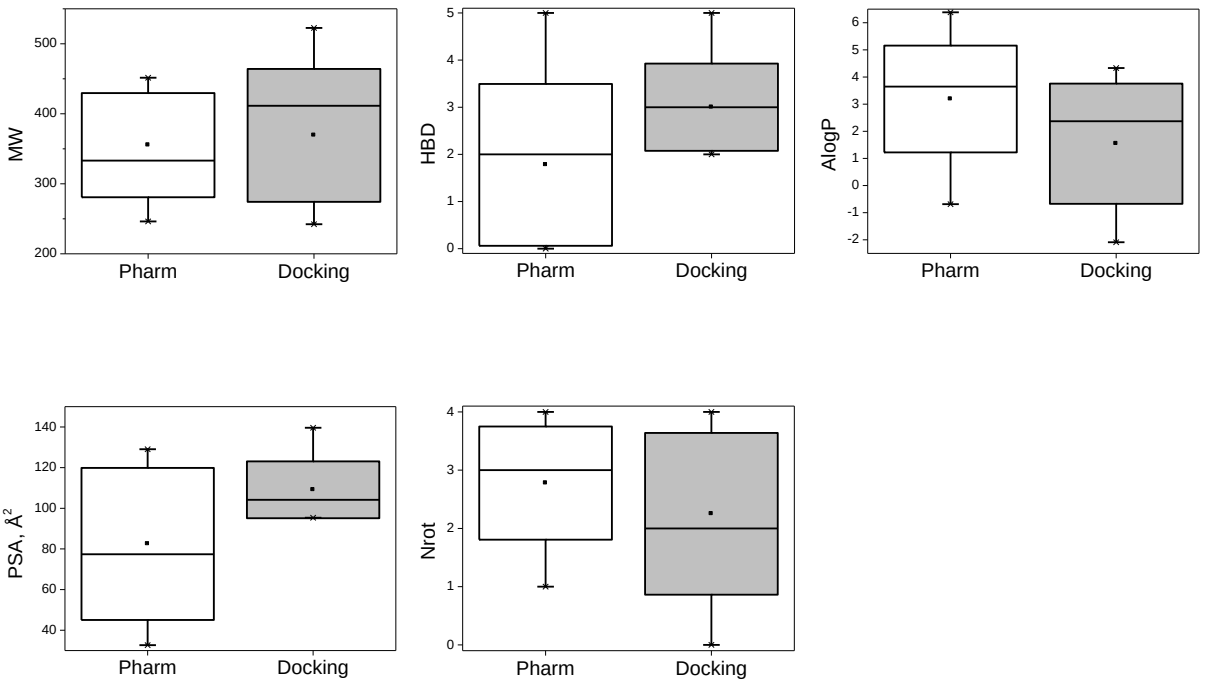


Figure 8: Molecular properties (molecular weight, MW; hydrogen-bond donor count, HBD; predicted log P, AlogP; polar surface area, PSA; and number of rotatable bonds, Nrot), computed by Pipeline Pilot, of ligands whose targets are only recovered by either pharmacophore-based search (Pharm) or docking-based profiling (Docking). The box delimits the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

We next compared the quality of the poses generated by either pharmacophore match or docking using the previously identified best profiling procotols for each method (rigid2, flex2, surflex2, plants2; Figure 5). Instead of reporting root mean square deviations (rmsd) to the X-ray pose, we plotted the similarity of predicted to experimentally observed protein- ligand interactions (Figure 10). The later measure was previously reported to be a much better indicator of pose quality than rmsd (Marcou *et al.* 2007).

Analysis of the best pose for the 128 Set 2 ligands from pharmacophore match or docking indicates that flexible pharmacophore fitting produces the best poses but with only a marginal superiority to rigid pharmacophore fit and docking (Figure 10). In 80% of the cases, all methods provides an orientation in which about 60% of protein-ligand interactions are conserved (Tc > 0.6), a threshold considered as acceptable for estimating the quality of a pose (Marcou *et al.* 2007).

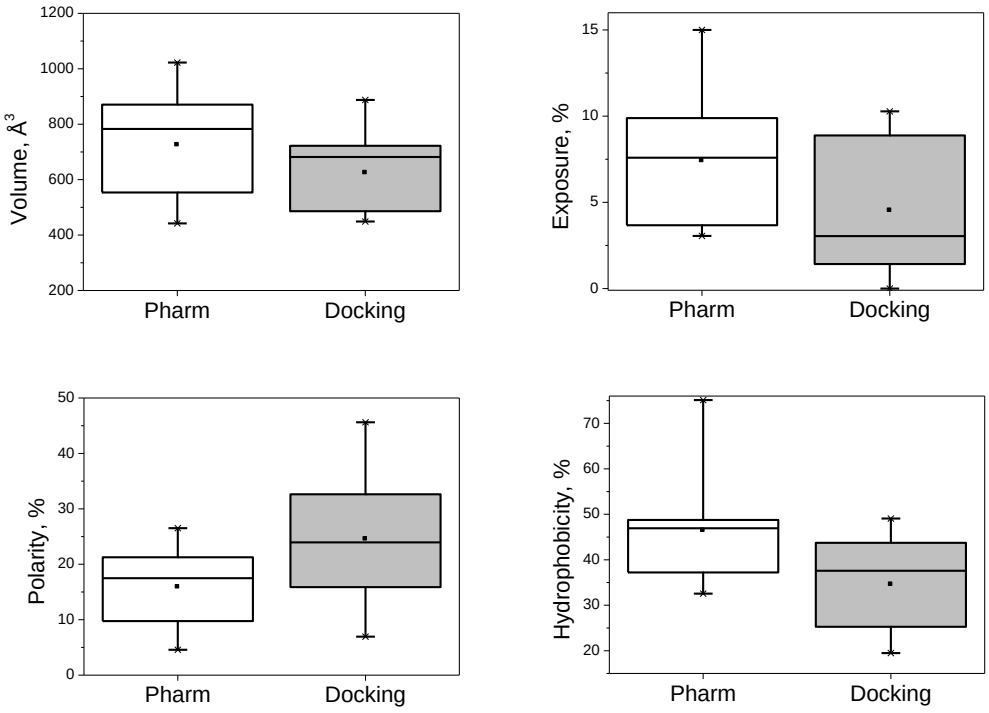


Figure 9: Molecular properties (volume, exposure, polarity, hydrophobicity), computed by VolSite (unpublished in-house program) of binding sites whose targets are only recovered by either pharmacophore-based search (Pharm) or docking-based profiling (Docking). The box delimits the 25th and 75th percentiles; the whiskers delimit the 5th and 95th percentiles. Median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

The poorer profiling performance of docking with respect to pharmacophore search (Figure 5) is therefore only attributable to scoring and not to insufficient conformational sampling of the ligand. Interestingly, flexible fit to a receptor−ligand pharmacophore does not provide a substantial advantage to the rigid fitting procedure, although the later is much faster (Table 3, Figure 2 of the Supporting Information).

Approach	Method	CPU time, s
Pharmacophore Search	Rigid	4
	Flex	70
Docking	Surflex	25
	Plants	20

Table 3: Median CPU Time (128 ligands of Set 2) for Profiling One Target by Pharmacophore and Docking Protocols. CPU time for a 3.16 Ghz Intel Core Duo E 8500 processor with 4 Go RAM.

In conclusion, receptor-ligand pharmacophore search can be considered as a reliable and fast alternative to molecular docking. We recommend this methodology for profiling targets for which few ligands but a 3-D structure is available, with the exception of profiling polar ligands to small, polar, and buried active sites for which molecular docking is preferable.

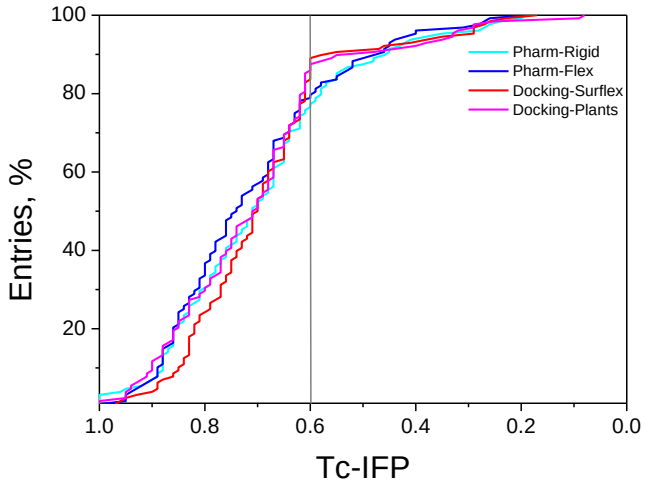


Figure 10: Quality of poses generated by pharmacophore match (rigid, flexible) and docking (Surflex, Plants), for the set of 159 sc-PDB diverse ligands. Similarity to the native X-ray pose is estimated by a Tanimoto coefficient on protein-ligand fingerprints (Tc-IFP). A single pose is retrieved for each method according to its score (adjusted fitvalue for pharmacophore search, pKd for Surflex, ChemPLP for Plants). Please note that raw docking poses are first processed to discard solutions for which Tc1-IFP is less than 0.6 and Tc2-IFP is less than 0.2 (see Methods). A solid vertical line at Tc1 = 0.6 indicates the threshold for acceptable solutions.

5. Conclusion

In this work, we describe a fully automated method to generate 3-D pharmacophore queries from protein-ligand X-ray structures, with an estimation of pharmacophore selectivity based on the number anticipated druglike hits. The protocol was applied to the sc-PDB data set of protein-ligand complexes to generate a database of 68 056 pharmacophores (PharmaDB) describing 2556 unique targets. This study offered us the opportunity to compare, for the first time, ligand-based and structure-based methods to profile a set of 157 diverse ligands against our panel of targets.

When applicable (more than 20 ligands known for each target), ligand-based methods (2-D and 3-D similarity search) clearly outperformed structure-based (pharmacophore, docking) profiling protocols. Pose accuracy is relatively similar for pharmacophore fitting and docking, whatever the protocol used. Because that accuracy is satisfactory for about 80% of ligands (rmsd < 2 Å, Tc-IFP > 0.6), we conclude that the lower performance of structure-based methods is due to the difficulty of scoring functions to discriminate correct poses from near-native decoys.

Failure of scoring functions in docking may be significantly rescued by re-ranking poses by interactions fingerprint similarity to a known reference (same protein cocrystallized with another ligand), which means that sampling and generating correct poses for most druglike compounds is not the major problem here.

Target flexibility is also partially addressed in our application because multiple copies of the same protein bound to various ligands are stored in the sc-PDB and therefore explicitly taken into account in both docking and pharmacophore-based profiling. Multiple reasons remain for explaining inaccurate scoring in docking, all of which have been extensively surveyed in recent reports (Novikov *et al.* 2011; Schneider *et al.* 2011; Smith *et al.* 2011).

Some are due to the necessary high throughput requested in preparing a large collection of heterogeneous binding sites for docking: possible inaccurate tautomeric state for some histidines, possible ligand-dependent flip of terminal amide bonds, and omission of ligand-dependent protein-bound water molecules. Some are intrinsic to any docking-based virtual screen: improper handling of dehydration and more globally entropic effects in absence of a slower but more rigorous energy function to rerank binding poses, neglecting the quality of the host protein structure in docking scores, and absence of accurate terms for weaker but sometimes important intermolecular interactions (e.g., weak hydrogen and halogen bonds, quadrupole-quadrupole interactions). Improper treatment of entropic effects and of peculiar protein-ligand interactions also applies to pharmacophore matches. Possible reasons for inaccurate fitting of true actives to structure-based pharmacophores have also been reviewed by many authors (Wolber *et al.* 2008; Spitzer *et al.* 2010; Leach *et al.* 2011).

Of particular concern in our application is the choice and placement of pharmacophoric features (notably hydrophobic features) and of exclusion spheres that are known to strongly influence pharmacophoric matches. We should point again that the herein derived conclusions have been drawn in a ligand profiling context, which is a particular application of virtual screening.

The present conclusion that ligand-centric approaches outperform structure-based methods in profiling is pragmatic and based on existing data at the PDB scale. We clearly demonstrate that the profiling accuracy of all methods is target and binding site dependent. Examining cases of successes and failures suggest the use of hybrid profiling workflows in which all methods should be applied depending on the protein-ligand context. This strategy presents the advantage to be data-driven and to significantly extend the applicability domain of ligand profiling to a wide array of different targets. Additional independent benchmarking studies as well as the increasing availability of high-quality bioactivity data will certainly help in refining the herein derived first conclusions on computational ligand profiling.

6. Contributions des collaborateurs

J. Li, J. Sutter, A. Stevens et H-O. Bertrand ont généré la base Pharmadb et le modèle prédictif de la sélectivité des pharamcophores.

7. Commentaires et validation expérimentale

Lors de cette étude, nous avons démontré que les pharmacophores d'interactions protéine-ligand était un outil utile pour du profilage notamment quand la cible dispose de peu de ligands connus.

Nous avons essayé de valider expérimentalement quelques cibles secondaires identifiées à partir de la matrice de profilage générée. Nous avons tenté de nous procurer les molécules profilées et d'essayer de trouver des laboratoires de recherche qui disposaient d'un test de liaison pour vérifier nos associations prédites. Malheureusement, dans la base PDB, très peu de ligands sont commercialisés. Nous sommes arrivés à nous procuré deux molécules (Figure 11) : l'Efaproxiral (HET code : RQ3) et le Tadalafil (HET code : CIA), pour lesquelles trois cibles ont été testées. Ces cibles secondaires ont été choisies uniquement parmi les cibles identifiées à l'aide des pharmacophores et parmi les 20 premières (1% des cibles) de la liste des cibles potentielles.

Les cibles retenues sont au final celles dont le site de liaison est dissimilaire à celui de la cible principale. La similarité est évaluée à l'aide du descripteur FuzCav (Weill *et al.* 2010) où un seuil de similarité inférieur à 0.16 a été utilisé pour retenir les cibles dissimilaires.

A **B**

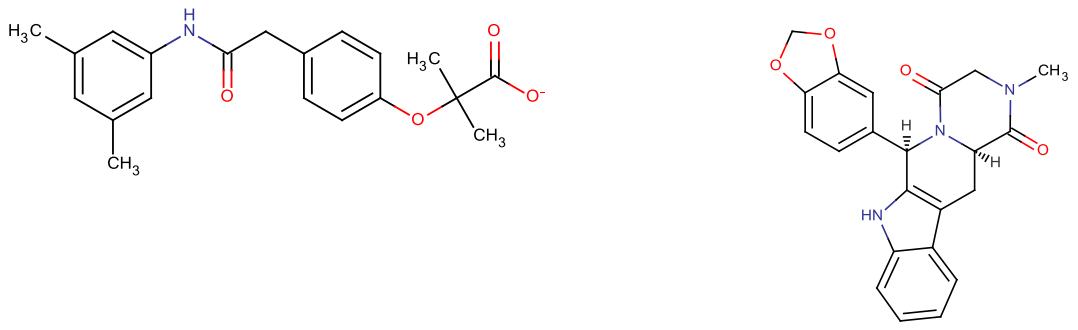


Figure 11 : Structures des molécules : **(A)** Efaproxiral. **(B)** Tadalafil.

Ces cas constituent des cas difficiles du fait que les sites de liaisons des cibles secondaires prédites ne sont pas similaires à la cible principale et donc des exemples intéressant à analyser.

7.1. Validation expérimentale de cibles secondaires de l'Efaproxiral

L'Efaproxiral est un modulateur allostérique de la protéine "Deoxyhemoglobin" (code PDB : 1g9v). Il est administré avant les radiothérapies pour sensibiliser les zones hypoxiques de tumeurs facilitant ainsi la libération d'oxygène à travers les hémoglobines, ce qui améliore la mortalité et réduit la reprise de la croissance tumorale (Source : Thomson Reuters Integrity).

Le tableau 4 présente les 8 cibles identifiées parmi les 20 premières de la liste du profilage et dont le site de liaison était dissimilaire à celui de l'entrée 1g9v. Cette liste recense des cibles très probables car elles ont un score de fit (*fitvalue*) intéressant et les poses produites de l'Efaproxiral dans ses sites de liaisons reproduisent plus de la moitié des interactions existantes avec le ligand natif de chaque protéine (valeurs de Tc1 et Tc2 dans tableau 4).

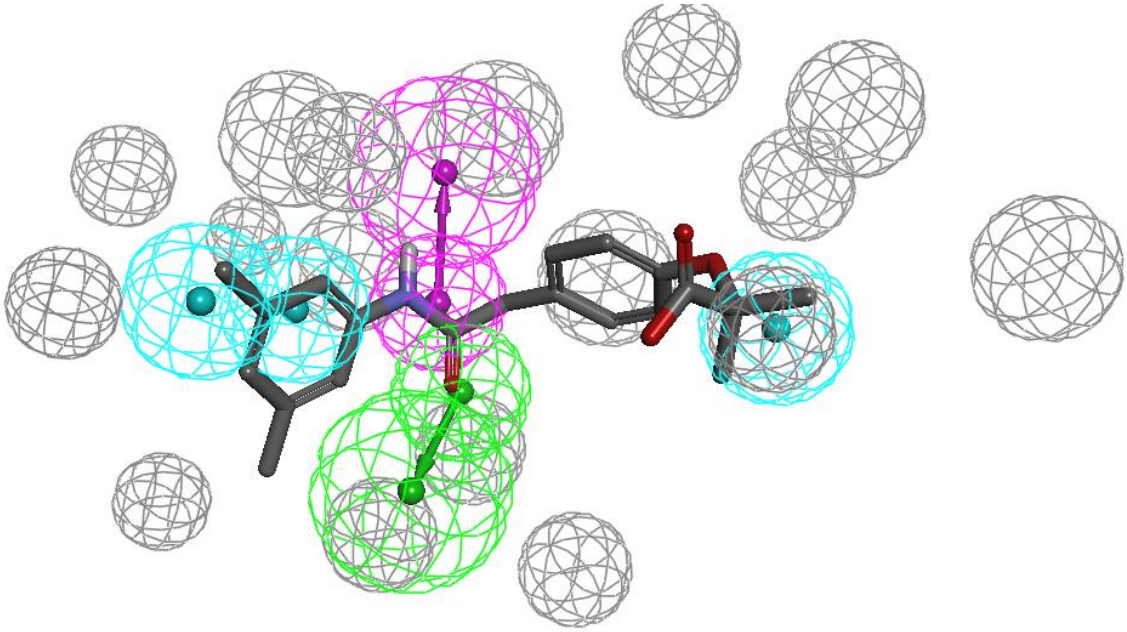
Nous avons pu réaliser deux tests expérimentaux, le premier sur la protéine "Epoxide hydolase 2" et le deuxième sur la protéine récepteur "Epherin type-a 7" dont le mode de liaison prédit est présent dans la figure 12.

PDB	N	Fitvalue	Nombre de motifs	Cible	Tc1	Tc2
3cwk	7	3.89515	4	**Cellular retinoic acid-binding protein 2**	0.75	1
3koo	5	3.86382	5	**Epoxide hydrolase 2**	0.7	1
2jav	6	3.76838	4	**Serine/threonine-protein kinase nek2**	0.65	0.5
2wel	6	3.74122	4	**Calcium/calmodulin-dependent protein kinase type ii subunit delta**	0.65	0.5
2hzn	1	3.73974	4	**Tyrosine-protein kinase abl1**	0.66	0.33
2zm3	6	3.68283	4	**Insulin-like growth factor 1 receptor**	0.72	0.33
3dko	4	3.66593	5	**Ephrin type-a receptor 7**	0.77	1
1opm	2	3.63765	4	**Peptidylglycine alpha-hydroxylating monooxygenase**	0.77	0.8

Tableau 4 : Liste des cibles obtenues pour l'Efaproxiral pour lesquels le site de liaison est dissimilaire avec la cible principale la "Deoxyhemoglobin". N est le numéro du pharmacophore qui a permis de sélectionner l'association ; Tc1 et Tc2 sont les coefficients de tanimoto sur les empreintes d'interactions IFP entre les interactions présentes dans le complexe cristallographique natif et l'Efaporoxiral dans la cavité de l'entrée.

Les tests biologiques réalisés ont permis de confirmer une association, celle de l'Efaproxiral avec la protéine "Epoxide hydrolase 2". Cette protéine se trouve dans le cytoplasme et métabolise les xénobiotiques en dégradant les époxides toxiques.

A



B

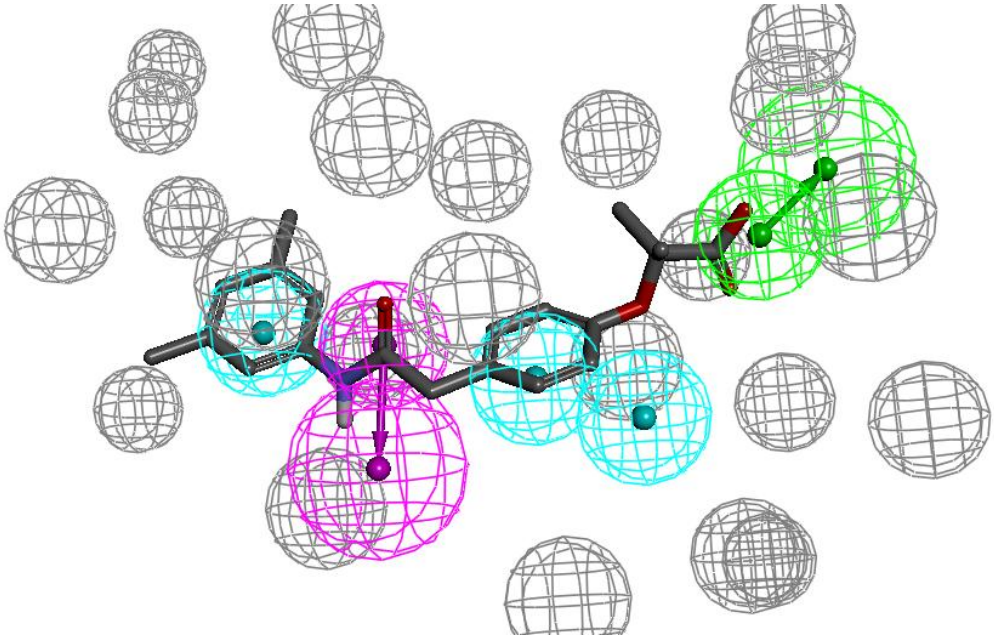


Figure 12 : Pose et mode d'interactions prédits pour l'Efaproxiral avec : **(A)** l'Epoxide hydrolase 2 à l'aide du pharmacophore numéro 5 de l'entrée 3koo. **(B)** le récepteur Ephrin type-a 7 à l'aide du pharmacophore numéro 4 de l'entrée 3dko.

Une inhibition de 30% de l'Epoxide hydrolase a été observée in vitro (Hahn *et al.* 2011) à une concentration d'Efaproxiral de 100 µM (*données E. Proschak., Université de Francfort*). Mais malheureusement aucune liaison n'a pu être observée à la kinase Ephrin type-a 7 (*données CEREP, ref. Essai 3058, Celle L'Evescault, France*).

7.2. Validation expérimentale de cibles secondaires du Tadalafil

La deuxième molécule testée est le Tadalafil. C'est un inhibiteur sélectif de la Phosphodiesterase (PDE5, code PDB : 1xoz) et un médicament pour soigner les troubles érectiles et l'hyperplasie de la prostate (Source : Thomson Reuters Integrity).

La liste des cibles secondaires obtenues est disponible dans le tableau 5, et nous avons pu tester cette molécule sur l'Inosine-5-monophosphate dehydrogenase (ou IMP dehydrogenase).

PDB	N	Fitvalue	Nombre de motifs	Cible	Tc1	Tc2
2zyb	8	3.49392	4	**Tyrosine-protein kinase lck**	0.75	1
3ekn	9	3.17078	4	**Insulin receptor**	0.68	0.33
3khj	1	3.11692	4	**Inosine-5-monophosphate dehydrogenase**	0.79	1
2r3f	4	3.7316	5	**Cell division protein kinase 2**	0.73	0.25
3et1	3	3.08627	4	**Peroxisome proliferator-activated receptor α**	0.65	0.5

Tableau 5 : Liste des cibles obtenues pour le Tadalafil pour lesquels le site de liaison est dissimilaire avec la cible principale la "Deoxyhemoglobin". N est le numéro du pharmacophore qui a permis de sélectionner l'association ; Tc1 et Tc2 sont les coefficients de tanimoto sur les empreintes d'interactions IFP entre les interactions présentes dans le complexe cristallographique natif et l'Efaporoxiral dans la cavité de l'entrée.

Quatre motifs ont été appariés entre le Tadalafil avec le 1er pharmacophore de l'entrée 3khj de l'IMP dehydrogenase. Le mode de liaison prédit est visible sur la figure 13. Cette prédiction a été validée expérimentalement par un test de liaison spécifique (MacPherson *et al.* 2010) à l'IMP dehydrogenase de *Cryptosporidium parvum* (*données L. Hesdtrom, Université Brandeis, U.S.A*). Une inhibition de 25% de l'IMP dehydrogenase étant observée à une concentration de ligand égale à 50µM.

Les interactions qu'on a identifié et mesuré expérimentalement sont certes faibles mais bien présentes. Ces cas représentent des cas difficiles car leurs sites de liaisons sont dissimilaires au site de la cible principale. Notons que ces cas n'ont pu être identifiés que par l'approche des pharmacophores ce qui valide leur utilité dans le cadre d'un profilage.

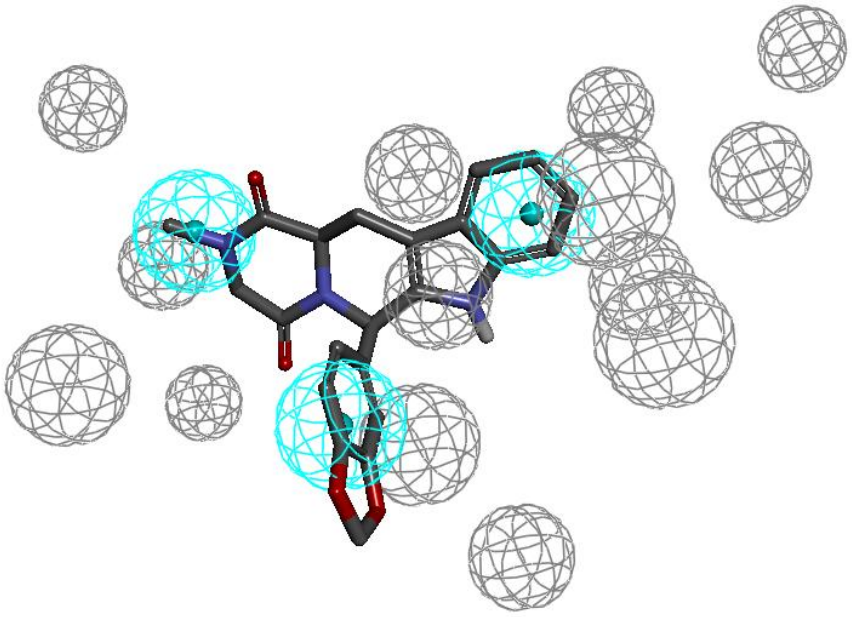


Figure 13 : Pose et mode d'interactions prédits pour le Tadalafil avec l'IMP dehydrogenase à l'aide du pharmacophore numéro 1 de l'entrée 3khj.

8. Références

Accelrys (2012). "Discovery Studio v.3.1.0; Accelrys Software Inc.: San Diego, CA 92121, U.S.A."

Accelrys (2012). "Pipeline Pilot v.8.5.0; Accelrys Software Inc.: San Diego, CA 92121, U.S.A."

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Defranchi, E., C. Schalon, M. Messa, F. Onofri, F. Benfenati and D. Rognan (2010). "Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements." PLoS One **5**(8): e12214.

Durrant, J. D., R. E. Amaro, L. Xie, M. D. Urbaniak, M. A. Ferguson, A. Haapalainen, Z. Chen, A. M. Di Guilmi, F. Wunder, P. E. Bourne and J. A. McCammon (2010). "A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology." PLoS Comput Biol **6**(1): e1000648.

Ekins, S. and A. J. Williams (2010). "When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold?" Drug Discov Today **15**(19-20): 812-815.

Ekins, S., A. J. Williams, M. D. Krasowski and J. S. Freundlich (2011). "In silico repositioning of approved drugs for rare and neglected diseases." Drug Discov Today **16**(7-8): 298-310.

Enyedy, I. and W. Egan (2008). "Can we use docking and scoring for hit-to-lead optimization?" Journal of Computer-Aided Molecular Design **22**(3): 161-168.

Ferrara, P., H. Gohlke, D. J. Price, G. Klebe and C. L. Brooks, 3rd (2004). "Assessing scoring functions for protein-ligand interactions." J Med Chem **47**(12): 3032-3047.

Gaulton, A., L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington (2011). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Res **40**: D1100-1107.

Hahn, S., J. Achenbach, E. Buscato, F. M. Klingler, M. Schroeder, K. Meirer, M. Hieke, J. Heering, E. Barbosa-Sicard, F. Loehr, I. Fleming, V. Doetsch, M. Schubert-Zsilavecz, D. Steinhilber and E. Proschak (2011). "Complementary screening techniques yielded

fragments that inhibit the phosphatase activity of soluble epoxide hydrolase." ChemMedChem **6**(12): 2146-2149.

Hartshorn, M. J., M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. Mooij, P. N. Mortenson and C. W. Murray (2007). "Diverse, high-quality test set for the validation of protein-ligand docking performance." J Med Chem **50**(4): 726-741.

Hopkins, A. L., J. S. Mason and J. P. Overington (2006). "Can we rationally design promiscuous drugs?" Curr Opin Struct Biol **16**(1): 127-136.

Keiser, M. J. and J. Hert (2009). "Off-target networks derived from ligand set similarity." Methods Mol Biol **575**: 195-205.

Keiser, M. J., V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth (2009). "Predicting new molecular targets for known drugs." Nature **462**(7270): 175-181.

Kellenberger, E., J. Rodrigo, P. Muller and D. Rognan (2004). "Comparative evaluation of eight docking tools for docking and virtual screening accuracy." Proteins **57**(2): 225-242.

Kinnings, S. L., N. Liu, N. Buchmeier, P. J. Tonge, L. Xie and P. E. Bourne (2009). "Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis." PLoS Comput Biol **5**(7): e1000423.

Korb, O., T. Stutzle and T. E. Exner (2009). "Empirical scoring functions for advanced protein-ligand docking with PLANTS." J Chem Inf Model **49**(1): 84-96.

Leach, A. R., V. J. Gillet, R. A. Lewis and R. Taylor (2011). "Three-dimensional pharmacophore methods in drug discovery." J Med Chem **53**(2): 539-558.

Li, Y. Y., J. An and S. J. Jones (2011). "A computational approach to finding novel targets for existing drugs." PLoS Comput Biol **7**(9): e1002139.

Liu, X., S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, S. Zheng, Z. Li, H. Li and H. Jiang (2010). "PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach." Nucleic Acids Res **38**(Web Server issue): W609-614.

MacPherson, I. S., S. Kirubakaran, S. K. Gorla, T. V. Riera, J. A. D'Aquino, M. Zhang, G. D. Cuny and L. Hedstrom (2010). "The Structural Basis of Cryptosporidium-Specific IMP Dehydrogenase Inhibitor Selectivity." J Am Chem Soc **132**(4): 1230-1231.

Marcou, G. and D. Rognan (2007). "Optimizing fragment and scaffold docking by use of molecular interaction fingerprints." J Chem Inf Model **47**(1): 195-207.

Markt, P., D. Schuster, J. Kirchmair, C. Laggner and T. Langer (2007). "Pharmacophore modeling and parallel screening for PPAR ligands." J Comput Aided Mol Des **21**(10-11): 575-590.

Martin, R. E., L. G. Green, W. Guba, N. Kratochwil and A. Christ (2007). "Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach." J Med Chem **50**(25): 6291-6294.

Meslamani, J., D. Rognan and E. Kellenberger (2011). "sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins." Bioinformatics **27**(9): 1324-1326.

Morphy, R. (2010). "Selectively nonselective kinase inhibition: striking the right balance." J Med Chem **53**(4): 1413-1437.

Muller, P., G. Lena, E. Boilard, S. Bezzine, G. Lambeau, G. Guichard and D. Rognan (2006). "In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors." J Med Chem **49**(23): 6768-6778.

Novikov, F. N., A. A. Zeifman, O. V. Stroganov, V. S. Stroylov, V. Kulkov and G. G. Chilov (2011). "CSAR scoring challenge reveals the need for new concepts in estimating protein-ligand binding affinity." J Chem Inf Model **51**(9): 2090-2096.

OpenEye (2012). "Filter v.2.0.2; OpenEye Scientific Software: Santa Fe, NM 87507."

OpenEye (2012). "ROCS v.3.1.2; OpenEye Scientific Software: Santa Fe, NM 87507."

Rogers, D. and M. Hahn (2010). "Extended-connectivity fingerprints." J Chem Inf Model **50**(5): 742-754.

Rogers, D. and A. J. Hopfinger (1994). "Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships." J Chem Inf Comput Sci **34**(4): 854-866.

Rognan, D. (2007). "Chemogenomic approaches to rational drug design." Br J Pharmacol **152**(1): 38-52.

Rognan, D. (2010). "Structure-based approaches to target fishing and ligand profiling." Mol Inf **29**: 176-187.

Rollinger, J. M., D. Schuster, B. Danzl, S. Schwaiger, P. Markt, M. Schmidtke, J. Gertsch, S. Raduner, G. Wolber, T. Langer and H. Stuppner (2009). "In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens." Planta Med **75**(3): 195-204.

Schneider, G. (2010). "Virtual screening: an endless staircase?" Nat Rev Drug Discov **9**(4): 273-276.

Schneider, N., S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, K. Beyer, H. Claussen, M. Gastreich, C. Lemmen and M. Rarey (2011). "Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function." J Comput Aided Mol Des.

Smith, R. D., J. B. Dunbar, Jr., P. M. Ung, E. X. Esposito, C. Y. Yang, S. Wang and H. A. Carlson (2011). "CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions." J Chem Inf Model **51**(9): 2115-2131.

Spitzer, G. M., M. Heiss, M. Mangold, P. Markt, J. Kirchmair, G. Wolber and K. R. Liedl (2010). "One concept, three implementations of 3D pharmacophore-based virtual screening: distinct coverage of chemical search space." J Chem Inf Model **50**(7): 1241-1247.

Spitzer, R. and A. N. Jain (2012). "Surflex-Dock: Docking benchmarks and real-world application." J Comput Aided Mol Des.

Steindl, T. M., D. Schuster, C. Laggner, K. Chuang, R. D. Hoffmann and T. Langer (2007). "Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models." J Chem Inf Model **47**(2): 563-571.

Steindl, T. M., D. Schuster, C. Laggner and T. Langer (2006). "Parallel screening: a novel concept in pharmacophore modeling and virtual screening." J Chem Inf Model **46**(5): 2146-2157.

Steuber, H., M. Zentgraf, C. La Motta, S. Sartini, A. Heine and G. Klebe (2007). "Evidence for a novel binding site conformer of aldose reductase in ligand-bound state." J Mol Biol **369**(1): 186-197.

Surgand, J. S., J. Rodrigo, E. Kellenberger and D. Rognan (2006). "A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors." Proteins **62**(2): 509-538.

Sutter, J., J. Li, A. J. Maynard, A. Goupil, T. Luu and K. Nadassy (2011). "New features that improve the pharmacophore tools from Accelrys." Curr Comput Aided Drug Des **7**(3): 173-180.

van Westen, G. J. P., J. K. Wegner, A. P. Ijzerman, H. W. T. van Vlijmen and A. Bender (2011). "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets." MedChemComm **2**: 16-30.

Vidal, D., R. Garcia-Serna and J. Mestres (2010). "Ligand-based approaches to in silico pharmacology." Methods Mol Biol **672**: 489-502.

Wang, F., D. Liu, H. Wang, C. Luo, M. Zheng, H. Liu, W. Zhu, X. Luo, J. Zhang and H. Jiang (2011). "Computational screening for active compounds targeting protein sequences: methodology and experimental validation." J Chem Inf Model **51**(11): 2821-2828.

Wang, W., X. Zhou, W. He, Y. Fan, Y. Chen and X. Chen (2011). "The interprotein scoring noises in glide docking scores." Proteins **80**(1): 169-183.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant (2011). "PubChem's BioAssay Database." Nucleic Acids Res.

Weill, N. and D. Rognan (2010). "Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites." J Chem Inf Model **50**(1): 123-135.

Wolber, G. and T. Langer (2005). "LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters." J Chem Inf Model **45**(1): 160-169.

Wolber, G., T. Seidel, F. Bendix and T. Langer (2008). "Molecule-pharmacophore superpositioning and pattern matching in computational drug design." Drug Discov Today **13**(1-2): 23-29.

Xie, L. and P. E. Bourne (2008). "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments." Proc Natl Acad Sci U S A **105**(14): 5441-5446.

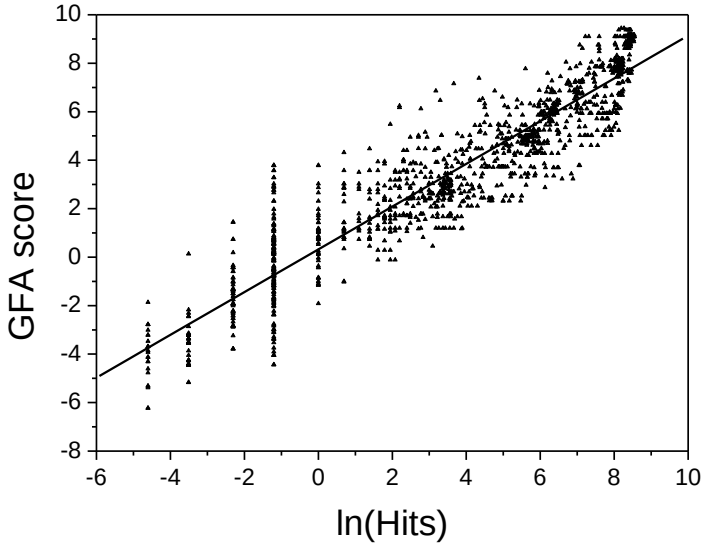
Yang, L., J. Chen and L. He (2009). "Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome." PLoS Comput Biol **5**(7): e1000441.

Yang, L., K. Wang, J. Chen, A. G. Jegga, H. Luo, L. Shi, C. Wan, X. Guo, S. Qin, G. He, G. Feng and L. He (2011). "Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome--clozapine-induced agranulocytosis as a case study." PLoS Comput Biol **7**(3): e1002016.

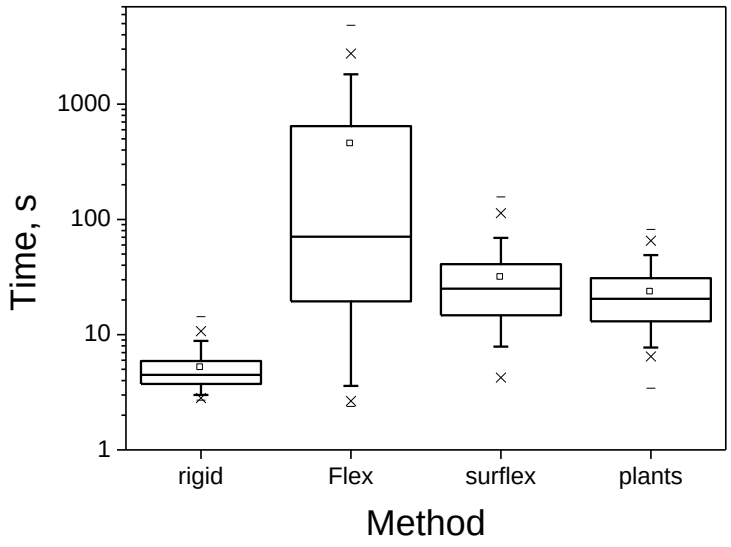
Yera, E. R., A. E. Cleves and A. N. Jain (2011). "Chemical structural novelty: on-targets and off-targets." J Med Chem **54**(19): 6771-6785.

9. Annexes

Supplementary Figure 1



Supplementary Figure 2



Supplementary Table 1: Pharmacophore descriptors used for training the GFA model to estimate selectivity

Descriptor Name	Descriptor Type
Desc1	Number of all features
Desc2	number of location features
Desc3	Has Shape (1) or not (0)
Desc4	Number of Excluded Volumes (EVs)
Desc5	Number of Hydrogen Bond Acceptors (HBAs)
Desc6	Number of Hydrogen Bond Donors (HBDs)
Desc7	Number of POSITIVE Ionizables
Desc8	Number of Negative Ionizables
Desc9	Number of Aromatic Ring features
Desc10	Number of Hydrophobic features
Desc11-Desc20	HBA-HBA feature distance [a]
Desc21-Desc30	HBA-HBD feature distance
Desc31-Desc40	HBA-POS feature distance
Desc41-Desc50	HBA-NEG feature distance
Desc51-Desc60	HBA-RA feature distance
Desc61-Desc70	HBA-HYD feature distance
Desc71-Desc80	HBD-HBD feature distance
Desc81-Desc90	HBD-POS feature distance
Desc91-Desc100	HBD-NEG feature distance
Desc101-Desc110	HBD-RA feature distance
Desc111-Desc120	HBD-HYD feature distance
Desc121-Desc130	POS-POS feature distance
Desc131-Desc140	POS-NEG feature distance
Desc141-Desc150	POS-RA feature distance
Desc151-Desc160	POS-HYD feature distance
Desc161-Desc170	NEG-NEG feature distance
Desc171-Desc180	NEG-RA feature distance
Desc181-Desc190	NEG-HYD feature distance
Desc191-Desc200	RA-RA feature distance
Desc201-Desc210	RA-HYD feature distance
Desc211-Desc220	HYD-HYD feature distance

[a] For each pair of feature types, the feature-feature distance is put into a distance bin (1-10), with a bin size of 2.0 Å. The distance bin count is used as the descriptor value. For instance, descriptors Desc11 to Desc20 are used for HBA-HBA distances. Desc11 is the number of HBA-HBA distances in the range 0-2.0 Å, Desc12 is the number of HBA-HBA distances in the range of 2.0-4.0 Å, and so on. Distances greater than 20.0 Å are counted in the last bin, i.e. Desc20. Similar descriptors are defined for the other types of feature-feature distances.

Supplementary Table 2: OpenEye Filter file used to select drug-like ligands

```
MIN_MOLWT      150        "Minimum molecular weight"
MAX_MOLWT      750        "Maximum molecular weight"
MIN_NUM_HVY    10         "Minimum number of heavy atoms"
MAX_NUM_HVY    55         "Maximum number of heavy atoms"
MIN_RING_SYS   0      "Minumum number of ring systems"
MAX_RING_SYS   7      "Maximum number of ring systems"
MIN_RING_SIZE  0      "Minimum atoms in any ring system"
MAX_RING_SIZE  20     "Maximum atoms in any ring system"
MIN_CON_NON_RING   0      "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING   20     "Maximum number of connected non-ring atoms"
MIN_FCNGRP     0      "Minimum number of functional groups"
MAX_FCNGRP     20     "Maximum number of functional groups"
MIN_UNBRANCHED 0      "Minimum number of connected unbranched non-ring atoms"
MAX_UNBRANCHED 8      "Maximum number of connected unbranched non-ring atoms"
MIN_CARBONS    5      "Minimum number of carbons"
MAX_CARBONS    40     "Maximum number of carbons"
MIN_HETEROATOMS    2      "Minimum number of heteroatoms"
MAX_HETEROATOMS    20     "Maximum number of heteroatoms"
MIN_Het_C_Ratio    0.10   "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio    1.0    "Maximum heteroatom to carbon ratio"
MIN_HALIDE_FRACTION    0.0    "Minimum Halide Fraction"
MAX_HALIDE_FRACTION    0.5    "Maximum Halide Fraction"
ADJUST_ROT_FOR_RING    true   "BOOLEAN for whether to estimate degrees of freedom in rings"
MIN_ROT_BONDS  0      "Minimum number of rotatable bonds"
MAX_ROT_BONDS  20     "Maximum number of rotatable bonds"
MIN_RIGID_BONDS    0      "Minimum number of rigid bonds"
MAX_RIGID_BONDS    50     "Maximum number of rigid bonds"
MIN_HBOND_DONORS 0    "Minimum number of hydrogen-bond donors"
MAX_HBOND_DONORS 6    "Maximum number of hydrogen-bond donors"
MIN_HBOND_ACCEPTORS 0    "Minimum number of hydrogen-bond acceptors"
MAX_HBOND_ACCEPTORS 10    "Maximum number of hydrogen-bond acceptors"
MIN_LIPINSKI_DONORS 0    "Minimum number of hydrogens on O & N atoms"
MAX_LIPINSKI_DONORS 5    "Maximum number of hydrogens on O & N atoms"
MIN_LIPINSKI_ACCEPTORS 0    "Minimum number of oxygen & nitrogen atoms"
MAX_LIPINSKI_ACCEPTORS 10    "Maximum number of oxygen & nitrogen atoms"
MIN_COUNT_FORMAL_CRG   0    "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG   3    "Maximum number of formal charges"
MIN_SUM_FORMAL_CRG -2    "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG 2    "Maximum sum of formal charges"
MIN_CHIRAL_CENTERS                  0                   "Minimum chiral centers"
MAX_CHIRAL_CENTERS                  4                   "Maximum chiral centers"
MIN_XLOGP      -2.0     "Minimum XLogP"
MAX_XLOGP      6.0      "Maximum XLogP"
MIN_SOLUBILITY     moderately     "Minimum solubility"
PSA_USE_SandP  true   "Count S and P as polar atoms"
MIN_2D_PSA     0.0      "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA     150.0    "Maximum 2-Dimensional (SMILES) Polar Surface Area"
AGGREGATORS    true     "Eliminate known aggregators"
PRED_AGG       false    "Eliminate predicted aggregators"
GSK_VEBER      true     "PSA>140 or >10 rot bonds"
MAX_LIPINSKI   2        "Maximum number of Lipinski violations"
MIN_ABS 0.5            "Minimum probability F>10% in rats"
PHARMACOPIA    true     "LogP > 5.88 or PSA > 131.6"
ALLOWED_ELEMENTS  H,C,N,O,F,S,Cl,Br,I,P
ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#specific, undesirable functional groups
RULE  0  quinone
RULE  0  pentafluorophenyl_esters
RULE  0  paranitrophenyl_esters
RULE  0  HOBT_esters
RULE  0  triflates
RULE  4  amino_acid
RULE  2  amine
RULE  4  primary_amine
RULE  4  secondary_amine
RULE  4  tertiary_amine
RULE  2  carboxylic_acid
```

RULE 0 lawesson_s_reagent
RULE 0 phosphoramides
RULE 0 beta_carbonyl_quat_nitrogen
RULE 0 acylhydrazide
RULE 0 cation_C_Cl_I_P_or_S
RULE 0 phosphoryl
RULE 0 alkyl_phosphate
RULE 0 phosphinic_acid
RULE 0 phosphanes
RULE 0 phosphoranes
RULE 0 imidoyl_chlorides
RULE 0 nitroso
RULE 0 N_P_S_Halides
RULE 0 carbodiimide
RULE 0 isonitrile
RULE 0 triacyloxime
RULE 0 cyanohydrins
RULE 0 acyl_cyanides
RULE 0 sulfonylnitrile
RULE 0 phosphonylnitrile
RULE 0 azocyanamides
RULE 0 beta_azo_carbonyl
RULE 0 polyenes
RULE 0 saponin_derivatives
RULE 1 cytochalasin_derivatives
RULE 4 cycloheximide_derivatives
RULE 1 monensin_derivatives
RULE 1 squalestatin_derivatives
RULE 0 acid_halide
RULE 0 aldehyde
RULE 0 alkyl_halide
RULE 0 anhydride
RULE 0 azide
RULE 0 azo
RULE 0 di_peptide
RULE 0 michael_acceptor
RULE 0 beta_halo_carbonyl
RULE 0 nitro
RULE 0 oxygen_cation
RULE 0 peroxide
RULE 0 phosphonic_acid
RULE 0 phosphonic_ester
RULE 0 phosphoric_acid
RULE 0 phosphoric_ester
RULE 0 sulfonic_acid
RULE 0 sulfonic_ester
RULE 6 halide
RULE 0 iodine
RULE 2 ketone
RULE 4 phenol
RULE 1 imine
RULE 1 methyl_ketone
RULE 1 alkylaniline
RULE 4 sulfonamide
RULE 1 sulfonylurea
RULE 0 phosphonamide
RULE 0 alphahalo_ketone
RULE 0 oxaziridine
RULE 1 cyclopropyl
RULE 2 guanidine
RULE 0 sulfonimine
RULE 0 sulfinimine
RULE 1 hydroxamic_acid
RULE 0 sulfinylthio
RULE 0 disulfide
RULE 0 enol_ether
RULE 0 enamine
RULE 0 organometallic
RULE 0 dithioacetal
RULE 1 oxime
RULE 0 isothiocyanate
RULE 0 isocyanate
RULE 3 lactone
RULE 3 lactam
RULE 1 thioester
RULE 1 carbonate
RULE 0 carbamic_acid
RULE 1 thiocarbamate
RULE 0 triazine
RULE 1 malonic
#other functional groups
RULE 2 alkyne
RULE 4 aniline
RULE 4 aryl_halide
RULE 2 carbamate
RULE 3 ester
RULE 5 ether
RULE 1 hydrazone
RULE 0 nonacylhydrazone
RULE 1 hydroxylamine
RULE 2 nitrile
RULE 2 sulfide

RULE 0 tricarbo_phosphene
RULE 0 epoxide
RULE 0 sulfonyl_halide
RULE 0 halopyrimidine
RULE 0 perhalo_ketone
RULE 0 aziridine
RULE 1 oxalyl
RULE 0 alphahalo_amine
RULE 0 halo_amine
RULE 0 halo_alkene
RULE 0 acyclic_NCN
RULE 0 acyclic_NS
RULE 0 SCN2
RULE 0 terminal_vinyl
RULE 0 hetero_hetero
RULE 0 hydrazine
RULE 0 N_methoyl
RULE 0 NS_beta_halothyl
RULE 0 propiolactones
RULE 0 iodoso
RULE 0 iodoxy
RULE 0 noxide
#groups of molecules
RULE 0 dye
#common functional groups
RULE 6 alcohol
RULE 4 alkene
RULE 4 amide
RULE 2 sulfone
RULE 2 sulfoxide
RULE 0 thiourea
RULE 1 thioamide
RULE 1 thiol
RULE 2 urea
RULE 0 hemiketal
RULE 0 hemiacetal
RULE 0 ketal
RULE 1 acetal
RULE 0 aminal
RULE 0 hemiaminal
#protecting groups
RULE 0 benzyloxycarbonyl_CBZ
RULE 0 t_butoxycarbonyl_tBOC
RULE 0 fluorenylmethoxycarbonyl_Fmoc
RULE 1 dioxolane_5MR
RULE 1 dioxane_6MR
RULE 1 tetrahydropyran_THP
RULE 1 methoxyethoxymethyl_MEM
RULE 2 benzyl_ether
RULE 2 t_butyl_ether
RULE 0 trimethylsilyl_TMS
RULE 0 t_butyldimethylsilyl_TBDMS
RULE 0 triisopropylsilyl_TIPS
RULE 0 t_butyldiphenylsilyl_TBDPS
RULE 1 phthalimides_PHT
RULE 2 arenesulfonyl

Supplementary Table 3: List of HET codes for the 157 ligands of the sc-PDB Diverse Set and their respective targets

HET_CODE	Targets
115	3-hydroxy-3-methylglutaryl-coenzyme a reductase
1CS	acetolactate synthase catalytic subunit
20A	camp-specific 3',5'-cyclic phosphodiesterase 4b
215	serine/threonine-protein kinase b-raf
2IG	Renin
356	dipeptidyl peptidase 4
3B9	dna ligase
3CC	carbonic anhydrase 2
3LP	glycylpeptide n-tetradecanoyltransferase
3QC	kinesin-like protein kif11
501	thrombin
521	tyrosine-protein phosphatase non-receptor type 1
55V	trimethoprim-sensitive dihydrofolate reductase
5RM	camp-specific 3',5'-cyclic phosphodiesterase 4b
61	metallo beta-lactamase
669	3-oxoacyl-[acyl-carrier-protein] synthase 3
675	urokinase-type plasminogen activator
6C3	macrophage colony-stimulating factor 1 receptor
760	acetylcholinesterase
783	trypsin
792	disintegrin and metalloproteinase domain-containing protein 17
839	wee1-like protein kinase
87Y	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
905	coagulation factor vii
915	lethal factor
961	retinoic acid receptor gamma
984	mitogen-activated protein kinase 10
A3M	dipeptidyl peptidase 4
A46	glycogen phosphorylase
A80	cellular retinoic acid-binding protein 1
AD3	Adenosylhomocysteinase iag-nucleoside hydrolase
AEE	epidermal growth factor receptor

AH1	protease
ALJ	chitinase
AO5	methionine aminopeptidase 2
AXX	probable serine/threonine-protein kinase pkng
AZM	carbonic anhydrase 9
	carbonic anhydrase 12
AZZ	deoxynucleoside kinase
	glycogen phosphorylase
	serum albumin
BAU	uridine phosphorylase 1
	uridine phosphorylase
BBT	neutrophil collagenase
BCZ	Neuraminidase
	sialidase-2
BDI	queuine trna-ribosyltransferase
BFS	scytalone dehydratase
BHY	glutamate receptor 2
BIG	5'-methylthioadenosine/s-adenosylhomocysteine nucleosidase
BIT	myosin-2 heavy chain
BRZ	glyceraldehyde-3-phosphate dehydrogenase
C4C	focal adhesion kinase 1
CA2	3-dehydroquinate dehydratase
CBT	beta-lactamase tem
CEI	Obelin
	renilla-luciferin 2-monooxygenase
CEL	prostaglandin g/h synthase 1
	carbonic anhydrase 2
CIA	cgmp-specific 3',5'-cyclic phosphodiesterase
CMB	coagulation factor x
CMF	RNA-directed RNA polymerase
CMU	thymidine phosphorylase
CRZ	fatty acid-binding protein
CT5	atp-dependent molecular chaperone hsp82
	heat shock protein hsp 90-alpha
D1L	acetyl-coa carboxylase
DBQ	activated cdc42 kinase 1

DD2	scavenger mrna-decapping enzyme dcps
DEO	macrophage metalloelastase
DES	estrogen-related receptor gamma
	estrogen receptor alpha
	transthyretin
DEX	glucocorticoid receptor
DZG	pyruvate kinase isozymes m1/m2
DZP	serum albumin
E4D	estrogen receptor alpha
E89	queuine trna-ribosyltransferase
ED2	protein farnesyltransferase/geranylgeranyltransferase type-1 subunit alpha
EI1	estrogen receptor alpha
EQI	estradiol 17-beta-dehydrogenase 1
ET	bacterial regulatory protein, tetr family
	hth-type transcriptional regulator qacr
F13	beta-lactamase
FDI	neuraminidase
FLP	cytochrome p450 2c9
FR4	adenosine deaminase
FRG	interleukin-2
FSN	thrombin
GB7	alpha-mannosidase 2
GEO	deoxynucleoside kinase
	deoxycytidine kinase
GIO	chitinase
GNT	Acetylcholinesterase
	soluble acetylcholine receptor
GRR	peroxisome proliferator-activated receptor gamma
GVR	udp-3-o-[3-hydroxymyristoyl] n-acetylglucosamine deacetylase
H11	corticosteroid 11-beta-dehydrogenase isozyme 1
H24	beta-secretase 1
H7J	cathepsin s
HA3	histone deacetylase 4
HEF	reverse transcriptase/ribonuclease h
HM5	methionine aminopeptidase 1
HR2	3-hydroxy-3-methylglutaryl-coenzyme a reductase

HUP	acetylcholinesterase
I84	aldose reductase
IAD	tryptophan synthase alpha chain
IC1	casein kinase i homolog 1
IMN	prostaglandin reductase 2 aldo-keto reductase family 1 member c3 prostaglandin g/h synthase 2 phospholipase a2
IMQ	iag-nucleoside hydrolase
IXM	cell division protein kinase 5 cdc2-like cdk2/cdc28 like protein kinase glycogen synthase kinase-3 beta
KAI	glutamate receptor, ionotropic kainate 2 glutamate receptor 4
LG7	androgen receptor
LI9	mitogen-activated protein kinase 14
LQQ	cyclin homolog
LS1	cell division protein kinase 2
MC9	vitamin d3 receptor
MD7	dihydroorotate dehydrogenase
MTI	purine nucleotide phosphorylase
NDR	progesterone receptor
NGH	stromelysin-2 macrophage metalloelastase
OA1	biotin carboxylase
OEF	thyroid hormone receptor beta
P1S	sam dependent isoflavanone 4'-o-methyltransferase/(+)-6a-hydroxymaackiain-3-0-methyltransferase
P1Z	prostaglandin reductase 2
P21	bifunctional protein glmu//udp-n-acetylglucosamine pyrophosphorylase/glucosamine-1-phosphate n-acetyltransferase
P34	poly [adp-ribose] polymerase 15 hypothetical exotoxin a
P4A	[pyruvate dehydrogenase [lipoamide]] kinase isozyme 4
PAF	pantothenate synthetase
PAU	pantothenate kinase

PBF	tyrosyl-trna synthetase
PFP	serine/threonine-protein kinase chk1
PH7	rna-directed rna polymerase
PM2	tryptase beta-2
PVB	cell division control protein 2 homolog
	uncharacterized protein srpk2
R6C	pyridoxal kinase
R78	serine/threonine-protein kinase plk1
R88	squalene--hopene cyclase
RNP	Lactotransferrin
	cellulase
ROF	camp-specific 3',5'-cyclic phosphodiesterase 4d
	camp-specific 3',5'-cyclic phosphodiesterase 4b
RQ3	hemoglobin subunit beta/hemoglobin subunit alpha
RRC	cell division protein kinase 2
	pyridoxal kinase
	cell division protein kinase 5
RXC	caspase-3
S22	dual specificity protein kinase ttk
SAG	amine oxidase [flavin-containing] b
SB8	peptide deformylase
SCT	thymidine kinase
SHM	Streptavidin
	avidin
SLX	reticuline oxidase
STC	beta-lactamase
SWA	alpha-1,2-mannosidase
	alpha-mannosidase 2
	putative alpha-1,2-mannosidase
T74	activated cdc42 kinase 1
TCD	retinoic acid receptor RXR-alpha
TDZ	fatty acid-binding protein
THM	thymidylate kinase
	glycogen phosphorylase
TIM	beta-2 adrenergic receptor
TNK	HIV-1 RT alpha-chain

TPR	thymidylate synthase
TSX	phosphoenolpyruvate carboxykinase
TTT	3c-like proteinase
TYR	prephenate dehydrogenase
	ribonucleoside-diphosphate reductase 1 subunit alpha
	chorismate mutase
	tyrosyl-trna synthetase
	cysteine synthase
VDN	camp-specific 3',5'-cyclic phosphodiesterase 4b
	cgmp-specific 3',5'-cyclic phosphodiesterase
VGA	glutamate racemase
VGB	o-glcnacase nagj
VGG	calmodulin-domain protein kinase 1
VIB	putative hmp/thiamine-binding protein ykof
	thiamine pyrophosphokinase
	thiamin pyrophosphokinase 1
XM5	ribosyldihydronicotinamide dehydrogenase [quinone]
XX5	angiotensin-converting enzyme 2
ZAM	enoyl-[acyl-carrier-protein] reductase [NADH]
ZMA	adenosine receptor alpha2a

Supplementary Table 4: Profiling accuracy of 128 ligands (Set 2) by 8 computational protocols: successful, target rank ≤ 25, green; ambiguous, 25 < target rank ≤ 50, orange; failed, target rank < 50, red. Numbers in cells indicate the rank of top-ranked true target entry.

HET	Rigid1	Rigid2	Flex1	Flex2	Surflex1	Surflex2	Plants1	Plants2
115	1	1	1	1	2	1	266	61
1CS	40	15	9	6	120	6	390	18
20A	50	21	29	11	188	19	484	37
215	1	1	1	1	189	12	177	9
2IG	1	1	7	4	2	1	248	15
356	9	10	48	22	180	28	310	118
3CC	1	1	1	1	314	153	916	89
3LP	605	803	670	735	1684	309	1100	361
3QC	1	1	1	1	55	3	52	4
501	18	12	1	1	14	149	98	26
521	20	10	2	1	6	1	54	6
5RM	1	1	1	1	110	10	17	3
61	47	10	118	25	593	449	173	194
669	98	36	494	139	1668	340	1121	245
675	1	1	1	1	107	7	46	6
6C3	13	18	15	20	118	15	44	4
760	387	291	360	141	54	73	12	210
783	3	5	6	3	23	3	10	3
792	16	7	31	11	223	28	370	68
839	1	1	1	1	44	5	85	13
87Y	1	1	1	1	1	1	1	1
905	2	2	5	4	27	2	5	1
915	2	2	1	1	21	5	289	20
961	1	1	1	1	1	1	1	1
984	4	2	6	2	6	1	94	8
A3M	28	9	48	16	1	1	125	9
A46	16	7	20	9	46	5	143	14
A80	9	4	3	3	70	7	169	23
AEE	1	1	1	1	9	1	302	19
AH1	1	3	1	1	6	3	1	1
ALJ	6	3	27	9	166	7	139	11
AO5	39	18	96	42	156	8	548	155
AXX	1	1	2	1	74	5	465	53
AZM	26	2	1	1	2283	228	1692	150
BBT	64	10	55	38	924	58	328	16

HET	Rigid1	Rigid2	Flex1	Flex2	Surflex1	Surflex2	Plants1	Plants2
BDI	15	3	20	4	15	5	97	35
BFS	1	1	1	1	368	20	451	67
BHY	280	258	93	261	22	4	34	6
BIG	1	1	2	3	1	1	3	2
BIT	3	1	1	1	1	1	9	1
BRZ	276	321	347	170	817	283	1398	338
C4C	59	20	85	28	743	34	1202	121
CA2	1	1	1	1	2	1	26	2
CBT	4	3	7	5	917	244	1554	262
CIA	42	17	9	42	58	11	22	3
CMB	45	17	43	24	49	7	125	10
CMF	2	2	1	1	40	2	158	18
CMU	1	1	1	1	72	7	214	21
CRZ	28	8	1	9	1767	183	1480	199
D1L	1	1	1	1	65	5	89	15
DBQ	1	1	1	1	533	58	376	256
DD2	12	5	337	110	16	1	5	3
DEO	79	17	1	1	43	3	105	17
DEX	1	1	1	1	2	1	3	2
DZG	619	135	120	24	407	438	18	534
DZP	1	1	1	1	298	123	319	229
E4D	1	1	1	1	46	3	292	16
E89	1	1	1	1	1	1	3	1
ED2	271	121	547	246	221	155	410	384
EI1	1	1	1	1	9	2	87	16
EQI	4	4	4	4	908	120	291	42
F13	272	50	256	45	410	187	1037	239
FDI	4	13	5	17	33	10	1249	133
FLP	1	1	3	1	462	69	831	183
FR4	4	2	11	2	54	96	504	76
FRG	95	84	87	28	833	56	884	73
FSN	1	2	1	2	19	4	12	3
GB7	1	1	1	1	255	134	872	69
GIO	23	7	23	7	2004	110	172	11
GRR	3	28	15	23	286	217	255	48
GVR	225	78	517	187	39	9	1963	301
H11	1	1	1	1	116	3	31	12
H24	27	101	45	29	59	2	641	48
H7J	46	17	177	62	1812	283	421	118
HA3	273	271	668	222	2	1	39	320
HEF	12	3	38	13	70	12	1	2

HET	Rigid1	Rigid2	Flex1	Flex2	Surflex1	Surflex2	Plants1	Plants2
HM5	177	77	283	111	1256	242	335	29
HR2	1	1	1	1	2	1	18	2
HUP	508	299	535	295	29	61	1	157
IAD	54	12	49	15	7	2	198	40
IC1	3	2	3	2	407	138	245	33
IMQ	2	1	4	2	3	5	1	1
KAI	1	1	1	1	20	1	196	17
LG7	1	1	6	1	36	6	9	3
LI9	1	1	1	1	7	2	96	13
LQQ	2	1	1	1	247	105	97	8
LS1	1	1	1	1	14	5	227	15
MC9	1	1	1	1	11	6	90	21
MD7	1	1	1	1	87	15	65	29
MTI	235	54	212	53	96	19	435	86
NDR	4	3	5	3	117	9	23	5
OA1	1	1	1	1	7	2	813	108
OEF	2	2	2	2	26	4	3	1
P1S	73	18	239	56	1799	322	1692	407
P1Z	81	28	26	9	1085	180	46	237
P21	1	1	6	1	677	106	1876	302
P4A	125	44	445	165	61	365	1238	467
PAF	433	205	3	170	166	72	518	135
PBF	147	30	225	42	391	54	746	109
PFP	4	4	7	4	795	88	98	29
PH7	7	1	57	16	141	109	352	224
PM2	385	110	406	154	269	21	1519	185
R6C	9	2	37	14	58	30	330	36
R78	42	51	1	4	69	5	75	9
R88	2	2	3	2	533	325	59	55
RQ3	66	26	96	40	2266	508	1835	641
RXC	97	445	292	402	1301	459	1738	536
S22	2	2	12	8	65	118	1160	102
SAG	985	316	847	244	71	335	14	26
SB8	2	8	13	20	27	3	1040	148
SCT	1	1	1	1	4	2	70	13
SLX	1	1	1	1	57	2	31	3
STC	8	6	1	1	365	90	276	186
T74	5	5	1	1	2	1	27	2
TCD	6	6	7	6	106	273	4	212
TDZ	35	16	21	18	464	46	963	218
TIM	1	1	1	5	11	1	6	1

HET	Rigid1	Rigid2	Flex1	Flex2	Surflex1	Surflex2	Plants1	Plants2
TNK	2	2	3	2	3	1	1	1
TPR	10	6	2	1	86	255	113	225
TSX	1	1	1	1	1	11	7	6
TTT	15	7	2	1	1095	262	2	1
VGA	915	448	469	114	1518	152	1587	172
VGB	1	1	1	1	169	19	212	21
VGG	1	1	1	1	3	3	130	13
XM5	1019	1472	1522	1820	1004	232	845	357
XX5	76	19	20	5	394	33	1097	746
ZAM	89	34	79	31	187	150	12	4
ZMA	169	49	521	123	995	85	1379	174

Chapitre 5 :

Développement d’un protocole hybride de profilage

1. Introduction

La disponibilité des données de bioactivité a contribué au développement des méthodes de profilage virtuel dans le but de prédire les associations possibles entre molécules et cibles à grand échelle. Ce profilage permet d'obtenir une liste de cibles potentielles à travers lesquelles on peut évaluer la promiscuité de la molécule, identifier des cibles associées à des effets pharmacologiques et en identifier d'autres pour anticiper d'éventuels effets secondaires ou toxicité.

Plusieurs approches existent pour obtenir un profil biologique d'une molécule (Rognan 2012) et certains groupes mettent à la disposition de la communauté scientifique des interfaces web pour pouvoir effectuer le profilage. C'est l'exemple du portail SEA (Keiser *et al.* 2007) et du portail ChemProt (Taboureau *et al.* 2011). Les deux approches utilisent des méthodes basées sur la similarité des ligands connus pour chaque cible. Dès lors, ces approches ne s'appliquent pas aux cibles pour lesquelles peu de ligands sont connus ou aux cibles orphelines.

Nous avons comparé les performances de certaines approches de criblage virtuel dans le cadre d'un profilage sur 2556 cibles (Meslamani *et al.* 2012) issues de la base sc-PDB (Kellenberger *et al.* 2006; Meslamani *et al.* 2011). Il en ressort de cette étude que certaines cibles sont plus faciles à identifier en employant certaines approches. Les méthodes basées sur les ligands sont performantes lorsque la cible possède plusieurs ligands connus. Concernant les approches basées sur la structure de la cible, l'arrimage était performant lorsque le site de liaison était polaire et enfoui. L'utilisation des pharmacophores d'interactions présentait des bons résultats pour les autres cas.

Dès lors, nous avons fait le choix de générer un flux automatique de profilage, qui a pour but de sélectionner la méthode de criblage virtuel la mieux adaptée afin d'évaluer la possibilité d'association entre une molécule et une cible.

Afin d'obtenir un profilage sur un large panel de cibles, trois bases de données de bioactivité ont été utilisées pour extraire tous les ligands connus. Les informations d'interactions ont été récupérées des bases ChEMBL (Gaulton *et al.* 2011), PubChem BioAssay (Wang *et al.* 2011) et IUPHAR-DB (Mpamhanga *et al.* 2012). Les cibles de la base de complexes cristallographiques sc-PDB ont été aussi incluses. Ceci permet d'intégrer des cibles pour lesquelles très peu de ligands sont connus.

2. Matériels et méthodes :

2.1. Préparation du jeu d'entrainement

Trois bases de données de bioactivité ont été utilisées pour extraire toutes les valeurs d'affinités possibles exprimées en K_i, K_d et IC_{50}. Ces bases sont ChEMBL, PubChem BioAssay et IUPHAR-DB. Tous les ligands des cibles répertoriées sont extraits avec la valeur d'affinité correspondante. Au démarrage du projet, PubChem BioAssay incluait la version 11 de la base ChEMBL et n'incluait pas la dernière mise à jour de la base IUPHAR-DB. Par conséquent, les trois bases ont été téléchargées et les données de chacune ont été traitées séparément.

2.1.1. ChEMBL

La version 12 de ChEMBL a été utilisée. Les associations protéine-ligand pour lesquelles une valeur d'affinité est présente sont extraites. Les affinités dont le seuil de confiance est supérieur à 7 ont été retenues. Ce seuil nous permet de garder les valeurs d'affinités pour lesquelles l'annotation de la cible est fiable. Le format *smiles* a été utilisé pour représenter les molécules, le code d'accès Uniprot (UniProt 2012) pour extraire le nom et l'annotation de la cible à partir de la base Uniprot et les affinités ont été exprimées en pK_i, pK_d et pIC_{50}. La base a été téléchargée sous format sql réservé au système de gestion de base de données MySQL (https://mysql.com). Des scripts Perl via le module DBI (http://search.cpan.org/~timb/DBI-1.622/DBI.pm) destinés à faciliter l'interfaçage aux bases de données, ont servi à extraire les informations et faire le traitement des données.

2.1.2. PubChem BioAssay

Tous les essais de la base PubChem BioAssay ont été téléchargés le 16 Décembre 2011. Un total de 586 272 d'essais a été ainsi traité. Les molécules, les cibles et les valeurs d'affinités ont été extraites de ces fichiers d'essais. Les molécules sont retenues à l'aide de leur numéro SID de la base PubChem Substance (Li *et al.* 2010). La structure est par la suite

téléchargée à partir de cette base. Les cibles ont été retenues grâce à leur numéro GI. Afin d'obtenir une annotation fiable et cohérente avec les données téléchargées depuis les autres bases, une table de correspondance a été utilisée afin de récupérer les numéros d'accession Uniprot correspondant aux numéros GI. Cette table est disponible sur cette adresse (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz). Les cibles qui ne possèdent pas d'identifiants GI dans la table ont été écartées de l'analyse. Les valeurs d'affinités ont tous été exprimées en pK_i, pK_d et pIC_{50}.
Le format XML a été choisi pour télécharger la base et un script Perl utilisant le module XML::Twig (http://search.cpan.org/~mirod/XML-Twig-3.40/Twig.pm) a été utilisé pour extraire les informations d'affinités.

2.1.3. IUPHAR-DB

La version du mois de Décembre 2011 de la base a été utilisée. Les molécules ont été extraite en chaine *smiles* et les protéines identifiées grâce à leur numéro d'accession Uniprot. Les affinités sont exprimées en pK_i, pK_d et pIC_{50}. Des scripts Perl ont été utilisés pour extraire les informations de la base PostgreSQL (www.postgresql.org).

2.1.4. sc-PDB

La version 2011 de la base sc-PDB a été utilisée. Celle-ci contient 9 877 entrées PDB qui incluent 3 034 cibles différentes. Chaque cible est annotée selon l'annotation Uniprot. Toutes les structures des cibles présentes vont être utilisées dans notre jeu d'entrainement pour le profilage à l'aide des méthodes basées sur les structures.

2.1.5. DrugBank

Afin de tester notre protocole de profilage, un jeu de validation externe a été choisi à partir de la base DrugBank 3.0 (Knox *et al.* 2011). Chaque médicament répertorié dans cette base possède une liste de cibles connues. La performance de notre protocole sera évaluée par le pourcentage de cibles retrouvé par le protocole pour chaque médicament du jeu de test externe choisi. Les associations médicament-cibles sont extraites à partir des fichiers XML de

la base à l'aide d'un script Perl. Les molécules sont générées à partir du fichier SDF proposé sur le site web de la base et les cibles sont identifiées avec leur numéro d'accession Uniprot.

2.1.6. Préparation des données du jeu d'entrainement

a) Préparation pour les méthodes basées sur la similarité 2D

Les molécules ont été standardisées de façon à éliminer la stéréo-isomérie à l'aide du programme Standardizer (ChemAxon 2011). Ceci permet d'éliminer les doublons car des empreintes circulaires 2D de graphe moléculaire ECFP_4 (Rogers *et al.* 2010) vont être utilisés. A l'aide du même outil, les solvants ont été supprimés et lorsque deux molécules étaient présentes, celle possédant la masse moléculaire la plus grande a été retenue. A l'aide du programme Filter (OpenEye 2011) les molécules qui ne possédaient pas d'atomes de Carbone ont été exclues. Les cycles de plus de 25 atomes ont été écartés dans le but d'exclure les molécules macrocycliques. Ensuite, les complexes métalliques ont été à leur tour écartés en spécifiant une liste de métaux à exclure. En vue de ne retenir que les molécules non peptidiques, une chaîne *smarts* ainsi qu'un seuil maximal de 12 angles de torsion ont été définis pour supprimer les molécules possédant plus de deux liaisons peptidiques et les longues chaines aliphatiques (hydrocarbures). Afin d'obtenir une représentation unique et homogène pour toutes les molécules issues des différentes bases, le tautomère le plus abondant pour chaque molécule a été généré à l'aide de l'application Quacpac (OpenEye 2011) et l'état d'ionisation défini à un pH physiologique de 7.4.

Les cibles ont toutes été représentées par leur numéro d'accession et leur nom Uniprot. Afin d'obtenir l'information sur l'espèce, les fichiers texte Uniprot correspondant à toutes les entrées ont été téléchargés et traités grâce à un script Perl.

Pour supprimer la redondance entre les valeurs d'affinités issues des trois bases, chaque association ligand-protéine-affinité a été définie par la chaine *smiles* du ligand, le nom Uniprot de la cible et l'affinité correspondante. Les doublons ont pu ainsi être écartés.

Si uniquement deux valeurs d'affinités différentes sont disponibles, la mesure la plus récente est retenue. La date correspond à celle de la publication qui est obtenue à partir de l'identifiant PubMed de celle-ci. Les détails des publications ont été acquis grâce à une

recherche sur Batch Entrez en spécifiant PubMed comme base de recherche (http://www.ncbi.nlm.nih.gov/sites/batchentrez).

Pour le cas des molécules ayant plusieurs valeurs d'affinités avec la même cible, la différence entre la valeur minimale pK_{min} et maximale pK_{max} est calculée. Si $\Delta|pK_{min} - pK_{max}| \leq 1$ alors la médiane est gardée comme valeur d'affinité, sinon l'association est rejetée. Ceci nous permet de retenir les associations avec des mesures fiables et de rejeter celles où les conditions expérimentales ou les erreurs de saisies sont probables.

En vue d'obtenir un nombre maximal de modèles pour des cibles humaines, celles qui ne possédaient pas plus de 25 ligands différents ont été supplémentées par des ligands de protéines orthologues. Si un ligand L_a de la protéine humaine P_H est testé sur une autre protéine orthologue (exemple : protéine bovine) P_O contenant des affinités pour des ligands L_O , et que : $\Delta|pK_i(L_H, P_H) - pK_i(L_H, P_O)| \leq 1$, alors les ligands L_O de la protéine P_O sont rajoutés à la protéine humaine P_H.

Les affinités récupérées des bases de bioactivité sont exprimées en pIC_{50}, pK_i et pK_d. Nous avons fait le choix de les considérer comme équivalentes. Nous stipulons que la variation des conditions de mesures n'affecte pas beaucoup la valeur d'affinité mesurée et que l'erreur qui peut exister est toutefois comparable à l'erreur expérimentale. Nous appellerons ces valeurs d'affinités pK_i dorénavant.

b) Préparation pour les méthodes basées sur la similarité 3D

La même procédure énoncée auparavant est appliquée à toutes les molécules extraites des trois bases de bioactivité, sauf que dans ce cas, la stéréo-isomérie est gardée. Les structures ont été générées à partir des chaines *smiles* et à l'aide du programme Corina (MolecularNetworks 2005).

2.1.7. Préparation des données du jeu de test

Les molécules extraites de la base DrugBank ont toutes été traitées de la même manière que les molécules des bases de bioactivité. Les chaines *smiles* ont servi à écarter les molécules qui étaient présentes dans l'ensemble d'entrainement. Les molécules retenues ont été soumises à une classification par un algorithme de graphe maximal commun à l'aide du logiciel MedChemStudio (SimulationsPlus 2012). La méthode choisie pour créer les classes fait appel aux fragments topologiques (énumération de tous les fragments possible entre 8 et 12 liaisons topologiques). Comme paramètres, nous avons choisi une homogénéité élevée (la taille du châssis moléculaire doit être proche de la taille de la molécule) et un niveau de redondance nul entre les classes (une molécule n'appartient qu'à une seule classe). Les cibles connues pour chaque molécule ont été soumises à une classification orthologique grâce à la base KEGG Orthology (Tanabe *et al.* 2012). Dès lors une sélection de molécules des classes de châssis différents a été effectuée tout en s'assurant que les protéines appartiennent à des familles différentes. Ce protocole permet d'obtenir un ensemble de validation hétérogène à la fois du point de vue moléculaire et du point de vue protéique afin d'évaluer au mieux les performances du protocole de profilage.

2.1.8. Stratégie de profilage

La figure 1 résume le fonctionnement du protocole automatique construit, appelé *Profiler*. La stratégie adoptée pour le profilage sur un grand nombre de cibles, consiste à utiliser une approche basée sur les ligands pour identifier les cibles possédant plus de 10 ligands différents. Dans le cas contraire, une méthode basée sur la structure de la cible (pharmacophore ou arrimage moléculaire) sera utilisée.

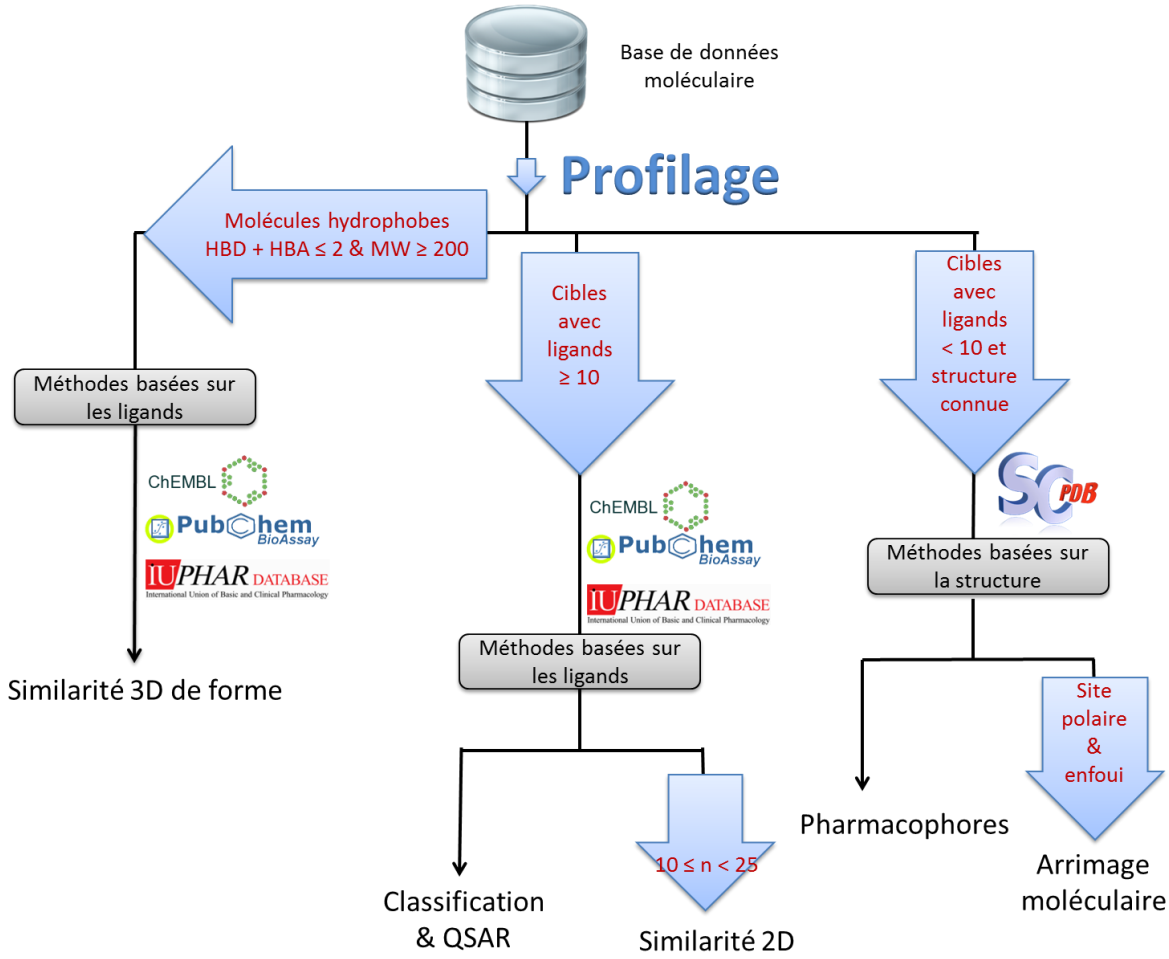


Figure 1 : Protocole de profilage automatique réalisé. Les molécules hydrophobes dont la masse moléculaire est supérieure à 200 et le nombre d'accepteur et donneurs de liaison H ≤ 2 sont profilées à l'aide de la similarité 3D de ROCS (OpenEye 2011). Pour les autres molécules, si la cible possède entre 10 et 25 ligands (nombre noté n), la similarité 2D à l'aide des empreintes ECFP_4 est évaluée. Dans les autres cas, l'association est prédite à partir d'un modèle de classification ou un modèle QSAR. Les cibles dont la structure 3D est disponible et qui possèdent moins de 10 ligands sont profilées à l'aide de l'arrimage si le site et polaire et enfoui, sinon à l'aide des pharmacophores.

a) Profilage par les modèles QSAR de régression et les modèles de classification

Nous avons décidé de profiler les molécules sur des cibles ayant plus de 25 ligands grâce à des modèles QSAR prédictifs et des modèles de classification SVM (la molécule prédite comme étant active ou inactive). Ces modèles sont développés pour chaque cible et à partir des ligands extraits des bases de données de bioactivité.

Les modèles QSAR ont été générés à l'aide des empreintes moléculaire ECFP_4 en utilisant les séparateurs à vaste marge avec le logiciel SVM^{light} 6.02 pour prédire les valeurs

d'affinité pKi. Ces modèles ont été validés grâce à une validation croisée *5-Fold* répétée trois fois, et ceux qui présentaient un coefficient de corrélation $Q^2 \geq 0.6$ ainsi qu'une MAE ≤ 1 (Tropsha 2010) ont été retenus (équations 1 et 2). Les modèles ne sont générés que pour les cibles de plus de 25 ligands avec une équidistribution des valeurs de leurs affinités sur les intervalles [4;6[, [6;8[et [8;10]. L'équidistribution est respectée si chaque intervalle dispose au minimum de 15% des données. Un seuil idéal de 33% aurait engendré l'élimination de beaucoup de modèles, raison pour laquelle le seuil à 15% a été choisi.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_{exp,i} - Y_{pred,i}\right)^2}{\sum_{i=1}^{n}\left(Y_{exp,i} - \langle Y\rangle_{exp}\right)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{exp} - Y_{pred}\right| \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{exp,i} - Y_{pred,i}\right)^2}{n}} \quad (3)$$

n est le nombre de molécules.

Y_{exp} est l'affinité expérimentale.

$\langle Y\rangle_{exp}$ est la moyenne des affinités expérimentales.

Y_{pred} est l'affinité prédite.

Quant aux autres cibles dont les modèles QSAR avaient des performances médiocres, un modèle de classification SVM a été généré en incluant des molécules supposées inactives. Ces molécules inactives ont été sélectionnées aléatoirement à partir de la base PubChem Substance (Li *et al.* 2010) et plus précisément du dépôt "NIH Substance Repository". Les molécules ont été traitées avec la même procédure que celle des molécules des bases de données de bioactivités. Les molécules inactives sélectionnées pour chaque modèle de classification et pour chaque cible doivent représenter 80% des molécules dans le modèle.

Le rappel, la précision et la F-mesure (équations 4,5 et 6) ont été calculés pour évaluer les modèles selon une procédure de validation croisée *5-Fold* et un jeu de validation externe représentant 1/5 des données a été utilisé pour une validation externe supplémentaire.

$$\text{Rappel} = \frac{VP}{VP + FN} \quad (4)$$

$$\text{Précision} = \frac{VP}{VP + FP} \quad (5)$$

VP : nombre de vrais positifs

FP : nombre de faux positifs

FN : nombre vrais négatifs

$$F\text{mesure} = 2 \times \frac{\text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}} \qquad (6)$$

Les modèles de classification retenus doivent respecter une F-mesure supérieure à 0.7 à la fois dans la validation croisée et la validation externe. Pour constituer l'ensemble des molécules actives pour chaque cible, un seuil de pK_i=5 a été utilisé pour séparer les molécules actives et inactives. Les molécules inactives pour lesquelles la valeur d'affinité $pK_i \in [4.5 , 5[$ ont été écartées pour éviter que des molécules analogues ne se retrouvent entre deux classes.

b) Profilage à l'aide de la similarité 2D utilisant les proches voisins

Les cibles dont le nombre de ligands varie entre 10 et 25 observent une possibilité d'association estimée grâce à une similarité 2D à l'aide de l'empreinte moléculaire ECFP_4. L'association est prédite comme possible si la similarité du ligand à profiler et de son plus proche voisin est supérieure ou égale à 0.5 et calculée grâce au coefficient de Tanimoto.

c) Profilage à l'aide de la similarité 3D des ligands

Nous avons remarqué lors du chapitre précédent que le profilage des molécules hydrophobes était plus fructueux lorsqu'une similarité 3D de forme est appliquée. Les molécules qui possèdent un nombre de donneurs et d'accepteurs de liaisons hydrogène inférieur à 2 ont été profilées à l'aide de la méthode ROCS (OpenEye 2011) sauf pour le cas des fragments. Ces petites molécules possédant une masse moléculaire inférieure à 200 Da sont très difficiles à profiler à l'aide de la similarité de forme et nous avons fait le choix de les profiler à l'aide des autres méthodes.

Une molécule est prédite comme active sur une cible, si elle est similaire à un des ligands extraits des bases de bioactivité et répertoriés pour chaque cible. Un maximum de 100 conformères par molécule ont été générés à l'aide du programme Omega2 (OpenEye 2011) et sauvegardés. Une comparaison est effectuée pour chaque molécule à profiler avec tous les ligands connus et actifs ($pK_i \geq 5$). Ne sont retenu que ceux pour lesquels un score de TanimotoCombo ≥ 1.4 et un score de ColorTanimoto qui suit l'équation 8. Cette équation permet de retenir les molécules similaires à la fois par leurs formes et leurs points pharmacophoriques.

$$\left(\frac{\text{TanimotoCombo}}{2} - 0.2\right) \leq \text{ColorTanimoto} \leq \left(\frac{\text{TanimotoCombo}}{2} + 0.2\right) \quad (8)$$

d) Profilage à l'aide de la structure de la cible

Les cibles de structure 3D connues et pour lesquelles moins de 10 ligands sont répertoriés ont été criblées soit par recherche pharmacophorique, soit par arrimage moléculaire. D'après la récente étude de profilage (Meslamani *et al.* 2012) détaillée dans le chapitre précédant, l'arrimage moléculaire est performant quand le site de liaison est polaire et enfoui. Pour les autre cas, les pharmacophores d'interactions sont convenables. Dès lors, un arbre de décision a été construit à l'aide de descripteurs de cavités déterminés à partir du programme VolSite (Desaphy *et al.* 2012). Ce programme se base sur une représentation en grille du site de liaison où des cubes de 1.5 Å d'arête sont définis sur toute la grille. Des descripteurs pharmacophoriques (donneurs/accepteurs de liaison H, hydrophobes, aromatiques, charge +/-) sont projetés sur chaque cube selon la propriété de l'atome de protéine le plus proche. Une propriété factice Du (*dummy*) est assignée aux cubes les plus lointains et en surface et permet d'évaluer l'accessibilité en surface du site de liaison.

L'arbre de décision a pour but la prédiction de la méthode à utiliser, à savoir l'arrimage ou les pharmacophores. Une matrice de profilage de 157 ligands divers sur 2 556 cibles (Meslamani *et al.* 2012) a été utilisée pour générer l'arbre de décision. Cet arbre a été construit à partir de 17 cas de profilage où la cible principale connue du ligand profilé appartenait au premier 1% des cibles bien classées, parmi lesquelles 9 cas affichant une cible principale bien classée uniquement selon la méthode des pharmacophores et 8 autres cas où la cible est bien classée uniquement à l'aide de l'arrimage. Les descripteurs utilisés pour construire l'arbre de décision sont les pourcentages des cubes présents de chaque propriété, déterminés par rapport au nombre de cubes total du site de liaison.

Une validation sur 34 cas de profilages de la même matrice a été utilisée où 7 cas de profilages ont été performant uniquement à l'aide de l'arrimage et 27 cas uniquement à l'aide des pharmacophores.

Nous avons utilisé le programme Surflex 2.514 (Spitzer *et al.* 2012) avec les paramètres par défaut pour effectuer l'arrimage. Pour chaque pose générée, des empreintes d'interactions (Marcou *et al.* 2007) sont calculées. Ces empreintes sont comparées à celles des empreintes

du complexe de la sc-PDB pour évaluer la reproduction des interactions cristallographiques. La première empreinte calculée se compose de huit interactions (un bit code pour une interaction : hydrophobe, aromatique, liaisons hydrogène, ioniques et interactions avec un métal). La deuxième empreinte se compose de cinq interactions polaires (liaisons H, ioniques et interaction avec un métal). Les poses d'arrimages retenues doivent avoir un score $pK_d \geq 3$, un score de crash ≥ -2 ainsi qu'un Tc1 ≥ 0.6 et Tc2 ≥ 0.5 (Tc1 est la similarité calculée à l'aide du coefficient de Tanimoto sur les premières empreintes d'interactions, et Tc2 sur les deuxièmes).

La base des pharmacophores Pharmadb de la sc-PDB 2011 (Meslamani *et al.* 2012) a été utilisée pour profiler les autres cibles à l'aide du logiciel DiscoveryStudio 3.1 (Accelrys 2012). L'algorithme de correspondance rigide a été utilisé (*rigid fit*) avec les paramètres par défaut. Les conformères de chaque molécule à profiler ont été générés à l'aide du programme Catalyst et l'algorithme *Fast* (Accelrys 2012). Un maximum de 100 conformères est retenu et les paramètres par défaut du programme sont utilisés. Le meilleur score de fit ajusté (équation 7) est retenu pour chaque cible à partir de tous ses pharmacophores dans la base. Les poses retenues doivent posséder au minimum un fit ajusté de 2.6.

$$\text{fit ajusté} = (\text{fit} \times \text{M})/\text{T} \qquad (7)$$

fit est le score de fit évalué par le programme.
M est le nombre de motifs concordant.
T est le nombre total de motifs dans le pharmacophore

3. Résultats et discussion :

3.1.Extractions des données des bases de bioactivité

Dans un premier temps, toutes les associations entre molécules et cibles présentant une valeur d'affinité exprimée en pK_i, pK_d et pIC_{50} ont été retenus (Tableau 1). Les données des bases ChEMBL et d'IUPHAR-DB ont été écartées de la base PubChem BioAssay et vice versa. Au cours du traitement, environ 15% des molécules ont été écartées. Le tableau 1 résume le nombre de molécules, cibles et affinités qui sont retenu pour la suite de l'analyse.

Base de données	Nombre de molécules		Nombre de cibles		Nombre de valeurs d'affinités	
	Avant	Après	Avant	Après	Avant	Après
ChEMBL	297668	243625	3699	2337	677416	454611
PubChem BioAssay	64405	61393	369	322	255512	75496
IUPHAR-DB	2575	1633	542	325	7767	3169

Tableau 1 : Nombre de molécules, cibles et valeurs d'affinités pour les trois bases de données ChEMBL, PubChem BioAssay et IUPHAR-DB qui possèdent des affinités exprimées en pK_i, pK_d et pIC_{50}. Les valeurs avant et après traitement des données sont reportées.

Les cibles retenues montrent une diversité de classes protéique KEGG comme le montre la figure 2.

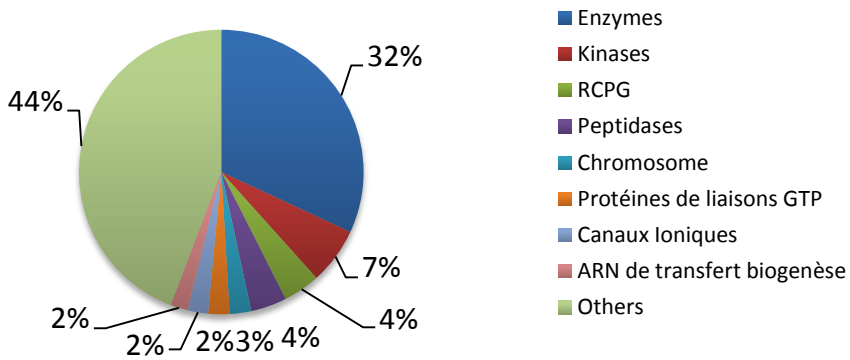


Figure 2 : Classification KEGG des cibles retenues.

3.2.Modèles de régression QSAR

Des modèles QSAR ont été générés pour toutes les cibles qui possèdent plus de 25 ligands différents et une équidistribution sur les intervalles d'affinités [4;6[, [6;8[et [8;10]. Les cibles humaines qui possédaient un nombre de ligands inférieur à 25 ont été supplémentées par les ligands d'autres cibles orthologues quand ceci était permis (cf . Matériel et méthodes 2.1.6). Le nombre de modèles QSAR (ou modèle SVR) à générer est de 271 modèles sur lesquelles 141 seront retenu comme ayant de bons paramètres statistiques.

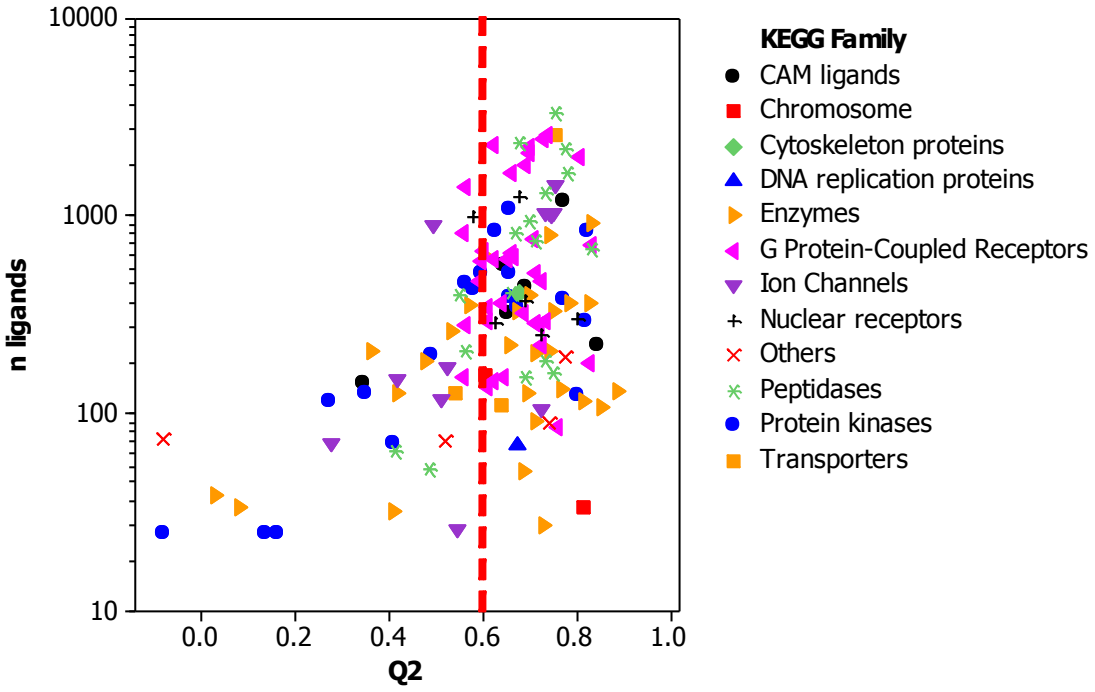


Figure 3 : Performances des 271 modèles QSAR en fonction du nombre de ligands et colorés par classes protéiques KEGG. L'échelle des ordonnées qui contient le nombre de ligand dans chaque modèle est logarithmique.

La figure 3 nous indique que la performance des modèles est reliée au nombre de molécules dans le jeu d'entrainement. L'espace chimique est bien fourni pour les cibles qui ont des bons modèles. On remarque aussi que les modèles qui n'ont pas obtenu de bons paramètres statistiques $Q^2 < 0.6$ correspondent soit à des protéines enzymatiques (quelques protéines kinases et peptidases) soit à des canaux ioniques qui ont un nombre plus faible de ligands disponibles. Leurs ligands ont des châssis moléculaires différents et c'est pour cette raison que les modèles QSAR n'arrivent pas à corréler l'affinité en fonction de la structure de ces molécules. La figure 4 montre la distribution des valeurs du coefficient de corrélation Q^2, l'erreur moyenne MAE et l'erreur type RMSE dans le cadre de la validation croisée *5-fold* répétée trois fois pour les 141 modèles QSAR retenus. On remarque que les enzymes et les RCPG constituent 62% des modèles. Ces protéines sont généralement très présentes dans les bases de données de bioactivité car elles sont les principales cibles thérapeutiques des médicaments. Plusieurs molécules ont donc été testées sur ces dernières expliquant ainsi leur surreprésentation.

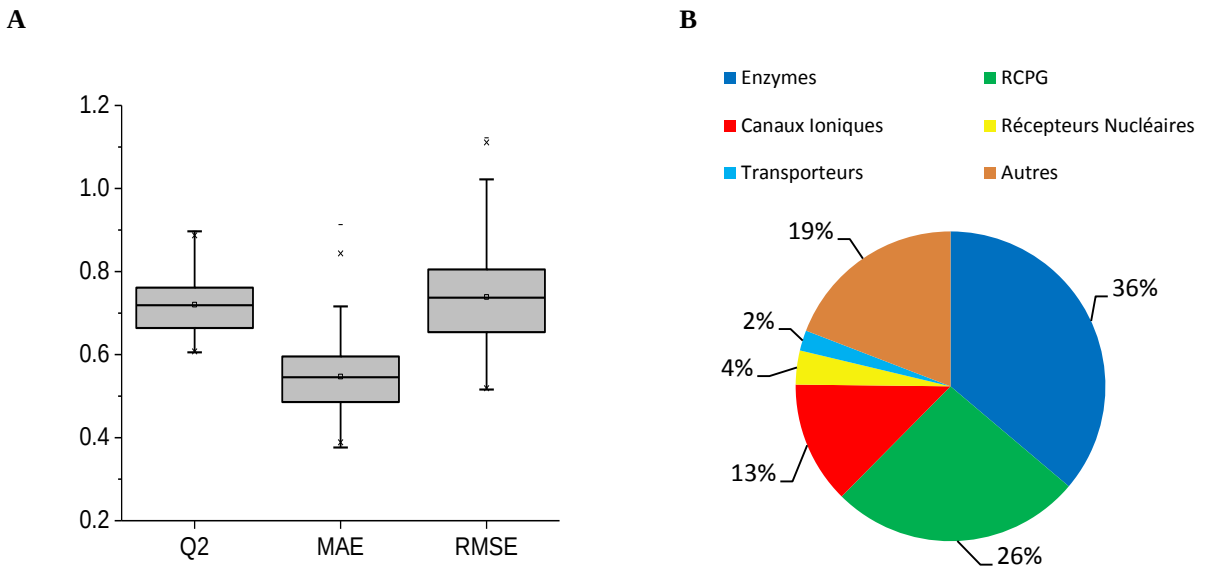


Figure 4 : (A) Distribution des valeurs des paramètres statistiques Q^2, MAE et RMSE des 141 modèles QSAR retenus possédant des valeurs $Q^2 \geq 0.6$ et MAE ≤ 1. **(B)** Pourcentage des 5 classes KEGG les plus représentées dans les 141 modèles retenus.

Les performances de nos modèles sont excellentes. A titre de comparaison, Vidal et al. (Vidal *et al.* 2010) ont profilé 13 médicament antipsychotiques sur 34 cibles dont 30 RCPG et 4 neurotransmetteurs. L'affinité de chaque molécule a été prédite en utilisant la moyenne des affinités pondérée par la distance des *k* plus proches voisins de la molécule. Les molécules du jeu d'entrainement et les valeurs d'affinités avec les cibles étaient extraites des bases ChEMBL (Gaulton *et al.* 2011), PDSP Ki (Roth *et al.* 2000), BindingDB (Liu *et al.* 2007) et IUPHAR-DB (Mpamhanga *et al.* 2012) (les versions ne sont pas mentionnées dans l'article). 65% des affinités prédites le sont avec une différence inférieure à une unité de pK_i. Nos modèles prédisent, pour les RCPG, 86% des affinités avec une différence inférieure à une unité de pK_i. Sachant que les données diffèrent entre les deux études, on peut néanmoins affirmer que nos modèles prédisent aussi bien que la méthode développée par Vidal et al.

3.3.Modèles de classification SVM

Les 1 227 cibles qui possédaient plus de 25 ligands actifs ($pK_i \geq 5$) ont été utilisées pour construire des modèles de classification SVM. Les cibles pour lesquelles des modèles de régression existent ont été écartées. Le paramètre statistique choisi pour évaluer la performance des modèles est la F-mesure. Certains modèles ne présentaient pas de valeurs de

F-mesure correctes (Figure 5). Par conséquent, un coefficient a été calculé pour évaluer la diversité des molécules actives pour chaque cible. Ce coefficient détermine la moyenne des dissimilarités entre les paires de molécules (Turner *et al.* 1997). Plus la diversité des molécules augmente, plus ce coefficient s'approche de 1. Les modèles qui possèdent une F-mesure < 0.7 présentent une grande diversité (entre [0.8;1[ ) expliquant ainsi l'échec de ces derniers. Cependant, on remarque que certains modèles performants (F-mesure > 0.8) avaient un nombre de ligands actifs assez conséquent et une bonne diversité. Cette observation est assez classique car les mesures statistiques tendent à surestimer les performances lorsqu'il s'agit d'effectuer une mesure globale sur des milliers de points.

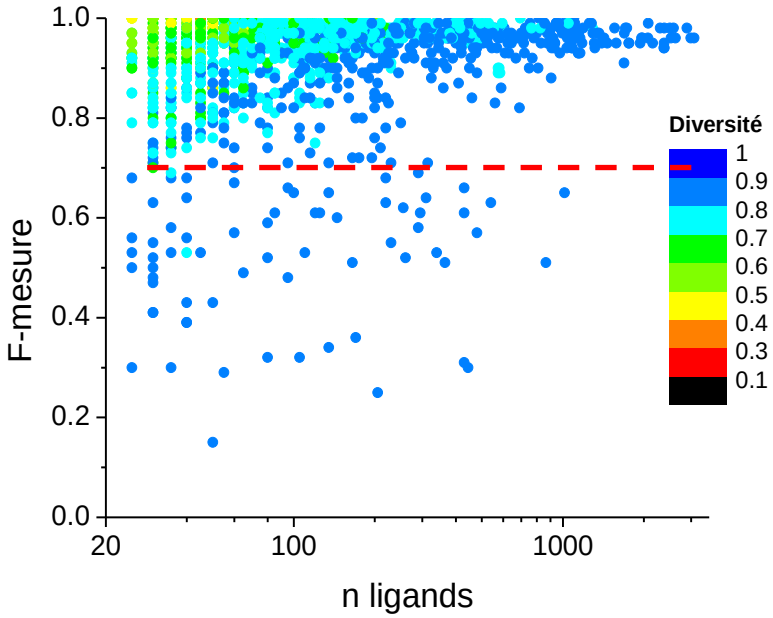


Figure 5 : Les valeurs de F-mesure pour les 1 227 modèles de classification pour la validation croisée 5-Fold répétée trois fois. Chaque point représente un modèle avec en abscisse le nombre de ligands actifs qu'il contient et la F-mesure correspondante à la validation croisée. Les points sont colorés par la diversité des ligands actifs.

Lorsqu'on analyse les mêmes cas d'échec en déterminant la classe protéique de chaque cible (Figure 6), on s'aperçoit que les mauvais modèles sont principalement ceux de protéines qui ne sont pas associés à des classes KEGG, des protéines kinases, un cytochrome P450 2B6 et trois récepteurs nucléaires (stéroidogenique 1, ror-alpha et isoforme cra_a du récepteur d'estrogène). Les espèces de ces protéines sont majoritairement humaines. Les protéines qui ne disposent pas de classes KEGG sont principalement des protéines G de régulation du

signal. Ces dernières se lient à une diversité de ligands pour induire des signaux biologiques d'où les mauvaises performances des modèles.

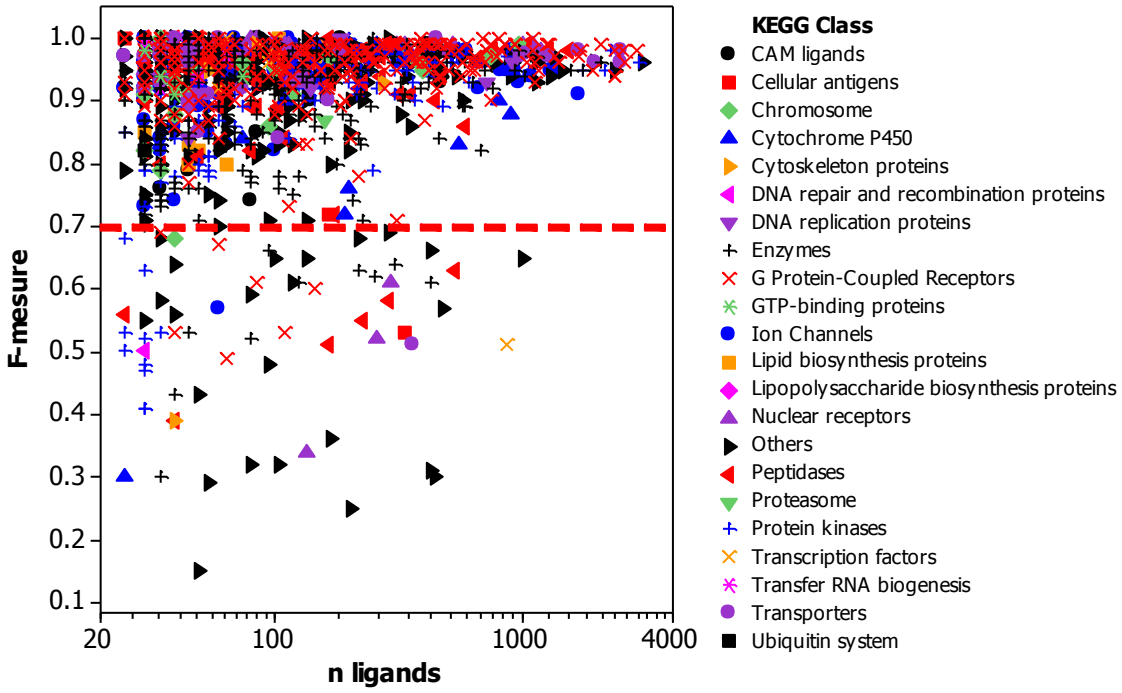


Figure 6 : Les valeurs de F-mesure pour les 1 227 modèles de classification pour la validation croisée 5-Fold répétée trois fois. Chaque point représente un modèle avec en abscisse le nombre de ligands actifs qu'il contient et la F-mesure correspondante à la validation croisée. Les points sont représentés par classe de protéines KEGG.

La validation croisée des 1 227 modèles ainsi que les résultats sur le test set nous indiquent que 54% des modèles possèdent F-mesure ≥ 0.7 à la fois en validation croisée *5-fold* répétée trois fois et en validation externe. Ces 667 modèles seront retenus pour le protocole de profilage (Figure 7).

A titre de comparaison, Strömbergsson et al. (Strombergsson *et al.* 2010) utilisent un modèle chémogénomique de classification généré à partir de 7087 complexes protéine-ligand (2 721 ligands et 340 cibles) extraits de la base BindingDB (Liu *et al.* 2007). La précision de ces modèles est de 82%. Sachant que ces complexes couvrent la majorité des cibles thérapeutiques, nous pouvons affirmer que nos modèles de classification sont de performance supérieure avec une précision moyenne de 92% sur les 667 modèles retenus.

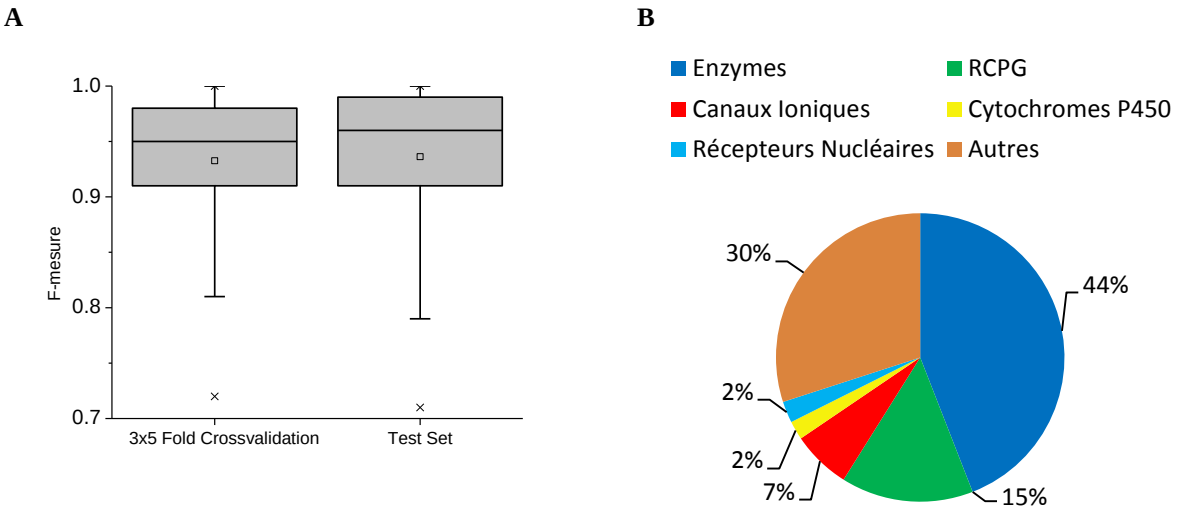


Figure 7 : (A) Distribution des valeurs du paramètre statistique F-mesure (Fm) pour les 667 modèles de classifications retenus possédant des valeurs F-mesure ≥ 0.7 et MAE ≤ 1. **(B)** Pourcentage des classes KEGG présentes pour ces modèles.

3.4. Similarité 2D par les proches voisins

Pour les 424 cibles restantes pour lesquelles les modèles QSAR et de classification SVM n'ont pas eu des paramètres statistiques convenables, une simple recherche par similarité du plus proche voisin déterminera le transfert d'annotation de cible du ligand retenu à la requête.

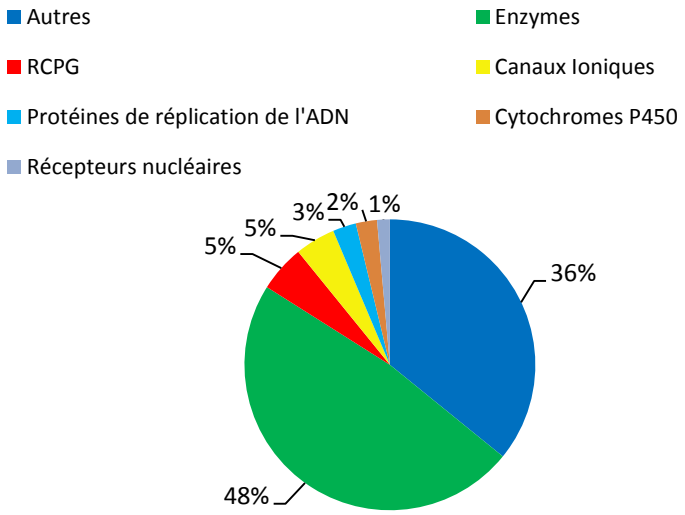


Figure 8 : Classes des 424 cibles pour du profilage à l'aide de la similarité 2D du proche voisin.

3.5. Arbre de décision pour les méthodes basées sur la structure 3D de la cible

Les cibles de structure 3D connue pour lesquelles moins de 10 ligands sont disponibles ont été profilées à l'aide d'une méthode basée sur leur structure 3D. La version de la base sc-PDB 2011 inclus 3 034 cibles différentes dont 2 549 pour lesquelles moins de 10 ligands étaient disponibles dans les bases de données de bioactivité. Par conséquent, ces cibles vont être profilées soit à l'aide des pharmacophores soit de l'arrimage moléculaire. Un arbre de décision a été créé pour déterminer la méthode à utiliser (c.f. matériels et méthodes). Dix-sept cibles dont le profilage était performant grâce à une de des deux méthodes (pharmacophore et arrimage) ont été utilisées pour construire l'arbre de décision présent dans la figure 9. L'arbre a été validé sur une matrice de profilage d'une étude précédente (Meslamani *et al.* 2012). 34 ligands profilés ont été utilisées pour cette analyse. 27 ligands présentaient un bon profilage (cible retrouvée parmi les premières 1% de la liste) à l'aide des pharmacophores et pas en arrimage, et 7 autres cas où c'était l'inverse. Les propriétés des cavités des entrés sc-PDB qui ont permis de récupérer la cible pour les ligands profilés sont utilisées pour la validation.

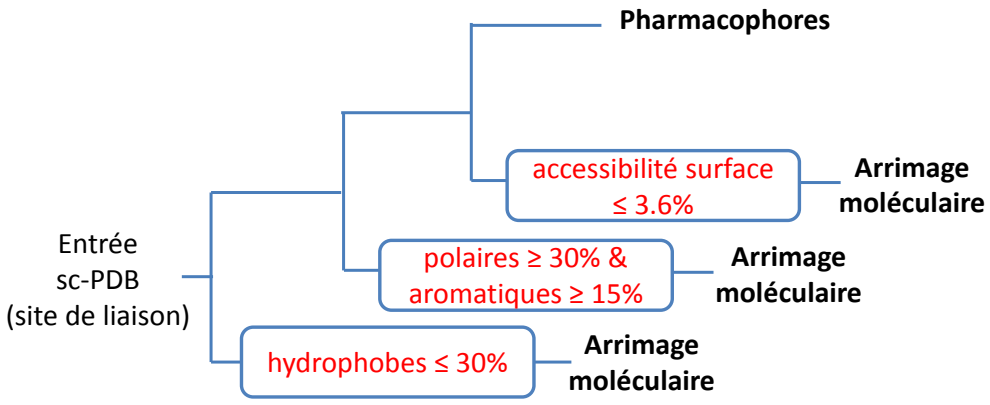


Figure 9 : Arbre de décision pour choisir entre la méthode de l'arrimage moléculaire et les pharmacophores. Le pourcentage des points de cavité hydrophobes, polaires, aromatiques et l'accessibilité sont déterminés à partir du programme VolSite (Desaphy *et al.* 2012) pour chaque site de liaison.

Un rappel de 0.74 a été obtenu pour les profilages où la cible principale était retrouvée uniquement à l'aide des pharmacophores et un rappel de 0.67 pour le profilage où la cible principale était retrouvée uniquement à l'aide de l'arrimage.

Validation	VP	FN	Rappel
Pharmacophores	20	7	0.74
Arrimage moléculaire	5	2	0.71

Tableau 3 : Résultats de la validation externe de l'arbre de décision

Ainsi, 1 551 cibles seront profilées à l'aide des pharmacophores, et 998 autres cibles à l'aide de l'arrimage moléculaire.

3.6. Validation externe du protocole à l'aide de la base DrugBank

Un total de 6 509 médicaments de la base DrugBank 3.0 ont été récupérés. 5 782 molécules ont été retenues après traitement des données. Après une élimination de molécules présentes dans le jeu d'entrainement, 1 204 molécules ont été obtenues et soumises à une classification par graphe maximal commun. Par conséquent, 591 classes de châssis moléculaires différents ont été obtenues. Des classes KEGG ont été assignées à ces molécules et une sélection de 119 d'entre elles a été faite de façon à obtenir un ensemble de validation hétérogène à la fois du point de vue du châssis moléculaire et du point de vue protéique afin d'évaluer au mieux les performances du protocole de profilage (liste de molécules et leur cibles sont présentes dans l'annexe 1).

Il est important de souligner que 49% des 1 204 molécules retenues sont des fragments (masse moléculaire ≤ 300 et AlogP ≤ 3 et nombre d'angles de torsion ≤ 5).

Douze des 119 molécules ont un nombre d'accepteurs et de donneurs de liaison hydrogène inférieur ou égal à deux, et par conséquent vont être profilés uniquement à l'aide de la similarité de forme ROCS. Les molécules sélectionnées présentent une bonne diversité selon les propriétés physicochimiques (Figure 10). On remarque que 73% des molécules ont une seule cible connue à retrouver à l'aide de notre protocole de profilage (Figure 10, F).

Les 119 molécules ont été profilées sur les 3 781 cibles contenues dans le protocole de profilage et un rappel évaluant ainsi le nombre de cibles principales retenues a été calculé pour chacune d'entre elle (Figure 11, A).

Nous remarquons que le protocole de profilage a réussi à identifier au moins une vraie cible pour les 44% des molécules profilées (Figure 11, A). Ceci correspond à 25 cibles sur les 221 possibles. Ce pourcentage est certes faible mais si on le compare au nombre total de cibles prédites, ceci reste relativement raisonnable (Figure 12). Le rappel pour chaque méthode de

criblage est reporté dans la figure 11, B. Celui-ci évalue les performances de chaque méthode de criblage à retrouver les vraies cibles qu'elles sont censées identifier. Nous apercevons que l'arrimage et la similarité 2D à l'aide des empreintes ECFP_4 n'ont pas des rappels très élevés. Ceci est en partie dû à l'inefficacité des seuils de scores choisis. Mais en général, l'échec peut être dû à plusieurs autres raisons. La première est que les molécules extraites de DrugBank et restantes pour la validation du protocole de profilage, sont dissimilaires à celles incluses dans le jeu d'entrainement. Les méthodes basées sur les ligands échouent alors à retrouver la cible connue.

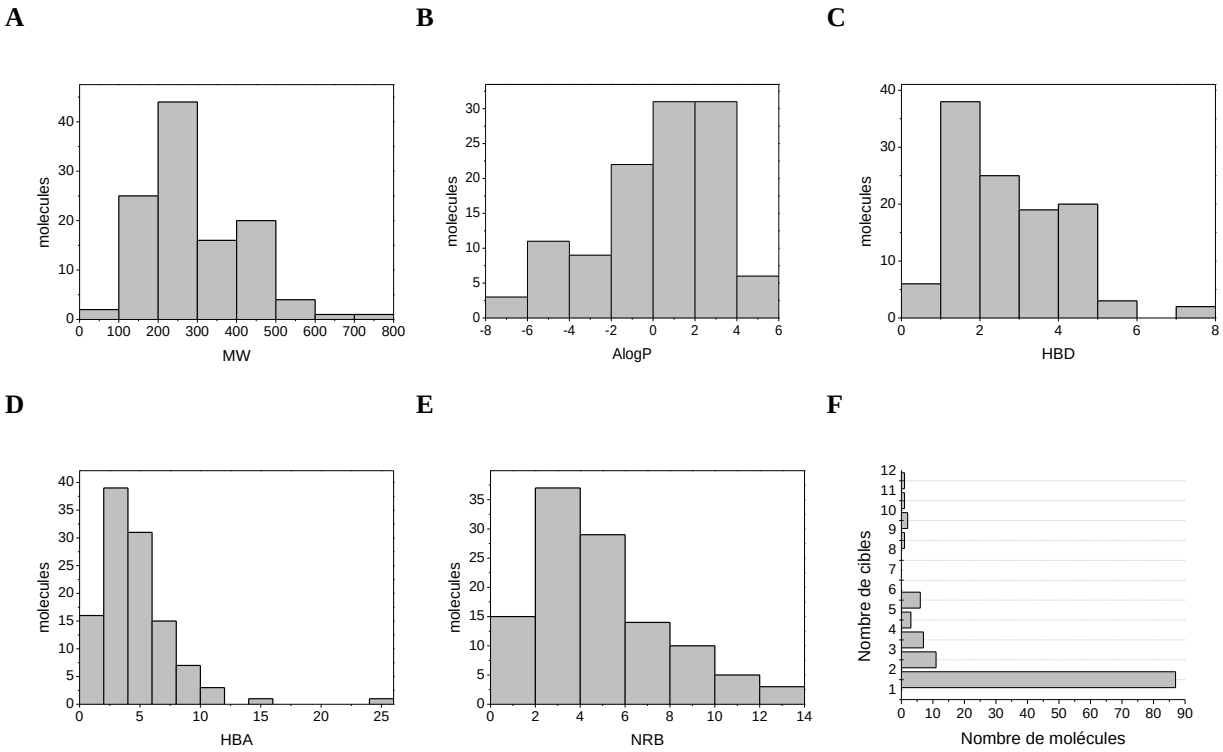


Figure 10 : Propriétés physico-chimiques des 119 molécules de DrugBank et leur nombre de cibles répertoriées dans la base DrugBank. **(A)** distribution de la masse moléculaire. **(B)** coefficient de partage AlogP. **(C)** nombre de donneurs de liaison H. **(D)** nombre d'accepteurs de liaison H. **(E)** nombre d'angles de torsions. **(F)** nombre de cibles des molécules.

Concernant les cibles qui devaient être retrouvées à l'aide de l'arrimage et des pharmacophores, ces méthodes étant basés sur le site de liaison, il nous est impossible d'affirmer que les ligands extraits de DrugBank se lient bel et bien au site défini dans la sc-PDB.

A **B**

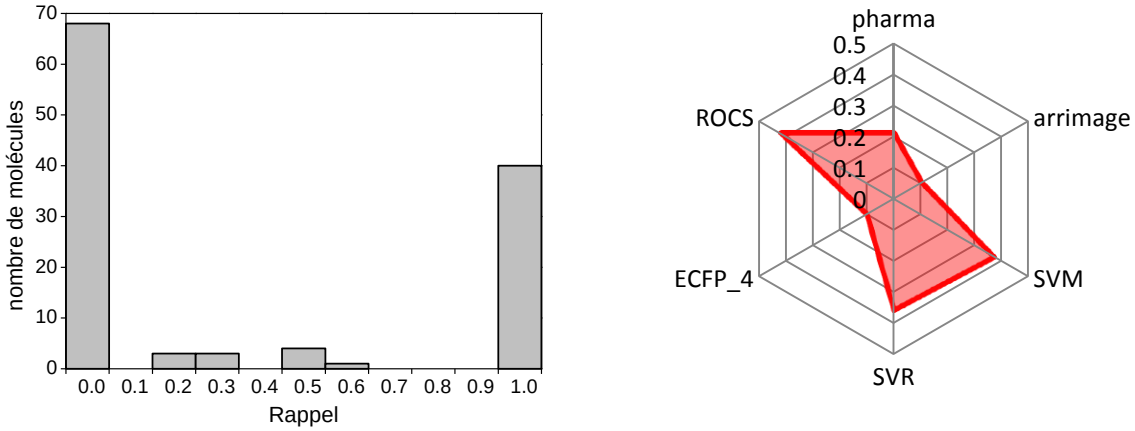


Figure 11 : (A) Rappel du nombre cibles retrouvées pour les 119 molécules profilées. **(B)** Rappel du nombre cibles trouvées pour chaque méthode de criblage.

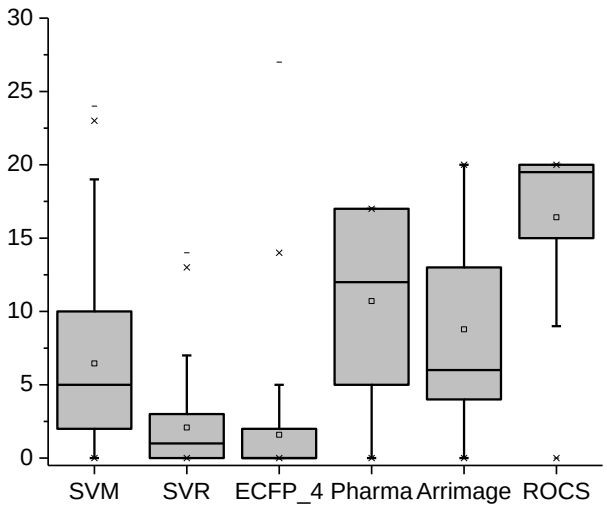


Figure 12 : Distribution des nombres de cibles retrouvés pour chaque méthode pour les 12 molécules hydrophobes pour la méthode ROCS, et des 107 molécules pour les méthodes de classification SVM, QSAR (SVR), similarité 2D à l'aide des empreintes ECFP_4, pharmacophores et arrimage moléculaire.

3.6.1. Performances des méthodes 2D basées sur les ligands

Nous avons essayé de comparer les propriétés des molécules qui ont été bien profilées et celles pour lesquelles les méthodes basées sur les ligands (modèles de régression, modèles de classification et similarité 2D) ont échoué (Figure 13). Un profilage est évalué comme bon si une des cibles répertoriées dans DrugBank est identifiée pour chaque molécule. Les cas de succès représentaient 38 molécules et les cas d'échec 68 molécules. On s'aperçoit que les molécules de faible poids moléculaire posent problème. Ces molécules sont généralement plus polaires d'après les valeurs de leurs coefficients de partage (Figure 13, C) et le nombre de cycles aromatiques (Figure 13, B). Cette observation peut s'expliquer par le fait qu'un grand nombre des molécules extraites des bases de données de bioactivité sont hydrophobes. Des molécules hydrophobes auront alors tendance à être bien profilées. Ces molécules reflètent la tendance des chimistes médicinaux à synthétiser des molécules qui contiennent des cycles aromatiques (Peters *et al.* 2012). Cependant, on peut remarquer qu'à partir des figures 13 D, E et F que ni la flexibilité ni le nombre de donneurs et d'accepteurs de liaisons hydrogène ne sont caractéristiques de différences entre les molécules où le profilage 2D a échoué et celles où il a réussi.

Dans le but de mettre en évidence notre méthode par rapport aux méthodes de profilage 2D actuelles, nous avons effectué des profilages de quelques molécules à l'aide du portail SEA (Keiser *et al.* 2007). Les molécules qui ciblent les grandes familles de protéines (Trypsine, RCPG) se profilent correctement à la fois dans SEA et dans notre protocole Profiler. Cependant il est un peu difficile de connaitre toute les cibles incluses dans SEA car le portail web ne le mentionne pas. Il faut alors essayer de profiler des molécules et essayer de voir la liste des cibles incluses de la sorte. De ce fait, nous avons pris au hasard deux molécules dont la cible connue est répertoriée à la fois dans SEA et dans Profiler. Ces molécules sont DB08292 et DB07178 (Figure 14). Les empreintes ECFP_4 et des prédictions à partir des ligands ChEMBL ont été choisi. Dans la base DrugBank, la molécule DB08292 se lie à la protéine de choc thermique HSP90 β et la molécule DB07178 à la protéine dipeptidyl peptidase 4.

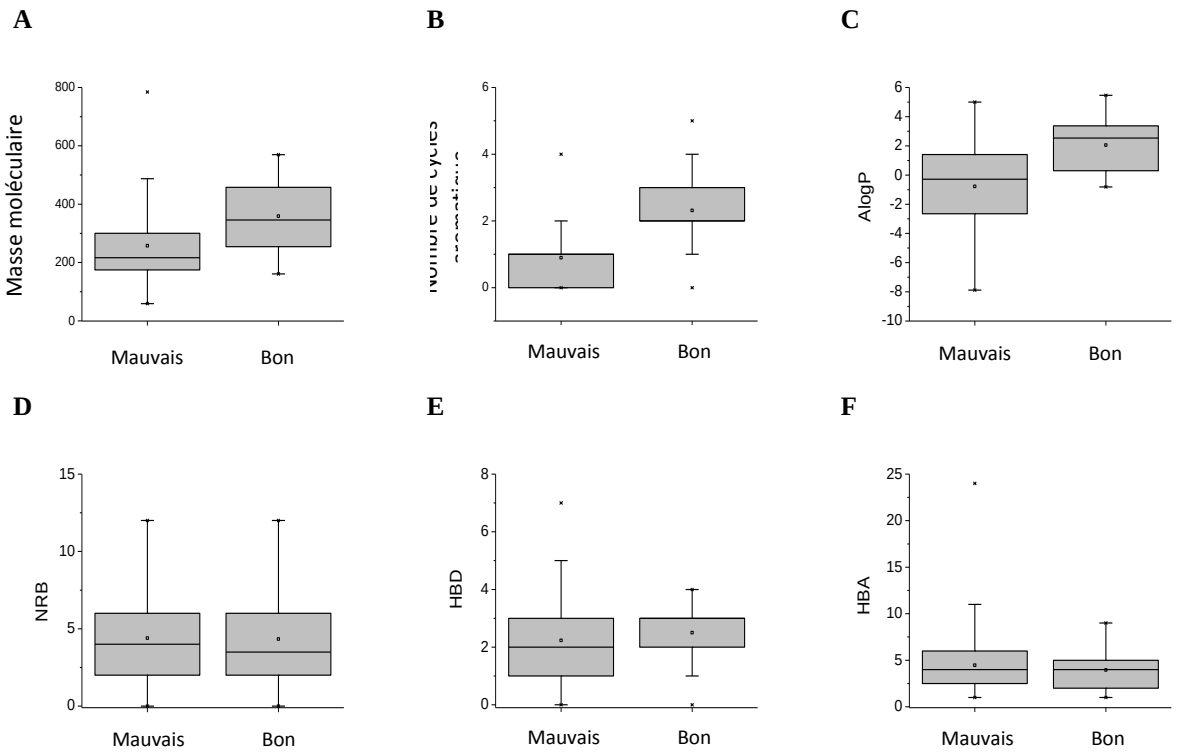


Figure 13 : Distribution des propriétés des 38 molécules qui ont été bien profilées à l'aide des méthodes 2D basées sur les ligands, ainsi que celles (68) où ces méthodes ont échoué. **(A)** Distribution des masses moléculaires. **(B)** Distribution des nombres de cycles aromatiques. **(C)** Distribution des coefficients de partage AlogP. **(D)** Distribution des nombres d'angles de torsions. **(E)** Distribution des nombres de donneurs de liaison H. **(F)** Distribution des nombres d'accepteurs de liaison H.

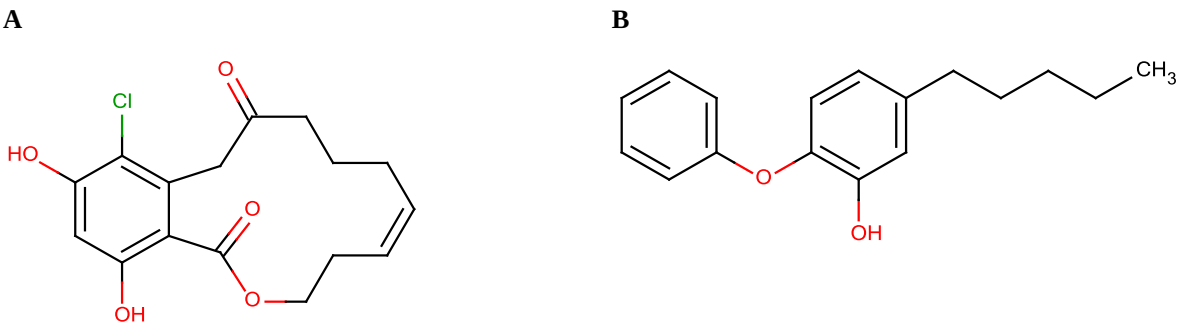


Figure 14 : (A) Molécule DB08292. **(B)** Molécule DB07178.

Molécule	Cible	Rang SEA	Prédiction Profiler 2D (*)	Nombre de cibles prédites SEA	Nombre de cibles prédites Profiler	Nombre des ligands de la cible dans le jeu d'entrainement SEA	Nombre des ligands du jeu d'entrainement Profiler
DB08292	HSP90 β	79	✓	124	18	130	80
DB07178	Dipeptidyl peptidase 4	✗	✓	162	42	57	1 790

Tableau 4 : Résultats du profilage des deux molécules DB08292 et DB07178 par la méthode SEA et profiler. (*) Profiler ne fournit pas de rang vu que plusieurs méthodes sont employées.

Notre protocole de profilage Profiler a réussi à identifier les deux cibles à travers les modèles de classification, par contre, la méthode SEA ne prédit la bonne cible pour la première molécule qu'à la 79ème place, et échoue pour la deuxième. Lors d'un criblage expérimental de la première molécule DB08292, il aurait fallu tester 79 cibles afin d'identifier celle qu'on cherche à trouver. Ceci n'est évidemment pas très efficace lors d'un profilage car la procédure devient très couteuse. Le protocole Profiler ne prédit avantageusement que 18 cibles au total pour cette même molécule.

La deuxième molécule n'est prédite que par notre protocole, qui a identifié 42 cibles potentielles dont 19 provenaient des méthodes 2D basées sur les ligands. La méthode SEA identifie 162 cibles mais la cible connue n'est malheureusement pas présente dans cette liste. Il est bien sûr très important de souligner que les jeux d'entrainement des méthodes SEA et Profiler sont différents, ceci pouvant être la cause de l'échec ou non d'une des méthodes.

On peut affirmer ainsi que notre protocole fournit à la fois une liste de cibles de taille raisonnable pour une éventuelle expérimentation mais également avec des prédictions fiables, même si le taux d'erreurs est difficilement évaluable car on ne dispose pas de toutes les valeurs d'associations possibles entre les molécules profilées et les cibles incluses dans notre jeu d'entrainement.

3.6.2. Performances de la similarité 3D avec ROCS

Le profilage à l'aide de la similarité de forme sur les molécules hydrophobes a obtenu un rappel moyen de 0.42 pour l'identification des cibles des 12 molécules profilées. Dans l'analyse qui suit, un profilage est évalué comme efficace si la moitié des cibles répertoriées dans DrugBank sont identifiées pour chaque molécule. Cinq molécules présentaient un

profilage performant et sept autres ont conduit à un échec. La figure 15, A nous montre que la taille des molécules n'affecte pas la performance du profilage. Par contre, le nombre d'angles de torsion et la surface polaire (Figure 15, B, C et D) de la molécule en sont un facteur déterminant.

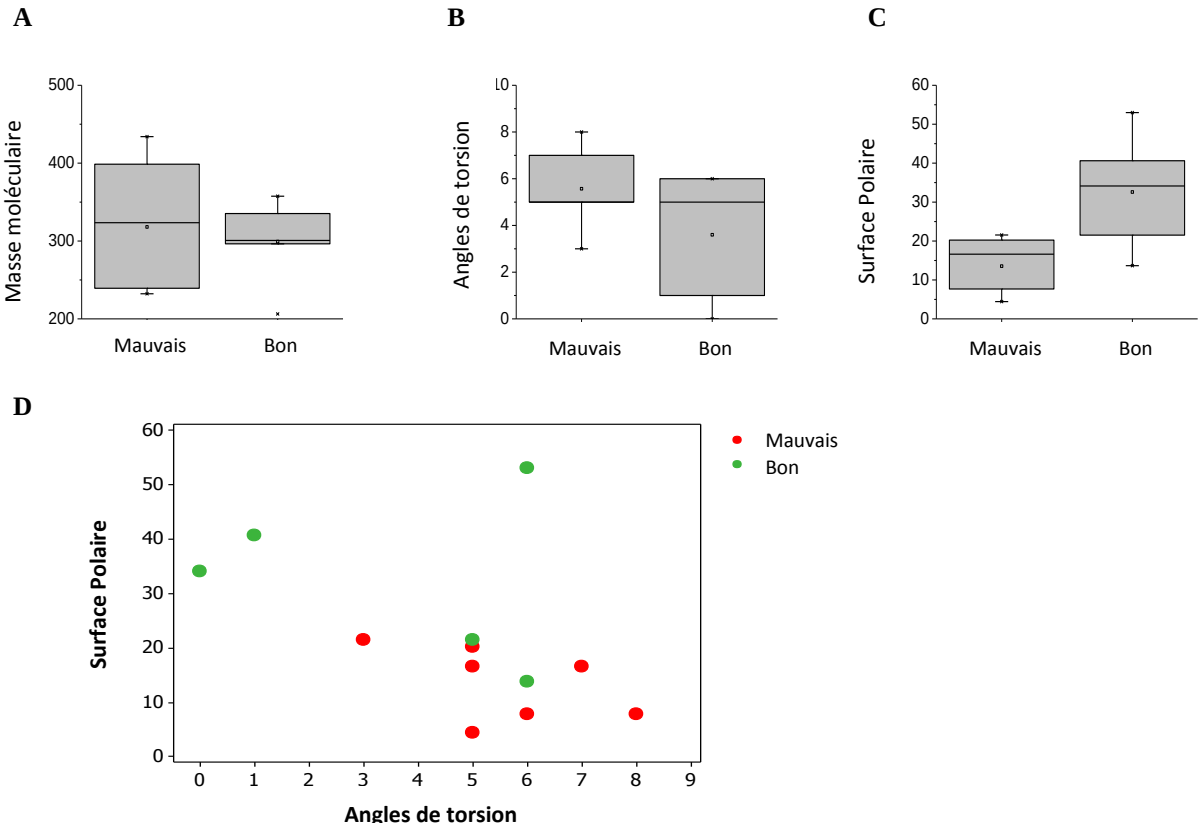


Figure 15 : Distribution des propriétés des 5 molécules qui ont été bien profilées à l'aide de la méthode 3D basée sur les ligands, ainsi que celles (7 molécules) où ces méthodes ont échoué. **(A)** distribution des masses moléculaires. **(B)** distribution des angles de torsions. **(C)** distribution des surfaces polaires en $Å^2$. **(D)** La surface polaire des ligands en $Å^2$ en fonction du nombre d'angles de torsion pour chaque molécule.

Dès qu'on franchit le seuil de 5 à 6 angles de torsions, on remarque que le profilage a tendance à échouer. Ceci est peut être une limite de l'échantillonnage conformationel qui n'a pas été suffisant. La figure 15, D nous montre que deux molécules ont été profilées efficacement malgré leurs nombres d'angles de torsion égal à 5 et 6 respectivement et avec une surface polaire < 20$Å^2$. Ces deux molécules ont pour identifiant DrugBank DB01237 et DB00937 (Figure 16A, C). En dépit du nombre d'angles de torsion présent dans les deux molécules, les formes 3D de leurs conformères (Figure 16B, D) sont équivalentes, avec une partie du châssis qui est plus rigide que l'extrémité formée par le diméthyl et diéthyl amine.

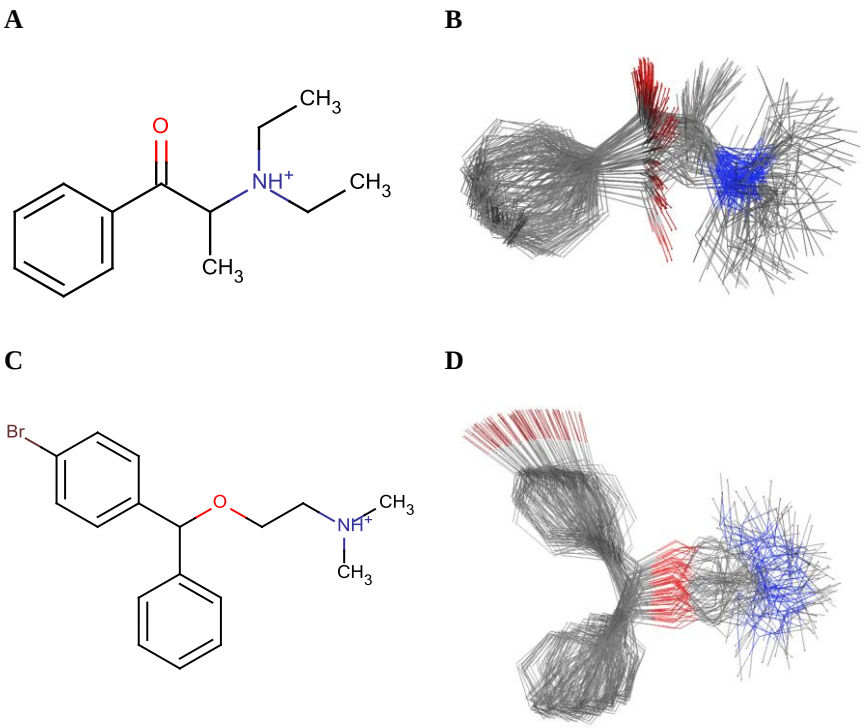


Figure 16 : **(A)** La molécule DB01237. **(B)** Conformations utilisées de la molécule DB01237. **(C)** La molécule DB00937. **(D)** Conformations utilisées de la molécule DB00937.

Malheureusement, nous n'avons pas eu la possibilité d'effectuer une comparaison avec une autre méthode de profilage de forme 3D disponible à la communauté scientifique. Le seul portail connu est celui de la méthode ReverseScreen3D (Kinnings *et al.* 2011) qui n'est malheureusement plus maintenu.

3.6.3. Performances des pharmacophores protéine-ligand

Un total de 20 molécules parmi les 117 choisies pour la validation présentent une cible à retrouver à l'aide des pharmacophores protéine-ligand issus des entrées de la base sc-PDB. Cependant, les cibles n'ont été identifiées que pour quatre d'entre elles. D'après les distributions des propriétés de ces ligands DrugBank profilés (Figures 17, A, B, C), on s'aperçoit que ni la taille, la flexibilité, ni les propriétés physicochimiques ne différent entre les molécules bien profilées et les autres où les pharmacophores ont enregistré un échec. Le volume, la taille des sites de liaisons et la sélectivité des pharmacophores pour les deux catégories de ligands bien profilées et les autres, ne sont pas non plus la cause de l'échec (Figure 17 E, F, G). Par contre si on regarde le nombre de cibles répertoriées dans la base

DrugBank pour ces ligands (Figure 17, D), on s'aperçoit que les cas d'échecs interviennent surtout quand les ligands sont associés à plusieurs cibles. Le protocole de profilage a tendance à sélectionner plus de cibles pour les ligands qui sont promiscuitaires et donc augmenter les chances de ne pas sélectionner celles qui sont connues.

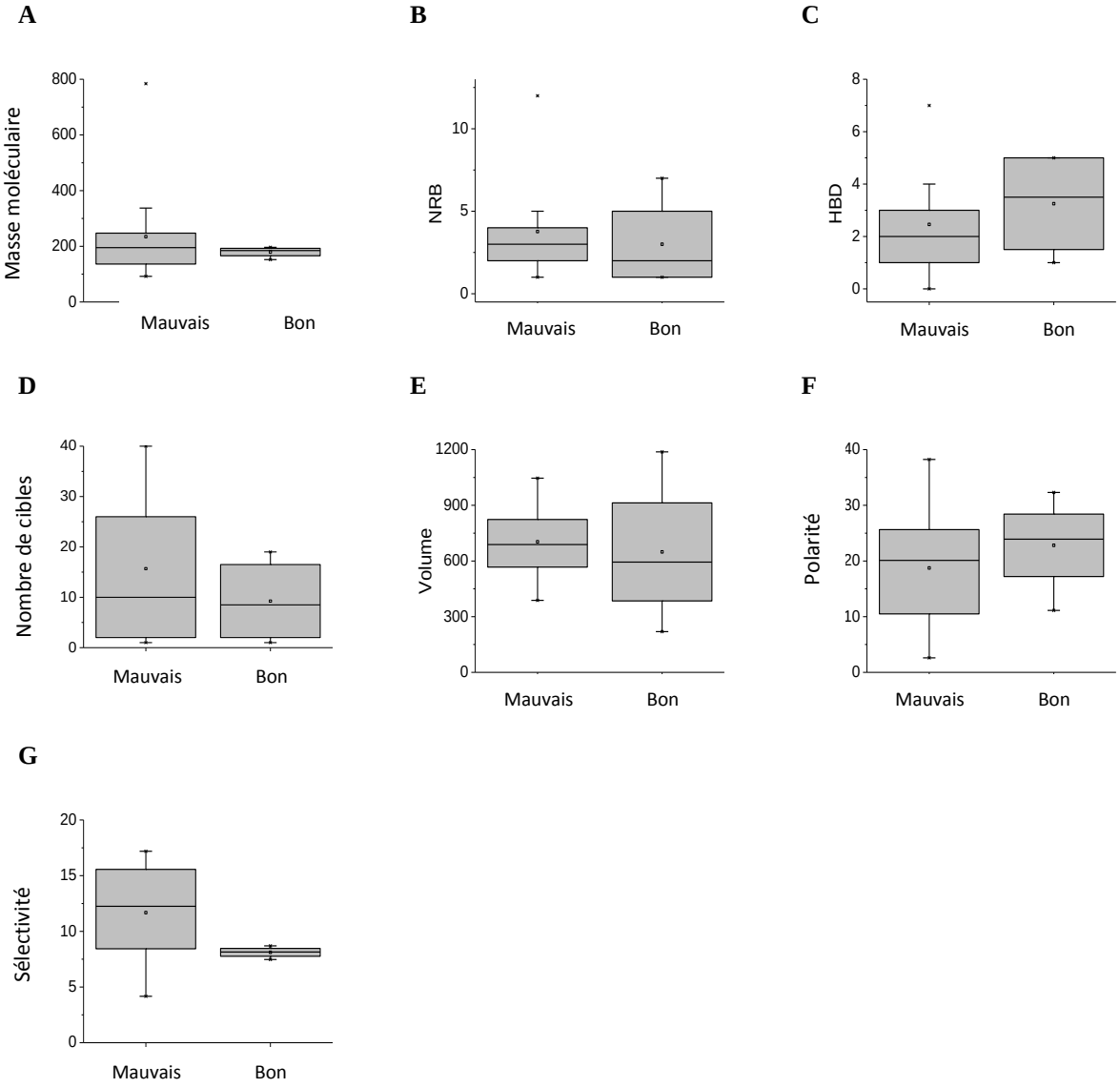


Figure 17 : Distribution des propriétés des cas d'échec et de réussite des profilages. **(A)** Masse moléculaire des ligands. **(B)** Nombre d'angles de torsions des ligands. **(C)** Nombre de donneurs de liaison H. **(D)** Nombre de cibles répertoriées dans DrugBank des molécules profilées. **(E)** Volume en $Å^3$ des sites de cibles des molécules profilées. **(F)** Pourcentage de la polarité des sites des molécules profilées. **(G)** Valeur de sélectivité (Meslamani *et al.* 2012) des modèles de pharmacophores des cibles des molécules profilées.

L'utilisation des pharmacophores dans le profilage présente quelques avantages par rapport aux profilages basés sur les ligands, et surtout pour les cibles qui ne disposent pas d'assez de ligands (<10 dans notre protocole). En voici un exemple avec la molécule DB03801 (Figure 18, A) qui se lie à 19 cibles selon la base DrugBank, dont l'une d'entre elles est la protéine "UDP-N-acetylmuramoylalanine--D-glutamate ligase" (code d'accession Uniprot : P14900). Cette protéine dispose de 8 ligands connus.

A **B**

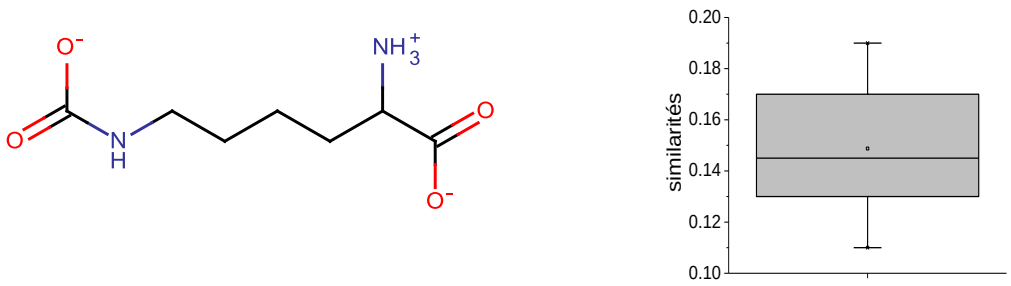


Figure 18 : (A) Structure de la molécule DB03801. (B) Similarités calculées à partir des coefficients de Tanimoto sur les empreintes ECFP_4 entre la molécule DB03801 et les 8 ligands connus pour la protéine P14900.

Les similarités calculées à partir des empreintes ECFP_4 entre la molécule DB03801 et les 8 autres ligands montrent que cette molécule présente une très faible similarité structurale avec les ligands existants (Figure 18, B). De ce fait, les méthodes basées sur la similarité 2D n'auront pas été capables d'identifier cette molécule et le profilage aurait échoué. Il est alors important d'utiliser la méthode des pharmacophores pour ces cibles qui ne disposent pas de beaucoup de ligands car cette méthode permet d'identifier des molécules qui ont des structures et des châssis différents des ligands connus.

3.6.4. Performances de l'arrimage moléculaire

Vingt-neuf des ligands sélectionnés à partir de la base DrugBank possédaient au moins une cible à identifier à l'aide de l'arrimage moléculaire avec Surflex. Le protocole de profilage n'en a identifié que trois cibles pour deux de ces molécules. Les raisons de l'échec tout comme pour les pharmacophores proviennent du fait que ces molécules sont promiscuitaires (Tableau 5). Les molécules pour lesquelles la cible n'a pas été identifiée sont

celles qui possédaient en moyenne 9.2 cibles contre 4.3 pour celles où le protocole avait réussi. Le nombre d'angles de torsions est aussi faible en moyenne pour les cas bien prédits avec 2.3 contre 4.6 pour les cas où la prédiction n'a pas été bonne. Les propriétés des sites de liaisons ne montrent pas de différences entre les deux cas. La performance de l'arrimage dans notre protocole est conséquemment dépendante de la nature du ligand sélectionné.

Propriétés	**Echec**		**Réussite**	
Ligands	Moyenne	Déviation standard	Moyenne	Déviation standard
Nombre de cibles DrugBank	9.2	11.2	4.3	0.9
Masse moléculaire	225.2	80.5	283.5	53.5
Angles de torsions	4.6	2.2	2.3	0.8
Surface polaire $Å^2$	118.5	34.9	162.8	20.8
Donneurs de liaison H	2.2	1.2	1.7	1.6
Accepteurs de liaison H	5.1	2	8.7	2
Sites de liaison	Moyenne	Déviation standard	Moyenne	Déviation standard
Volume $Å^3$	729.5	218.6	653.6	328.6
Polarité %	27.7	8.8	37	11

Tableau 5 : Propriétés des ligands DrugBank et des sites de liaisons de leurs cibles du jeu d'entrainement.

Nous pouvons affirmer que notre arbre de décision est performant car les cas d'échec des profilages dépendent le plus souvent de la nature des ligands profilés et non pas des propriétés des sites de liaisons, observation remarquée à la fois à partir des profilages à l'aide des pharmacophores et à l'aide de l'arrimage moléculaire.

4. Conclusion

A présent, les méthodes de criblage virtuel sont nombreuses et variées. Nous avons constaté que toutes les approches de criblage sont intéressantes et que chacune d'entre elle à ses propres limites. Ces dernières dépendent en général de la nature de la molécule et celle de la cible pour laquelle on essaye de prédire l'association. C'est pourquoi nous avons identifié lors d'une précédente étude de profilage (Meslamani *et al.* 2012) les cas de succès de quelques méthodes de criblage afin de les utiliser dans un cadre approprié selon la nature de la molécule à profiler et celle de l'éventuelle cible.

Il est certes difficile de valider un protocole de profilage si les molécules utilisées n'ont pas toutes été testées expérimentalement sur toutes les cibles. Néanmoins, le protocole a réussi à identifier au moins une cible connue pour 51 des 117 médicaments sélectionnés à partir de la base DrugBank. Des incertitudes demeurent sur les annotations biologiques des cibles répertoriées dans cette banque pour certaines molécules qui sont promiscuitaires, d'autant plus que les valeurs d'affinités ne sont pas disponibles. Toutefois, nous avons remarqué que notre protocole avait des performances similaires si ce n'est meilleures que celles qui existent à l'instar de la méthode SEA sur deux cas que nous avons montré. Le protocole de profilage bénéficie de plusieurs méthodes de criblage virtuel adaptées à chaque molécule et à chaque cible ce qui nous permet d'être beaucoup plus efficace dans les prédictions.

Le but de tout profilage consiste à récupérer une liste de cibles assez réduite mais pertinente et c'est ce que nous avons essayé de mettre en place.

Nous avons démontré qu'avec notre protocole, il est désormais possible d'obtenir un profil complet pour chaque molécule. Une validation expérimentale de quelques cibles retrouvées sur des molécules orphelines en ferait une validation idéale. Le profil biologique obtenu peut servir à relier les cibles prédites avec celles qui sont impliquées dans des effets secondaires ou de toxicité afin de les anticiper.

Afin d'améliorer les performances du protocole, il serait intéressant de trouver des règles statistiques dans le but de corriger les seuils de scores choisis et non pas d'utiliser des seuils fixes pour chaque méthode.

5. Contributions

Ce travail a été réalisé avec la contribution de deux de mes collègues. François Martz a écrit les scripts pour extraire les informations des bases ChEMBL et IUPHAR-DB. Ricky Bhajun a effectué la validation croisée des 141 modèles de régression et écrit le script pour la correspondance entre les codes d'accession Uniprot et la classification orthologique KEGG.

6. Références

Accelrys (2012). "Discovery Studio v.3.1.0; Accelrys Software Inc.: San Diego, CA 92121, U.S.A."

ChemAxon (2011). "Standardizer, JChem 5.7.2. ChemAxon Kft. Budapest, Hungary."

Desaphy, J., K. Azdimousa, E. Kellenberger and D. Rognan (2012). "Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes." J Chem Inf Model.

Gaulton, A., L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington (2011). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Res **40**: D1100-1107.

Keiser, M. J., B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet (2007). "Relating protein pharmacology by ligand chemistry." Nat Biotechnol **25**(2): 197-206.

Kellenberger, E., P. Muller, C. Schalon, G. Bret, N. Foata and D. Rognan (2006). "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank." J Chem Inf Model **46**(2): 717-727.

Kinnings, S. L. and R. M. Jackson (2011). "ReverseScreen3D: a structure-based ligand matching method to identify protein targets." J Chem Inf Model **51**(3): 624-634.

Knox, C., V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart (2011). "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic Acids Res **39**(Database issue): D1035-1041.

Li, Q., T. Cheng, Y. Wang and S. H. Bryant (2010). "PubChem as a public resource for drug discovery." Drug Discov Today **15**(23-24): 1052-1057.

Liu, T., Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic Acids Res **35**(Database issue): D198-201.

Marcou, G. and D. Rognan (2007). "Optimizing fragment and scaffold docking by use of molecular interaction fingerprints." J Chem Inf Model **47**(1): 195-207.

Meslamani, J., J. Li, J. Sutter, A. Stevens, H. O. Bertrand and D. Rognan (2012). "Protein-Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling." J Chem Inf Model.

Meslamani, J., D. Rognan and E. Kellenberger (2011). "sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins." Bioinformatics **27**(9): 1324-1326.

MolecularNetworks (2005). "Corina 3.1. Molecular Networks GmbH - Computerchemie."

Mpamhanga, C. P., J. L. Sharman, A. J. Harmar and I. Nc (2012). "How to Use the IUPHAR Receptor Database to Navigate Pharmacological Data." Methods Mol Biol **897**: 15-29.

OpenEye (2011). "Filter 2.1.1. OpenEye Santa Fe, NM ".

OpenEye (2011). "Omega 2.4.6. OpenEye Santa Fe, NM."

OpenEye (2011). "QUACPAC 1.5.0. OpenEye Santa Fe, NM."

OpenEye (2011). "ROCS 3.1.2. OpenEye Santa Fe, NM."

Peters, J. U., J. Hert, C. Bissantz, A. Hillebrecht, G. Gerebtzoff, S. Bendels, F. Tillier, J. Migeon, H. Fischer, W. Guba and M. Kansy (2012). "Can we discover pharmacological promiscuity early in the drug discovery process?" Drug Discov Today **17**(7-8): 325-335.

Rogers, D. and M. Hahn (2010). "Extended-connectivity fingerprints." J Chem Inf Model **50**(5): 742-754.

Rognan, D. (2012). "Computational approaches to target fishing and ligand profiling." AIP Conference Proceedings **1456**(1): 157-164.

Roth, B. L., E. Lopez, S. Patel and W. K. Kroeze (2000). "The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches?" Neuroscientist **6**(4): 252-262.

SimulationsPlus (2012). "MedChem Studio 3.0. Simulations Plus, Inc. Lancaster, CA."

Spitzer, R. and A. N. Jain (2012). "Surflex-Dock: Docking benchmarks and real-world application." J Comput Aided Mol Des.

Strombergsson, H., M. Lapins, G. J. Kleywegt and J. L. E. S. Wikberg (2010). "Towards Proteome-Wide Interaction Models Using the Proteochemometrics Approach." Molecular Informatics **29**(6-7): 499-508.

Taboureau, O., S. K. Nielsen, K. Audouze, N. Weinhold, D. Edsgard, F. S. Roque, I. Kouskoumvekaki, A. Bora, R. Curpan, T. S. Jensen, S. Brunak and T. I. Oprea

(2011). "ChemProt: a disease chemical biology database." Nucleic Acids Res **39**(Database issue): D367-372.

Tanabe, M. and M. Kanehisa (2012). "Using the KEGG Database Resource." Curr Protoc Bioinformatics **Chapter 1**: Unit1 12.

Tropsha, A. (2010). "Best Practices for QSAR Model Development, Validation, and Exploitation." Molecular Informatics **29**(6-7): 476-488.

Turner, D. B., S. M. Tyrrell and P. Willett (1997). "Rapid Quantification of Molecular Diversity for Selective Database Acquisition." J Chem Inf Comput Sci **37**(1): 18-22.

UniProt, C. (2012). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Res **40**(Database issue): D71-75.

Vidal, D. and J. Mestres (2010). "In Silico Receptorome Screening of Antipsychotic Drugs." Molecular Informatics **29**(6-7): 543-551.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant (2011). "PubChem's BioAssay Database." Nucleic Acids Res.

7. Annexes

Annexe 1 : Liste des 119 molécules DrugBank utilisées pour la validation du protocole.

DrugBank id	Uniprot AC	Uniprot Name
DB00507	P94692	Pyruvate-ferredoxin oxidoreductase
DB01769	P24941	Cell division protein kinase 2
DB01769	P48147	Prolyl endopeptidase
DB01769	P52699	Beta-lactamase IMP-1
DB01769	P30305	M-phase inducer phosphatase 2
DB01769	P18031	Tyrosine-protein phosphatase non-receptor type 1
DB01789	P43379	Cyclomaltodextrin glucanotransferase
DB01805	P07478	Trypsin-2
DB01805	P07477	Trypsin-1
DB01805	P00782	Subtilisin BPN'
DB01872	P84141	Chondroitinase
DB01872	Q54873	Hyaluronate lyase
DB01872	Q59288	Chondroitinase AC
DB01985	Q9RGX9	Beta-agarase A
DB02018	P20231	Tryptase beta-2
DB02142	P22256	4-aminobutyrate aminotransferase
DB02142	P10724	Alanine racemase
DB02142	P19938	D-alanine aminotransferase
DB02142	P06986	Histidinol-phosphate aminotransferase
DB02164	Q9UJM8	Hydroxyacid oxidase 1
DB02561	Q9HYN5	Fucose-binding lectin PA-IIL
DB02713	O31616	Glycine oxidase
DB02725	P12995	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
DB03205	P13650	Quinoprotein glucose dehydrogenase-B
DB03205	P27505	Pyrroloquinoline-quinone synthase
DB03382	P48147	Prolyl endopeptidase
DB03382	O67040	Exopolyphosphatase
DB03382	P37062	NADH peroxidase
DB03584	P37698	Endoglucanase F
DB03584	O85465	Endoglucanase 5A
DB03584	Q79G13	POSSIBLE CELLULASE CELA1
DB03709	O06934	UDP-galactopyranose mutase
DB03709	P00747	Plasminogen
DB03709	P41052	Membrane-bound lytic murein transglycosylase B
DB03760	Q9WY54	Aminomethyltransferase
DB03801	P24941	Cell division protein kinase 2
DB03801	P10724	Alanine racemase
DB03801	P15925	Folylpolyglutamate synthase
DB03801	P39377	Isoaspartyl dipeptidase
DB03801	P14900	UDP-N-acetylmuramoylalanine--D-glutamate ligase

DB03814	P0A9B2	Glyceraldehyde-3-phosphate dehydrogenase A
DB03814	P09211	Glutathione S-transferase P
DB03814	P23946	Chymase
DB03814	Q88H32	Ornithine cyclodeaminase
DB03814	P28845	Corticosteroid 11-beta-dehydrogenase isozyme 1
DB03814	P20932	L(+)-mandelate dehydrogenase
DB03814	Q9X5N2	Prolyl endopeptidase Pep
DB03814	P17169	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]
DB03814	P06276	Cholinesterase
DB03814	P06276	Cholinesterase
DB04077	Q9NZK7	Group IIE secretory phospholipase A2
DB04077	P84077	ADP-ribosylation factor 1
DB04077	P09211	Glutathione S-transferase P
DB04077	O60760	Glutathione-requiring prostaglandin D synthase
DB04077	O14717	tRNA
DB04077	P37173	TGF-beta receptor type-2
DB04077	P00325	Alcohol dehydrogenase 1B
DB04077	P23367	DNA mismatch repair protein mutL
DB04077	P05181	Cytochrome P450 2E1
DB04077	P08183	Multidrug resistance protein 1
DB04138	P38489	Oxygen-insensitive NAD(P)H nitroreductase
DB04447	Q9UQM7	Calcium/calmodulin-dependent protein kinase type II alpha chain
DB04447	P96618	Holo-[acyl-carrier-protein] synthase
DB04447	P04035	3-hydroxy-3-methylglutaryl-coenzyme A reductase
DB04447	O43708	Maleylacetoacetate isomerase
DB04447	P21340	Protease synthase and sporulation negative regulatory protein PAI 1
DB04687	Q93LD7	Phosphotriesterase
DB06952	Q51948	2-hydroxychromene-2-carboxylate isomerase
DB08261	P0A3R9	Neocarzinostatin
DB08266	Q5SHR6	DNA-directed RNA polymerase subunit alpha
DB08432	P00582	DNA polymerase I
DB08605	P0AEK4	Enoyl-[acyl-carrier-protein] reductase [NADH]
DB00296	Q9Y5Y9	Sodium channel protein type 10 subunit alpha
DB00296	P10635	Cytochrome P450 2D6
DB00296	P08684	Cytochrome P450 3A4
DB00296	P20813	Cytochrome P450 2B6
DB00296	P05177	Cytochrome P450 1A2
DB00961	Q9Y5Y9	Sodium channel protein type 10 subunit alpha
DB01031	P00918	Carbonic anhydrase 2
DB01031	P00915	Carbonic anhydrase 1
DB01255	P35368	Alpha-1B adrenergic receptor
DB01260	P04150	Glucocorticoid receptor
DB01411	Q9Y271	Cysteinyl leukotriene receptor 1
DB01411	P19838	Nuclear factor NF-kappa-B p105 subunit
DB01411	P01375	Tumor necrosis factor

DB01411	P08684	Cytochrome P450 3A4
DB01411	P11712	Cytochrome P450 2C9
DB01442	P23975	Sodium-dependent noradrenaline transporter
DB01442	Q05940	Synaptic vesicular amine transporter
DB01442	P31645	Sodium-dependent serotonin transporter
DB01442	P27338	Amine oxidase [flavin-containing] B
DB01442	P21397	Amine oxidase [flavin-containing] A
DB01649	Q96C86	Scavenger mRNA-decapping enzyme DcpS
DB01738	P05771	Protein kinase C beta type
DB01766	Q13526	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1
DB01766	P00734	Prothrombin
DB01827	P17931	Galectin-3
DB01939	P07477	Trypsin-1
DB02108	P07858	Cathepsin B
DB02112	P07477	Trypsin-1
DB02288	P07477	Trypsin-1
DB02354	P07477	Trypsin-1
DB02371	P09871	Complement C1s subcomponent
DB02685	P07858	Cathepsin B
DB02872	Q00987	Ubiquitin-protein ligase E3 Mdm2
DB02898	P24941	Cell division protein kinase 2
DB03016	P07477	Trypsin-1
DB03024	P00797	Renin
DB03028	P07339	Cathepsin D
DB03096	P07339	Cathepsin D
DB03253	P27487	Dipeptidyl peptidase 4
DB03314	P09488	Glutathione S-transferase Mu 1
DB03373	P07477	Trypsin-1
DB03384	P55210	Caspase-7
DB03683	P14780	Matrix metalloproteinase-9
DB03729	P00749	Urokinase-type plasminogen activator
DB03959	P05230	Heparin-binding growth factor 1
DB03959	Q9Y663	Heparan sulfate glucosamine 3-O-sulfotransferase 3A1
DB03963	P16753	Capsid protein P40
DB03978	P42790	Pseudomonalisin
DB04058	P15085	Carboxypeptidase A1
DB04059	P00749	Urokinase-type plasminogen activator
DB04107	P07477	Trypsin-1
DB04144	Q00987	Ubiquitin-protein ligase E3 Mdm2
DB04215	P07477	Trypsin-1
DB04316	P15085	Carboxypeptidase A1
DB04336	P07477	Trypsin-1
DB04442	P07477	Trypsin-1
DB04491	P27487	Dipeptidyl peptidase 4
DB04491	P00782	Subtilisin BPN'

DB04664	P20813	Cytochrome P450 2B6
DB04673	P00742	Coagulation factor X
DB06262	P18825	Alpha-2C adrenergic receptor
DB06262	P25100	Alpha-1D adrenergic receptor
DB06262	P18089	Alpha-2B adrenergic receptor
DB06262	P08588	Beta-1 adrenergic receptor
DB06262	P35368	Alpha-1B adrenergic receptor
DB06262	P07550	Beta-2 adrenergic receptor
DB06262	P13945	Beta-3 adrenergic receptor
DB06262	P08913	Alpha-2A adrenergic receptor
DB06262	P23975	Sodium-dependent noradrenaline transporter
DB06838	P00734	Prothrombin
DB06840	P07477	Trypsin-1
DB06844	P24941	Cell division protein kinase 2
DB06844	P20248	Cyclin-A2
DB06857	P00749	Urokinase-type plasminogen activator
DB06880	P27487	Dipeptidyl peptidase 4
DB06898	Q15596	Nuclear receptor coactivator 2
DB06898	P03372	Estrogen receptor
DB06918	P07477	Trypsin-1
DB06923	P07477	Trypsin-1
DB07026	P39900	Macrophage metalloelastase
DB07165	P00734	Prothrombin
DB07178	P0A5Y6	Enoyl-[acyl-carrier-protein] reductase [NADH]
DB07353	P00734	Prothrombin
DB07356	P27487	Dipeptidyl peptidase 4
DB07440	P00734	Prothrombin
DB07521	P00734	Prothrombin
DB07605	P00742	Coagulation factor X
DB07905	P28482	Mitogen-activated protein kinase 1
DB08024	P27487	Dipeptidyl peptidase 4
DB08171	P0A574	3-oxoacyl-[acyl-carrier-protein] synthase 3
DB08184	P07477	Trypsin-1
DB08247	P24941	Cell division protein kinase 2
DB08265	P0AEK4	Enoyl-[acyl-carrier-protein] reductase [NADH]
DB08270	P43235	Cathepsin K
DB08292	P08238	Heat shock protein HSP 90-beta
DB08309	P24941	Cell division protein kinase 2
DB08309	P20248	Cyclin-A2
DB08346	P08238	Heat shock protein HSP 90-beta
DB08398	Q15596	Nuclear receptor coactivator 2
DB08398	P03372	Estrogen receptor
DB08539	P24941	Cell division protein kinase 2
DB08745	P00742	Coagulation factor X
DB08746	P00742	Coagulation factor X

DB08751	P01112	GTPase HRas
DB08752	P25774	Cathepsin S
DB08755	P25774	Cathepsin S
DB08783	P18031	Tyrosine-protein phosphatase non-receptor type 1
DB00865	Q05940	Synaptic vesicular amine transporter
DB00865	P08913	Alpha-2A adrenergic receptor
DB00865	P08684	Cytochrome P450 3A4
DB00865	P20813	Cytochrome P450 2B6
DB00937	P23975	Sodium-dependent noradrenaline transporter
DB00990	P11511	Cytochrome P450 19A1
DB00990	P11511	Cytochrome P450 19A1
DB00990	P08684	Cytochrome P450 3A4
DB01191	P28335	5-hydroxytryptamine 2C receptor
DB01191	P31645	Sodium-dependent serotonin transporter
DB01191	P33261	Cytochrome P450 2C19
DB01191	P10635	Cytochrome P450 2D6
DB01191	P05181	Cytochrome P450 2E1
DB01191	P11712	Cytochrome P450 2C9
DB01191	P05177	Cytochrome P450 1A2
DB01191	P11509	Cytochrome P450 2A6
DB01429	Q14524	Sodium channel protein type 5 subunit alpha
DB01429	P62158	Calmodulin
DB01429	P10635	Cytochrome P450 2D6
DB08363	P04637	Cellular tumor antigen p53
DB00349	P78334	Gamma-aminobutyric acid receptor subunit epsilon
DB00349	Q99928	Gamma-aminobutyric acid receptor subunit gamma-3
DB00349	O14764	Gamma-aminobutyric acid receptor subunit delta
DB00349	O00591	Gamma-aminobutyric acid receptor subunit pi
DB00349	P31644	Gamma-aminobutyric-acid receptor subunit alpha-5
DB00349	Q8N1C3	Gamma-aminobutyric acid receptor subunit gamma-1
DB00349	P18505	Gamma-aminobutyric-acid receptor subunit beta-1
DB00349	P33261	Cytochrome P450 2C19
DB00349	P33260	Cytochrome P450 2C18
DB00349	P08684	Cytochrome P450 3A4
DB00349	P20813	Cytochrome P450 2B6
DB00354	P35367	Histamine H1 receptor
DB00354	P11229	Muscarinic acetylcholine receptor M1
DB01012	P41180	Extracellular calcium-sensing receptor
DB01012	P10635	Cytochrome P450 2D6
DB01012	P08684	Cytochrome P450 3A4
DB01012	P05177	Cytochrome P450 1A2
DB01070	P11473	Vitamin D3 receptor
DB01237	P35367	Histamine H1 receptor
DB01439	P35372	Mu-type opioid receptor
DB01439	P41145	Kappa-type opioid receptor

DB01439	P41143	Delta-type opioid receptor

Conclusion

Durant cette thèse, j'ai eu l'occasion de travailler sur une thématique fort intéressante qui est la chémogénomique et plus précisément le profilage biologique. Ceci m'a permis d'apprendre à manipuler des millions de données et à les analyser. Grace aux projets du laboratoire, j'ai acquis un savoir-faire dans les méthodes d'apprentissages ainsi que dans la plupart des outils de chémoinformatique et de recherche de médicaments assistée par ordinateur.

J'ai eu le privilège d'évoluer au sein d'un laboratoire qui m'a fourni les outils et l'expertise nécessaire pour mener à bien les projets qui m'étaient confiés. J'ai notamment eu l'occasion de participer à des manifestations scientifiques de renommé mondiale en chémoinformatique comme le 18ème EuroQSAR en 2010 et les deux dernières écoles d'été de chémoinformatique (2010 et 2012).

Les différents projets auxquels j'ai participé ont abouti à la création d'un protocole de profilage automatique qui sélectionne la méthode de criblage virtuelle la mieux adaptée selon les propriétés du ligand à profiler et de la cible à cribler. Une implémentation de ce protocole dans le logiciel Pipeline Pilot (http://accelrys.com/products/pipeline-pilot) est en cours. Ceci nous permettra de profiler les molécules à travers une interface web générée à l'aide de l'outil *webport* de Pipeline Pilot.

Il serait par ailleurs intéressant d'utiliser des bases comme SIDER (Kuhn *et al.* 2006) qui répertorient les effets secondaires pour 996 médicaments avec des méthodes d'apprentissage à l'aide du profil biologique obtenu par notre protocole. En effet, on peut utiliser ce profile comme descripteur pour chaque médicament et essayer de le relier à chaque effet secondaire à l'aide d'une machine d'apprentissage.

Références :

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* **2010,** 6:343. Epub 2010 Jan 19.