Modélisation des interactions protéines–effecteurs et développement de méthodes de prédiction de ces interactions

Natacha Cerisier

Université de Paris

École doctorale Pierre Louis de Santé Publique à Paris :

Épidémiologie et Sciences de l'Information Biomédicale (ED 393)

Laboratoire Biologie Fonctionnelle Adaptative, INSERM UMR 1133, équipe « Approches computationnelles pour le profilage pharmacologique »

Modélisation des interactions protéines–effecteurs et développement de méthodes de prédiction de ces interactions.

Par Natacha CERISIER

Thèse de doctorat de Bioinformatique

Dirigée par Pr Anne-Claude CAMPROUX

Présentée et soutenue publiquement le 4 novembre 2019

Pr Bernard OFFMANN	Université De Nantes	Rapporteur
Pr Pascal BONNET	Université d'Orléans	Rapporteur
Dr Claire MINOLETTI	Sanofi-Aventis	Examinatrice
Dr Jean-Christophe GELLY	Université Paris Diderot	Examinateur
Pr Guillaume ACHAZ	Université Paris Diderot	Examinateur
Pr Anne-Claude CAMPROUX	Université Paris Diderot	Directrice de thèse

RESUME

Modélisation des interactions protéines–effecteurs et développement de méthodes de prédiction de ces interactions.

Le processus de développement d'un médicament est long et coûteux puisqu'il dure plusieurs dizaine d'années. Le principal motif de rejet des médicaments pour la mise sur le marché est l'apparition des effets secondaires. De nos jours de nombreuses méthodes computationnelles sont développées pour éviter au maximum ces effets qui peuvent être dangereux. Il a maintenant été démontré que certains médicaments peuvent interagir avec plusieurs cibles différentes, entraînant ces effets secondaires. Ces effets peuvent être de nature négative, menant à un retrait du marché, ou positive, menant à une réutilisation de ce médicament pour une autre pathologie. Afin de mieux comprendre les mécanismes des effets secondaires, il est primordial de comprendre les mécanismes complexes des interactions entre les médicaments et leur cible au niveau d'un site de liaison dans les phases préliminaires de découverte de médicaments.

Ce projet de thèse a pour but de contribuer à améliorer la compréhension des mécanismes de liaison entre des candidats médicaments et des cibles, en quantifiant les cibles capables d'interagir avec plusieurs candidats médicaments et enfin en développant un outil de prédiction de ces interactions à l'aide d'outils de bio-informatique structurale et statistiques. Nous avons travaillé en utilisant les informations structurales disponibles sur les cibles pour étudier (i) les différentes caractéristiques physico chimiques et géométriques des sites de liaison et des molécules à l'aide de descripteurs en prenant une famille d'intérêt thérapeutique pour exemple : les urokinases, (ii) la capacité des cibles à interagir avec plusieurs candidats médicaments et leur propriétés particulières et enfin (iii) un protocole statistique permettant de prédire les interactions entre cibles et candidats médicaments est en cours de développement.

(i) Nos résultats ont montré qu'il existe des méthodes efficaces pour détecter et caractériser les sites de liaison, ainsi que pour caractériser les molécules qui s'y lient. Nous avons pu développer une méthode qui établit les correspondances entre les profils des sites de liaison et des molécules. Cette méthode a été testée sur une famille de protéines intervenant dans de nombreuses voies de signalisation des cancers et dont de nombreuses structures 3D sont disponibles. Cette étude nous a permis de mieux comprendre les propriétés importantes dans les interactions entre les protéines et les candidats médicaments.

(ii) Nous avons aussi mis en évidence qu'une majorité des protéines (80 %) pouvaient interagir avec plusieurs candidats médicaments au niveau de sites de liaison dits promiscuous. Cette promiscuité est très étudiée pour les molécules mais peu pour les sites des liaisons des protéines. La taille, l'hydrophobicité et la flexibilité sont les propriétés déterminantes des sites de liaison et des molécules dans les interactions. Nous avons mis en avant l'importance de la promiscuité des sites de liaison dans l'étude et la compréhension de la polypharmacologie.

(iii) Enfin, un protocole de prédiction des interactions est en cours de développement. Il se base sur les approches protéochémométriques développées précédemment fournissant des profils d'interaction et leur comparaison, et propose une chimiothèque réduite de composés «candidat-médicaments» commercialisés, pour une cible déterminée. Ce protocole n'est pas entièrement automatisé mais a déjà été testé sur les urokinases et est en cours sur la cible NS1, de façon exploratoire.

Mots-clefs : interactions, prédiction, profilage statistique, polypharmacologie, promiscuité, médicaments, effets secondaires, site de liaison, ligand.

ABSTRACT

Modeling of protein–effector interactions and development of methods for predicting these interactions.

The process of developing a drug is long and costly as it lasts for several decades. The main reason for rejecting drugs of marketing approval is the side effects. Nowadays many computational methods are developed to avoid these effects, which can be dangerous. It has now been shown that some drugs can interact with many different targets, causing these side effects. These effects can be negative, leading to withdrawal from the market, or positive, leading to reuse of this drug for another pathology. In order to better understand the mechanisms of side effects, it is important to understand the complex mechanisms of interactions between drugs and their target in binding site in the early stages of drug discovery.

The purpose of this thesis project is to contribute to improving the understanding of the binding mechanisms between drug and targets by quantifying the targets able to interact with several drug candidates and characterizing drug-target interactions. Finally, we developed a tool for predicting these interactions using structural bioinformatics tools and statistics. To do this, we have worked using available structural information on targets to study (i) the different physicochemical and geometric characteristics of both binding sites and molecules using descriptors by taking a family of therapeutic interest for example, the urokinase, (ii) the ability of targets to interact with only one or several drug candidates and their specific properties, and finally (iii) a statistical protocol for predicting interactions between targets and drug candidates, that is under development.

(i) Our results have shown that there are effective methods for detecting and characterizing binding sites, as well as for characterizing molecules. We have been able to develop a method that matches the profiles of binding sites and molecules. This method has been tested on a family of proteins involved in many cancer signaling pathways and many 3D structures are available. This study has allowed us to better understand the important properties in the interactions between proteins and drug candidates.

(ii) We also found that a majority of binding sites (80 %) could interact with several drug candidates, so-called promiscuous binding sites, using a database of high-quality complexes. This promiscuity is very much studied for molecules but little less for protein binding sites. Size, hydrophobicity and flexibility are the determining properties of binding sites and molecules in interactions. This high binding site promiscuity is therefore an important factor to take into account in the understanding of polypharmacology.

(iii) Finally, a protocol of interaction prediction is under development. It is based on the previously developed proteochemometric approaches providing interaction profiles, their comparison and proposing a reduced library of " drug-candidate" marketed compounds for a

specific target. This protocol is not fully automated but has already been tested on urokinase and is underway on the NS1 target, in an exploratory manner.

Keywords: interactions, prediction, statistical profiling, polypharmacology, promiscuity, drugs, side effects, binding site, ligand.

REMERCIEMENTS

Tout d'abord je tiens à remercier ma directrice de thèse, le Pr Anne-Claude Camproux de m'avoir permis de réaliser ma thèse, pour la confiance et tout le temps qu'elle m'a accordé durant ces trois ans. Je remercie Dr Bruno Villoutreix, directeur du laboratoire MTi lors de mon arrivée ainsi que Pr Jean-Marie Dupret, directeur du laboratoire BFA.
Merci à l'université Paris Diderot, au Ministère de l'Éducation Nationale de la Recherche et de Technologie et l'École Doctorale 393 Pierre Louis de santé publique pour le financement de ma thèse.

Je tiens particulièrement à remercier le Pr Bernard Offmann et le Pr Pascal Bonnet d'avoir accepté de rapporter ce manuscrit de thèse. Je remercie aussi le Dr Claire Minoletti, le Dr Jean-Christophe Gelly et le Pr Guillaume Achaz d'avoir pris le temps d'examiner mes travaux de thèse.

Un grand merci à tous ceux avec qui j'ai eu la chance de partager le bureau 420 : Inès, Ikram, Dhoha, Baptiste, Alejandro, Dhebia, Sookie, Katia, Pierre, Bryan et Loïc, pour tous ces moments de joie, nos discussions, votre soutien et vos encouragements. Je suis heureuse d'avoir fait un bout de chemin avec vous !

Merci à tous les membres de l'équipe ; merci à toi Leslie de m'avoir donné le goût de la recherche lors de mon premier stage, j'ai beaucoup appris avec toi, c'était un plaisir. Laurence, mille mercis pour ton soutien sans faille durant toutes ces années. Merci Olivier, alias « Professeur », pour les encouragements, très souvent sous forme de carrés de chocolat. Colette, merci pour tes mots toujours sympathiques. Delphine, je te remercie pour ton aide sur NS1 et nos discussions réconfortantes. Merci Michel pour cette collaboration sur DISC et toutes ces discussions scientifiques enrichissantes. Merci à Anne de m'avoir donné goût aux statistiques et de m'avoir permis de m'essayer au monitorat. Merci à Karine aussi pour ses passages, furtifs mais appréciés, au bureau. Merci aux membres du laboratoire, permanents et non-permanents pour votre aide et votre soutien.

Merci aussi à Alexandre et Mélaine, « les anciens », et Manon et Quentin pour votre aide et vos collaborations. Je remercie Tarun, Yassine, Hubert et Nicolas, qui m'ont permis de grandir scientifiquement lors mon stage de M2.
Je remercie aussi les étudiants qui ont fait preuve de patience lors de mes premiers pas en tant que monitrice.

À mes camarades de promos, je voudrais vous remercier du fond du cœur. Laure, Solenne et Léa, je vous dois tellement ! Merci d'être là dans tous les bons comme les mauvais moments. Un merci à la team « BRLA » de P7 pour ces moments sportifs, à Laurent pour m'avoir coachée (et supportée) toutes ces années et à Margaux et Pierre pour avoir rendu mon handball bien plus drôle. Cécile, merci de m'avoir permis de m'évader tous les week-end à dos de cheval, grâce à toi j'ai rencontré deux cavalières et amies en or, Tatiana et Audrey. Merci à Pierre L. et Pierre H. d'avoir relu ce manuscrit avec un œil expert et Guillaume pour tes conseils de rédaction.

Merci à tous mes amis qui, de près ou de loin, ont contribué à cette réussite : Agathe, Lisa & Eliot, Marie & Damien, Jérôme & Delphine, Marie-Laure & Fred, Sam, Eslem, Pauline & Adrien, Alexia, et tous ceux que je n'ai pas cités.

Bien sûr, ce travail n'aurait jamais pu aboutir sans le soutien, l'amour, et les encouragements constants de toute ma famille. Je remercie particulièrement mes parents, pour m'avoir donné toutes les clefs pour réussir. Cette réussite, c'est aussi la vôtre. Merci aussi à toute ma belle-famille pour votre soutien et tous ces moments de joie.

Enfin, je remercie Lazar qui m'a suivi dans cette aventure sans hésiter. Merci pour ton amour inconditionnel, tes mots réconfortants, ton soutien quotidien et pour avoir cru en moi plus que moi-même, depuis le début.

ABREVIATIONS

ANSM : Agence Nationale de Sécurité du Médicament

ADME-TOX : Absorption Distribution Métabolisme Excrétion – Toxicité

AMM : Autorisation de Mise sur le Marché

ARN : Acide Ribonucléique

DBN : *Deep Belief Network*

DISC : *Distributional Sphericity Coefficient*

DSE : Dossier de Santé Électronique

EC : *Enzyme Commission*

FC : Facteurs de Croissance

FP : *Fingerprint*

IAP : Inhibiteur de l'Activateur du Plasminogène

MEC : Matrice Extra-Cellulaire

MMP : Métalloprotéase matricielle

MOAD : *Mother Of All Databases*

MTD : *Multi Target Drug*

PDB : *Protein Data Bank*

RBM : *Restricted Boltzmann Machine*

RMN : Résonnance Magnétique Nucléaire

SIDER : *Side Effect Resource*

TSH : *Thyroid Stimulating Hormon*

uPA : *urokinase Plasminogen Activator*

2D : deux dimensions

3D : tri-dimensions

TABLE DES MATIERES

LISTE DES ARTICLES

1. **Cerisier N**, Regad L, Triki D, Camproux AC, Petitjean M. Cavity Versus Ligand Shape Descriptors: Application to Urokinase Binding Pockets, *J Comput Biol.* 2017 Jun 1. doi: 10.1089/cmb.2017.0061

2. **Cerisier N***, Regad L*, Triki D, Petitjean M, Flatters D, Camproux AC. Statistical Profiling of One Promiscuous Protein Binding Site: Illustrated by Urokinase Catalytic Domain. *Mol Inform.* 2017 Jul 11. doi: 10.1002/minf.201700040. (*co-premiers auteurs)

3. **Cerisier N**, Petitjean M, Regad L, Bayard Q, Réau M, Badel A and Camproux A-C, High Impact: The Role of Promiscuous Binding Site in Polypharmacology, *Molecules*. 2019 Jul 10;24(14). pii: E2529. doi: 10.3390/molecules24142529.

LISTE DES FIGURES

LISTE DES TABLES

INTRODUCTION

Les interactions entre les protéines et les molécules sont à la base de nombreuses voies de signalisation qui régissent le bon fonctionnement des organismes biologiques. Les médicaments sont des molécules aux propriétés particulières qui ont pour cible un certain type de protéines impliquées dans une pathologie. Depuis quelques décennies, de nombreux médicaments sont reconnus pour interagir avec plusieurs types de protéines différents, impliquant des effets secondaires non maîtrisés. Bien qu'imprévus, ces effets peuvent être bénéfiques pour l'organisme et traiter une autre pathologie, la molécule est alors « recyclée » pour une autre utilisation, appelée le *repurposing*. Mais bien souvent ils peuvent être néfastes et conduire à l'arrêt du traitement.

Cette faculté qu'a une molécule à interagir avec des cibles différentes est appelée la promiscuité (*drug promiscuity* en anglais). La découverte de la promiscuité des molécules a chamboulé les processus de développement de médicaments et elle fait l'objet de nombreuses études sous de nombreux angles. Il est désormais admis que la promiscuité des molécules est un phénomène fréquent qui est à l'origine de la polypharmacologie : l'étude et l'utilisation des molécules médicamenteuses qui interagissent avec plusieurs cibles et voies de signalisation.

Depuis peu, le nombre de données disponibles dans les bases de données de structures protéiques tridimensionnelles a fortement augmenté. En effet, au début de ces travaux de thèse, la PDB était composée de 130 000 structures alors qu'elle en compte actuellement plus de 154 000. Profitant de ce nombre grandissant, de plus en plus d'études se portent maintenant sur l'étude des sites de liaison. Il est davantage montré que l'étude des sites de liaison est une approche prometteuse dans la compréhension de la promiscuité des molécules. La caractérisation et la comparaison des sites de liaison sont des axes majeurs. Il a été montré que leurs structures et leur similarité ont une plus grande influence sur la promiscuité des molécules que les propriétés physico-chimiques et la flexibilité des molécules elles-mêmes. Récemment, le concept de promiscuité des cibles a été introduit : une cible peut interagir avec plusieurs molécules différentes, rendant encore plus complexe l'étude de la polypharmacologie.

Cependant, peu d'études se portent sur la caractérisation conjointe des sites de liaison et des molécules qui interagissent conjointement. De nombreux outils sont développés dans le but de prédire ces interactions et donc les effets secondaires probables, mais ne prennent en compte que partiellement les deux composantes de l'interaction et sont directement orientés vers le docking.

Dans ce manuscrit, je décris nos différentes études menées, basées sur la caractérisation physico-chimique et géométrique conjointe des sites de liaison et des molécules dans le but de prédire les interactions.

Tout d'abord, dans une partie état de l'art, je détaille les différentes étapes de la conception de médicament, les principes des effets secondaires, du *repurposing* en donnant des exemples concrets de médicaments, ce qui est déjà connu et les méthodes existantes pour pallier

à leurs effets néfastes ou optimiser les effets bénéfiques découverts. Je présente par la suite les données structurales dont la communauté scientifique dispose et les méthodes pour les obtenir. Ensuite, je décris l'état de l'art portant sur les mécanismes d'interactions entre les protéines et les molécules (effecteurs), la polypharmacologie, les méthodes in silico de caractérisation et de prédiction de ces interactions et des partenaires multiples. Une description plus précise des méthodes existantes portant sur l'étude des poches protéiques est fournie. Les poches, aussi appelées cavités, sont les atomes de la protéine qui sont en interaction avec le(s) ligand(s) et qui correspondent donc au site de liaison.

Dans un second chapitre, je détaille l'étude conjointe des sites de liaison et des molécules correspondantes appliquée à une protéine choisie pour son intérêt thérapeutique : les Urokinases. L'analyse conjointe du double espace nous permet de décrire les deux acteurs de l'interaction et de mieux comprendre leurs mécanismes d'interaction. Une classification hiérarchique des complexes nous a permis de mettre en évidence une correspondance entre les profils des sites de liaison et ceux de leurs ligands. Cette caractérisation des sites de liaison et des molécules actives, combinée aux nombreuses données, nous a permis d'établir des profils types d'interactions existant pour les urokinases, ce qui justifient d'étendre cette approche dans un but de prédiction des partenaires des interactions. Cette application sur une protéine nous a permis d'établir et d'améliorer notre protocole de caractérisation conjointe afin de pouvoir l'étendre à un nombre de données plus large.

Dans le chapitre suivant, je décris notre étude des interactions multipartenaires. Pour cela, la caractérisation conjointe des profils de poches et de ligands a été appliquée sur un jeu de données de haute qualité choisi parmi les grands jeux de données. Cette caractérisation de nombreuses interactions nous a permis de porter une attention particulière aux interactions multipartenaires, phénomène appelé « promiscuité ». Les acteurs d'interactions multipartenaires sont appelés par anglicisme *« promiscuous »*. Elle est déjà fréquemment étudiée pour les ligands, mais encore peu pour les sites de liaison. MOAD, un jeu de données de complexes de haute qualité, nous a permis de quantifier cette promiscuité, pour les sites de liaison. Les propriétés influant sur cette promiscuité des sites de liaison ont été mises en exergue et nous avons démontré que cette dernière, fréquente, contribue fortement à l'explication de la promiscuité des ligands et à la polypharmacologie. Cette caractérisation de la promiscuité des sites de liaison est importante et pourrait avoir des implications dans le processus de développement de médicaments.

Enfin, dans le dernier chapitre, je présente le protocole de prédiction des profils d'interaction à partir de la caractérisation des poches et des ligands, développée précédemment. Ce protocole est en cours de mise en place et d'automatisation, mais il permet de fournir une chimiothèque réduite de molécules commercialisées qui seraient des candidats probables à la liaison d'une structure tridimensionnelle donnée. La drugabilité et la promiscuité des sites de liaison sont bien sûr prises en compte dans l'étude. En plus du détail du protocole, les premières applications de ce protocole sur deux protéines d'intérêt thérapeutiques : l'urokinase, déjà

étudiée précédemment et intervenant dans les voies de signalisation de l'inflammation et une protéine du virus de la grippe (NS1) qui favorise la réplication du virus.

Chapitre 1. ÉTAT DE L'ART

1.1 La découverte de médicaments

1.1.1 Définition de médicament

Selon le code de la Santé publique (Article L5111-1, modifié par la Loi n°2007-248), un médicament est défini comme suit :

« On entend par médicament toute substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que toute substance ou composition pouvant être utilisée chez l'homme ou chez l'animal ou pouvant leur être administrée, en vue d'établir un diagnostic médical ou de restaurer, corriger ou modifier leurs fonctions physiologiques en exerçant une action pharmacologique, immunologique ou métabolique. »

Mais cette définition n'est que le résultat d'une évolution des nombreuses définitions depuis la première apparition des médicaments, jusqu'à aujourd'hui. La première apparition de la notion de médicament est due à Paracelse au XVIe siècle qui recommanda l'utilisation d'un médicament précis pour traiter chaque maladie, relaté dans différents journaux traitant de l'histoire des sciences (Feder 1993; Pagel 1982). Il met aussi en évidence la corrélation entre la dose ingérée et l'effet produit affirmant ainsi que « c'est la dose seule qui fait le poison ».

En 1803, une nouvelle découverte permis une avancée dans le traitement des maladies par médicament : l'extraction de la morphine à partir de végétaux (Wu and Wittick 1977) puis, 50 ans plus tard, de l'aspirine ainsi que d'autres principes actifs toujours utilisés de nos jours comme l'insuline (1920) par Banding et Mc Léod (colauréats du prix Nobel de médecine en 1923).

Au regard des effets possibles de ces molécules, les autorités ont mis en place des codes de bonnes pratiques afin d'assurer la sécurité des personnes à qui les médicaments étaient administrés. Cela a donné lieu notamment à la création de l'actuelle Agence nationale de sécurité du médicament (ANSM) pour la France, qui a pour principale mission d'évaluer, de contrôler, d'inspecter et d'informer les professionnels de santé et les utilisateurs de médicaments (ANSM 2018). Elle assure aussi la sécurité et le bon déroulement de certaines étapes du développement de médicaments.

1.1.2 Les phases de développement

En France, pour qu'une molécule devienne un médicament, il faut compter plus de 12 ans et un coût moyen de plus d'un milliard d'euros. Une fois le besoin précis de développer un médicament identifié, le long processus de recherche et développement peut alors être entamé. Il est composé de 10 étapes avant le lancement puis de plusieurs étapes de pharmacovigilance, résumées en figure 1.

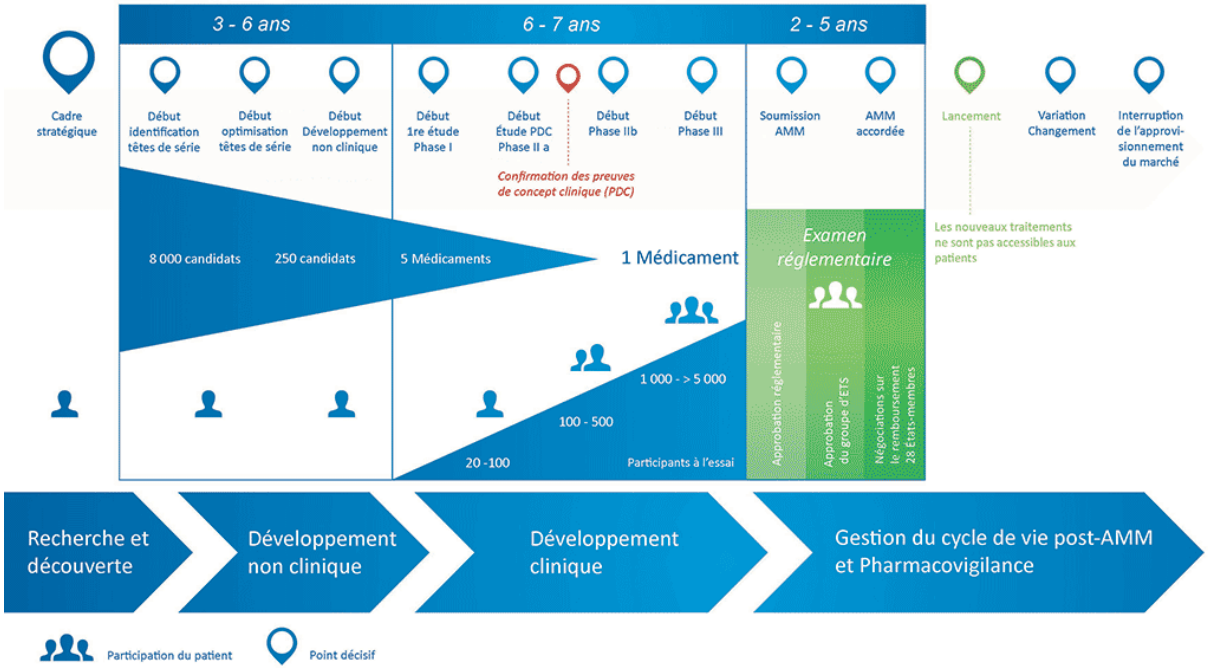


Figure 1 : Détails des étapes de développement de médicaments extraite du site de l'Académie européenne des patients (https://www.eupati.eu). À noter, les nombres de candidats médicaments ou de participants à l'essai peuvent varier selon le type de médicament recherché.

1.1.2.1 L'étape de recherche et découverte

La première étape consiste donc à identifier le besoin de médicament. Dans la plupart des cas, ce besoin provient d'études menées sur une pathologie dont le traitement n'existe pas ou n'est pas optimal. Il convient alors d'identifier les différentes voies de signalisation déficientes impliquées dans cette pathologie et ses acteurs : cellules, protéines et molécules. Ce sont ces acteurs qui seront les cibles des molécules médicamenteuses afin de corriger les dysfonctionnements. Ces dysfonctionnements sont de différents types, par exemple dans les cancers l'organisme produit une trop grande quantité d'une molécule qui signale aux cellules de se diviser anormalement ; dans le cas du diabète, la production d'insuline est insuffisante ou les cellules n'y réagissent pas normalement. Les molécules visant à « réparer » les voies de signalisation doivent donc agir de manière différente et spécifique à chaque dysfonctionnement.

De nos jours, cette étape de recherche peut être complétée par des méthodes bioinformatiques et chémoinformatiques (Lee, Lee, and Kim 2016; Lo et al. 2018; Xia 2017). Ces méthodes permettent de mieux comprendre les mécanismes d'interactions entre les protéines et les molécules. Elles permettent, si les données sont disponibles et les structures résolues, de réduire le temps et les coûts des expérimentations *in vitro*, en testant de nombreuses hypothèses et en orientant les recherches vers des voies plus prometteuses.

1.1.2.2 L'étape de développement non-clinique (préclinique)

Cette étape consiste à trouver la ou les molécule(s) qui auraient une activité sur la cible choisie, appelée dans la figure 1 « identification têtes de série ». De très nombreuses molécules sont testées par criblage à haut débit afin de supprimer les molécules inactives et de réduire le nombre de candidats (Shterev et al. 2018). Ensuite, la phase d'optimisation de ces têtes de série

consiste à modifier les molécules afin d'améliorer leur effet sur la cible voulue. Encore une fois, des méthodes bioinformatiques et chémoinformatiques peuvent intervenir afin de faciliter l'optimisation (Hopkins et al. 2014; Borrotti et al. 2014). Les molécules modifiées sont testées sur des cultures cellulaires et tissulaires afin d'évaluer leur toxicité potentielle, leurs propriétés pharmacologiques et pharmacocinétiques : l'absorption, la distribution, le métabolisme et l'élimination, en plus de la toxicité (appelé aussi propriétés ADME-TOX). Des tests sur les animaux sont ensuite réalisés (Andrade et al. 2016). Ces études sont cruciales pour la suite du développement et aucune méthode bioinformatique ne peut se substituer à cette étude puisque les molécules doivent être testées sur des organismes vivants (hors homme). Elles permettent aussi de déterminer les doses d'efficacité et de toxicité des molécules (Andrade et al. 2016) et les éventuels effets secondaires (détaillés en partie 1.1.3). Sont aussi testés les pouvoirs mutagène et cancérogène des molécules et leurs effets sur la descendance. Cette phase réduit fortement le nombre de candidats médicaments et seulement un médicament sur quinze passe au développement clinique.

1.1.2.3 L'étape de développement clinique

Durant les essais cliniques, découpés en trois phases, la molécule est testée chez l'homme (sujets volontaires). Ces tests sont régis très strictement en France par la Loi Huriet du 20 décembre 1988. Les trois phases se déroulent comme suit :

- Phase I : étude de la tolérance. Différentes concentrations sont administrées à des sujets sains volontaires jusqu'à l'apparition des signes d'intolérance. Cette phase permet aussi de tester les propriétés ADME-TOX chez l'homme et de déterminer la dose à utiliser dans la phase suivante (Storer 1989). Dans le cas particulier des médicaments oncologiques, ces derniers peuvent être testés en phase I sur des sujets malades, dont la molécule est l'unique traitement potentiel.
- Phase II : étude de l'efficacité. Elle-même séparée en deux parties, la phase II consiste à tester la posologie recommandée en phase précédente sur des sujets sains (phase II a) pour valider les preuves de concept clinique, et sur des sujets modérément atteints par la maladie ciblée (phase II b). Cette phase II permet de mieux étudier les propriétés pharmacologiques et pharmacocinétique du candidat médicament et de préciser la posologie optimale pour laquelle le ratio entre l'effet désiré et les effets secondaires est le meilleur. Dans le cas des médicaments oncologiques, la phase II peut être réalisée sur des sujets malades.
- Phase III : essais comparatifs. La posologie optimale misz en place en phase II est testée en comparaison avec les autres traitements reconnus efficaces pour cette maladie ou bien un placebo si aucun autre traitement n'existe (Buyse 2016). Durant cette phase, le candidat médicament est testé sur un grand nombre de patients malades (jusqu'à plusieurs milliers). Si son efficacité est prouvée, il peut alors être qualifié de médicament et prétendre à une autorisation de mise sur le marché (AMM). Le principal but de la phase III est de calculer le ratio bénéfice/risque.

La principale cause d'échec des essais cliniques est le manque d'efficacité des molécules (Harrison 2016; Fogel 2018).

1.1.2.4 L'étape administrative d'Autorisation de Mise sur le Marché

La réglementation complexe régissant la conception du médicament impose une étape administrative qui peut durer entre 2 et 5 ans. Elle consiste à présenter un dossier de plusieurs milliers de pages auprès des autorités compétentes (en France l'ANSM) qui évaluera le médicament : résultats des essais précliniques et cliniques, rapport bénéfice/risque, etc.

Il existe quelques dérogations aux AMM : elles peuvent être conditionnelles (autorisation d'un an), exceptionnelles (dossier incomplet, mais médicament nécessaire), accélérées (intérêt majeur de santé publique) ou temporaire d'utilisation (AMM non française pour des maladies sans traitement). Une AMM n'est pas définitive et peut être suspendue ou retirée à tout moment (article R. 5121-47 du Code de la Santé Publique) pour des raisons de santé, économiques ou simplement parce que le médicament n'a pas été commercialisé pendant trois ans (articles R. 5121-36-1 et R. 5121-102 du code de la santé publique, issus respectivement des décrets n° 2008-435 et n° 2008-436 du 6 mai 2008).

1.1.2.5 L'étape de pharmacovigilance et suivi AMM

Le contrôle des médicaments ne s'arrête pas à son lancement sur le marché, les autorités continuent de surveiller leurs usages et de prévenir leurs effets secondaires. Pour ce faire, des centres de pharmacovigilance ont été mis en place. Ils procèdent à des prises de mesures collectives, des enquêtes, des analyses de risques, etc.

Un des exemples de l'action de la pharmacovigilance est celui du Médiator® (molécule benfluorex), approuvé en 1976 pour traiter le diabète de type II, utilisé aussi comme traitement amaigrissant, mais retiré en 2009 en raison de ses effets secondaires morbides impactant la fonction cardio-vasculaire (A. Weill et al. 2010).

1.1.3 Les effets secondaires et le *repurposing*

Un effet secondaire est une réaction involontaire et souvent imprévue suivant l'administration à dose thérapeutique d'un médicament. Les premières notions d'effet secondaire sont apparues avec les premiers développements de médicaments, au début du XXe siècle. Il y a plusieurs types d'effets secondaires : ils peuvent être indésirables, bénéfiques ou neutres (c'est-à-dire ni néfastes, ni bénéfiques). Les effets secondaires peuvent être classés selon 3 critères :

- Leur nature : aucune spécificité d'organe, réaction aiguë, subaiguë ou chronique, bénigne ou grave, précoce ou tardive,
- Leur gravité : modérée (arrêt du médicament à discuter), sévère (arrêt du médicament, nécessite des soins supplémentaires), grave (arrêt du médicament, menace pour la vie voire décès),
- Leur fréquence (que l'on retrouve sur les notices explicatives) : très fréquent (≥ 1⁄10), fréquent (1⁄10 à 1⁄100), peu fréquent (1⁄100 à 1⁄1000), rare (1⁄1000 à 1⁄10000) et très rare (< 1⁄10000).

Comme mentionné dans la section précédente, les effets secondaires sont déterminants lors du processus de développement du médicament et sont une des causes de nombreux échecs de la mise sur le marché (Fogel 2018).
Concernant les mécanismes qui régissent les effets secondaires, tous ne sont pas connus, c'est pourquoi la plupart des effets secondaires sont découverts « par hasard ». Ils résultent d'une interaction secondaire sur une cible secondaire impliquant un effet différent de celui principalement désiré. De nombreuses études relatent les effets secondaires de nombreux médicaments et de thérapies. Il a été montré que les effets secondaires partagés entre médicaments peuvent être utilisés pour prédire des cibles communes (Campillos et al. 2008).

En France, récemment, de nombreux signalements d'effets secondaires indésirables ont été révélés concernant le médicament Lévothyrox®. Cet évènement, appelé la « crise du Lévothyrox® » survient après l'amélioration de la formule du médicament par le laboratoire Merck. Les changements ne concernaient que les excipients utilisés afin d'assurer la stabilité du médicament et pas la molécule active, la lévothyroxine, prescrite en cas d'hypothyroïdie pour freiner la sécrétion de l'Hormone qui Stimule la Thyroïde, la TSH en anglais (Ianiro et al. 2014). L'ancien excipient, le lactose (à l'origine d'allergies ou d'intolérances) a été remplacé par du mannitol (E421) et de l'acide citrique, substances sans danger et déjà utilisées dans d'autres médicaments. Plusieurs hypothèses sont alors émises sur l'origine des effets secondaires ressentis dont la plus probable est la modification de l'absorption. Les effets secondaires sont les suivants : aggravation possible de troubles cardiaques et réactions allergiques. Les effets dus à une modification de la dose absorbée sont les suivants : tachycardie, tremblements, insomnie, excitabilité, fièvre, sueurs, nausées, vomissements, amaigrissement rapide, diarrhée (pour l'hyperthyroïdie) et fatigue, déprime et sautes d'humeur, difficulté de concentration, troubles de la parole, cheveux et ongles cassants, prise de poids inexpliquée, ballonnement et constipation, frisson et rythme cardiaque ralenti (symptômes de l'hypothyroïdie). A ce jour, l'ANSM n'a pas donné d'explication sur les raisons de cette augmentation des effets secondaires ressentis, qui concerne environ 0,75 % des patients traités.

Aussi, une attention particulière est portée aux femmes enceintes à qui on prescrit des médicaments. Les principes actifs peuvent toucher le futur enfant à tous les stades de développement (embryon, fœtus et enfant), considéré comme effet secondaire du médicament, on dit alors que le médicament est tératogène. Comme exemple, l'isotrétinoïne (Revue Prescrire 2005) (traitement de l'acné) ou encore l'un des énantiomères du thalidomide (Knobloch, Jungck, and Koch 2017)(sédatif et anti-nauséeux), qui engendrent des malformations congénitales graves pour le fœtus.

Quand les effets secondaires sont neutres, voire bénéfiques, il existe maintenant un procédé appelé *« repurposing »* (Ashburn and Thor 2004). Le principe est simple : utiliser un médicament déjà existant en dehors de son champ d'application de l'indication médicale d'origine. Repositionner les médicaments existants pour de nouvelles indications peut offrir un meilleur compromis « risque/bénéfice » par rapport aux autres stratégies de développement de médicaments.

Parmi les exemples de réussite en matière de repositionnement, citons:

- L'aspirine, substance active de nombreux médicaments aux propriétés antalgiques et anti-inflammatoires est aussi utilisée comme antiagrégant plaquettaire (Raber et al. 2019) ;
- Le sildénafil, commercialisé sous le nom de Viagra® était développé pour le traitement de l'angine de poitrine. Les essais cliniques n'ont pas démontré d'effet significatif sur l'angine de poitrine, mais un effet secondaire inattendu : la provocation d'érection. Le laboratoire Pfizer décida donc de repositionner le sildénafil pour le traitement des troubles de l'érection, alors dépourvue de médicament. (Goldstein et al. 2019);
- Le bupropion (Wellbutrin®), un antidépresseur, est également utilisé comme aide au sevrage tabagique (Wilkes 2008).;
- Le minoxidil, utilisé en tant qu'hypertenseur puissant a montré des propriétés intéressantes dans le traitement de la repousse de cheveux (Rossi et al. 2012);

Tous ces exemples, ainsi que ceux répertoriés en table 1, montrent que les mécanismes d'action des molécules ne sont pas tous connus et maîtrisés. Il est donc primordial de développer des méthodes capables de prévoir les effets, qu'ils soient néfastes, afin d'éviter les scandales sanitaires, ou bénéfiques, afin d'améliorer le repositionnement de médicaments existants, d'accélérer leur développement et de traiter de nouvelles maladies.

Table 1 : Exemples de médicaments repositionnés (Yella et al. 2018).

Médicament	Indication originale	Nouvelle indication
Allopurinol	Cancer	Goutte
Amantadine	Grippe	Maladie de Parkinson
Amphotéricine	Antifongique	Leishmaniose
Arsenic	Syphilis	Leucémie
Aspirine	Inflammation, douleur	Antiplaquettaire
Atomexetine	Dépression	TDAH
Bimatoprost	Glaucome	Croissance des cils
Bromocriptine	Maladie de Parkinson	Diabète de type I
Bupropion	Dépression	Sevrage tabagique
Colchicine	Goutte	Péricardite récurrente
Colesevelam	Hyperlipidémie	Diabète de type II
Dapsone	Lèpre	Paludisme
Disulfiram	Alcoolisme	Mélanome
Doxépine	Dépression	Antiprurigineux
Eflornithine	Dépression	TDAH
Finastéride	Hyperplasie bénigne de la prostate	Calvitie masculin
Gabapentine	Épilepsie	Douleur neuropathique
Gemcitabine	Antiviral	Cancer
Lomitapide	Lipidémie	Hypercholestérolémie familiale
Méthotrexate	Cancer	Psoriasis, polyarthrite rhumatoïde

Miltefosine	Cancer	Leishmaniose viscérale
Minoxidil	Hypertension	Chute de cheveux
Naltrexone	Addiction aux opioïdes	Sevrage de l'alcool
Naproxen	Inflammation, douleur	Maladie d'Alzheimer
Nortriptyline	Dépression	Douleur neuropathique
Premetrexed	Mésothéliome	Cancer du poumon
Propranolol	Hypertension	Prophylaxie de la migraine
Raloxifène	Contraceptif	Ostéoporose
Sildénafil	Angine	Dysfonction érectile; hypertension pulmonaire
Thalidomide	Nausées matinales	Lèpre; le myélome multiple
Trétinoïne	Acné	Leucémie
Zidovudine	Cancer	VIH/SIDA
Zileuton	Asthme	Acné

1.1.4 Les méthodes in silico *drug design*

Dans ce contexte, les progrès de la génomique, des méthodes de chémoinformatique et bioinformatique offrent de nouvelles possibilités de recherche et de développement de médicaments. Depuis la première moitié des années 2000, l'importance de l'utilisation de modèles informatiques a été soulignée (Ekins 2004).

Les données telles que l'expression des gènes, les interactions médicament–cible, les réseaux de protéines, les dossiers de santé électroniques (DSE), les rapports d'essais cliniques et les rapports d'événements indésirables liés aux médicaments s'accumulent rapidement et deviennent de plus en plus accessibles et normalisés. Il faut cependant être capable de traiter ces données complexes et bruitées afin d'accélérer la découverte de médicaments et de générer de nouvelles perspectives sur leurs mécanismes, leurs effets indésirables et leurs interactions.

Ces défis peuvent être relevés avec des méthodes de « pharmacologie computationnelle », résumés dans l'étude de Hodos et al., (Hodos et al. 2016) et la figure 2.

Le premier axe que nous allons traiter est la prédiction des interactions entre un médicament et une (ou plusieurs) cible(s). Ces interactions sont fondamentales pour l'efficacité du médicament, mais aussi pour les deux autres axes d'étude (détaillés plus tard). A l'aide d'outils tels que les pharmacophores, la modélisation moléculaire, l'étude de la similarité entre médicaments et cibles ou les réseaux de neurones, des programmes sont développés afin de prédire avec le plus de fiabilité possible ces interactions. On peut notamment citer les outils suivants : PDTPS qui utilise des méthodes de Machine Learning (Meng et al. 2017), ProBiS (Konc and Janežič 2010) et SuperPred qui utilisent la similarité (Nickel et al. 2014), ChemMapper qui propose aussi des applications pour le recyclage des médicaments et les effets secondaires (Gong et al. 2013) et DINIES (Yamanishi et al. 2014) et DTINet (Luo et al. 2017), dont la prédiction est basée sur les réseaux de neurones. La technique de l'amarrage moléculaire (appelé *docking* en anglais) est aussi très utilisée depuis ces vingt dernières années et fait partie intégrante des protocoles de prédiction des interactions protéine–ligand (Leach, Shoichet, and Peishoff 2006).

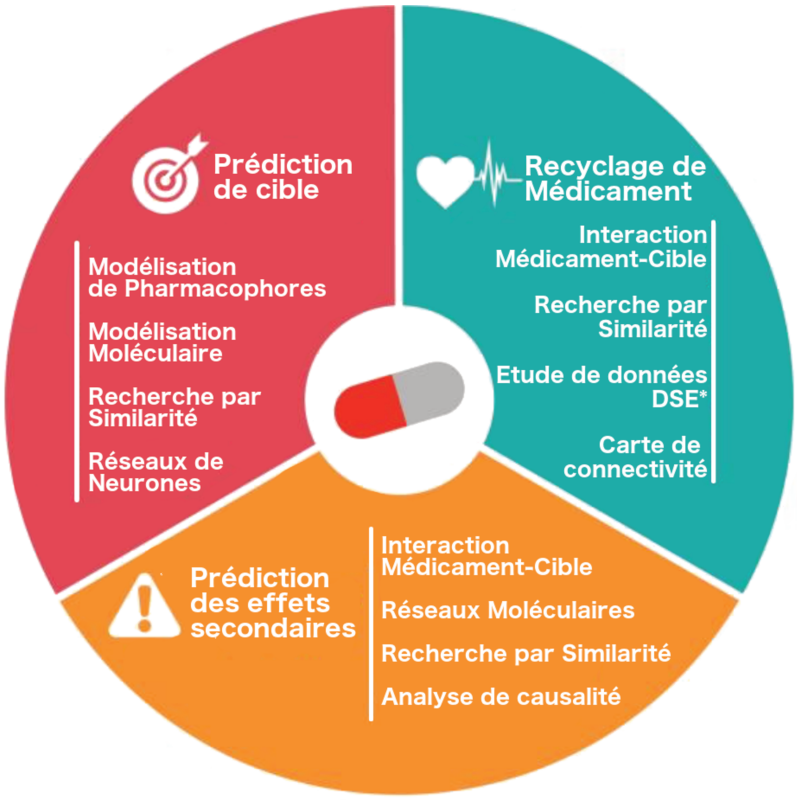


*Figure 2 : Détails des trois axes majeurs de développement de méthodes de pharmacologie computationnelle et des pistes d'étude utilisées. Extraite de Hodos et al. et traduite de l'anglais. *DES : dossier de santé électronique*

La prédiction des effets secondaires, comme discuté plus tôt dans ce manuscrit, est primordiale. Certains outils y sont dédiés comme des bases de données d'effets secondaires telle que the Side Effect Resource (SIDER) (Kuhn et al. 2016) et de nombreuses études traitent de ce sujet telles que les études de Yang et al. (Yang and Agarwal 2011) qui utilisent les effets secondaires des médicaments comme caractéristiques pour prédire leurs indications ou encore Ye et al. (Ye, Liu, and Wei 2014) qui ont identifié de nouvelles indications en partant de l'hypothèse que des profils d'effets secondaires similaires pourraient partager des propriétés thérapeutiques similaires.

Enfin, le recyclage de médicaments est aussi très étudié. Voici un exemple de quelques études récentes sur le sujet : Bisgin et al. (Bisgin et al. 2014) qui ont développé un modèle suggérant des indications alternatives pour des médicaments en prenant en compte les effets secondaires, Li et al. (Li and Lu 2013) qui ont développé une approche systématique en utilisant des chaînes causales dans des réseaux de médicaments-maladies, Xu et al, (Xu et al. 2015) qui valident un repositionnement de médicament en utilisant les DES.

Une revue récente de D. Rognan (Didier Rognan 2017) fourni un grand nombre d'outils très utiles à la conception de pipeline de découverte de médicaments avec des candidats-médicaments innovants et sûrs. Les champs traités sont diverses : Validation de cible, criblage haut-débit, criblage virtuel, identification et optimisation de « têtes de séries », docking, *etc*.

1.2 Les données structurales

Les approches informatiques qui traitent des structures tridimensionnelles des médicaments ou des protéines se basent sur des données structurales.

1.2.1 Les bases de données

La principale source de ces données est la Protein Data Bank (PDB, http://www.rcsb.org/, dernier accès 06-2019) (Berman et al. 2002). C'est une banque de données gratuite qui a été créée en 1971 en tant qu'archive centrale de toutes les données de structures protéiques déterminées expérimentalement. En juin 2019, la PDB comptait 153 328 structures moléculaires biologiques, dont 142 183 protéines (et dont 42 057 protéines humaines), et ce nombre ne cesse d'augmenter. Cette banque de données est utilisée par plus d'un million d'utilisateurs uniques chaque année.

D'autres banques de données existent, mais ne sont que des sous-banques de la PDB. On citera en exemple, the Mother Of All Databases (MOAD) (L. Hu et al. 2005; Ahmed et al. 2015) qui regroupe les noms des structures de protéines de bonne résolution et en interaction avec un ligand biologiquement valide (non cofacteur ni agent cristallisant, ions ni métaux), ou PDBsum (Laskowski et al. 2018) qui regroupe des structures exclusivement protéiques de la PDB ainsi que d'autres informations croisées d'autres bases de données et analyses des structures.

Les travaux de cette thèse sont basés sur des données structurales, aussi appelées fichiers PDB. Ce sont des fichiers textes qui regroupent les coordonnées cartésiennes de chaque atome de tous les composants de la structure 3D (protéine(s), ligand(s), molécule(s) d'eau, ion(s), etc.), ainsi que les informations techniques de la structure (nom, méthode de résolution, organismes, etc.).

1.2.2 Les méthodes de résolution

Il existe plusieurs méthodes pour obtenir les données structurales. La plus utilisée est la cristallographie par rayons X qui représente plus de 89 % des structures de la PDB. Cette méthode est basée sur les propriétés de diffraction des rayons X sur des cristaux de matière (ici, les atomes des protéines et molécules). Un faisceau de rayons X est envoyé sur le cristal de la structure à résoudre. Ce faisceau est dévié dans des directions spécifiques qui, une fois mesurées, permettent d'obtenir des informations précises sur la « maille » (zone de l'espace qui permet de créer le motif cristallin) dont sa densité électronique. Cette densité nous renseigne sur la position moyenne des atomes du cristal et leur nature. Cette technique est limitée par la taille de la protéine à cristalliser. La qualité des modèles issus de la cristallographie à rayons X est évaluée par le critère de résolution. Elle permet de quantifier la netteté du modèle, en Angström. Les structures avec une résolution inférieure ou égale à 2 Å sont considérées de haute qualité (Lamb, Kappock, and Silvaggi 2015). Une résolution supérieure à 4 Å ne permet pas de positionner correctement les atomes lourds.

La seconde méthode de résolution la plus utilisée selon les structures de la PDB (~ 8 %) est la spectroscopie par résonnance magnétique nucléaire (RMN) en solution. Le principe consiste à soumettre un échantillon à un très fort champ magnétique ce qui permet de faire

réagir par rayonnement radiofréquence les noyaux des atomes d'hydrogènes, produisant un signal électrique interprété par un ordinateur. Les informations tirées du signal renseignent sur la nature et le nombre d'atomes (et ceux avoisinant), les liaisons chimiques, les distances interatomiques, les angles dièdres, la conformation et la mobilité moléculaire, etc. Cette méthode est très coûteuse et le coût augmente encore plus lorsqu'il faut résoudre de grosses macromolécules., mais a pour avantage d'être non destructrice et qui permet d'obtenir des informations sur la dynamique des arrangements conformationnels des macromolécules biologiques.

La troisième méthode de résolution de structure la plus utilisée (2 %) est la microscopie électronique. Le principe est similaire au microscope optique, les lentilles optiques étant remplacées par des lentilles électromagnétiques, permettant un zoom bien plus puissant que celui en microscopie optique. L'échantillon est donc traversé ou balayé par un faisceau d'électrons et récupéré sur une plaque électrosensible ou un détecteur d'électrons. En traversant l'échantillon, le faisceau perd des électrons en fonction de la composition de l'échantillon. En le balayant, l'échantillon va émettre des électrons secondaires qui dépendent de la composition de l'échantillon. Le faisceau recueilli permet ensuite de visualiser l'image de l'échantillon. Cette méthode est une alternative à la cristallographie à rayons X pour les protéines membranaires difficiles à cristalliser, mais requiert une préparation très précise des échantillons.

Enfin, il existe d'autres méthodes de résolution de structures 3D qui représentent moins d'un pour cent des structures de la PDB et que je ne détaillerai pas ici : La RMN solide, la cristallographie électronique, la diffraction de fibres, la diffraction de neutrons, ou des méthodes hybrides, mélangeant plusieurs méthodes existantes.

Peu importe la méthode utilisée, la résolution de la structure est un critère de qualité important pour caractériser les interactions. La possibilité de résoudre une interaction avec une molécule est aussi primordiale dans l'apprentissage des interactions dans le but de les prédire.

1.3 Les mécanismes d'interactions protéine-effecteur

La reconnaissance moléculaire est le processus d'interaction des macromolécules biologiques entre elles ou avec de petites molécules. Ces interactions ont une grande spécificité et une affinité élevée, elles forment donc un complexe spécifique. Cela constitue la base de tous les processus biologiques chez les organismes vivants (Du et al. 2016).

Les protéines, qui constituent une catégorie très importante de macromolécules, jouent un grand nombre de rôles, notamment structurels (cytosquelette), mécaniques (muscle), biochimiques (enzymes) et de signalisation cellulaire (hormones). Essentiellement, les protéines réalisent leurs fonctions biologiques par leur(s) interaction(s) physique(s) directe(s) avec d'autres molécules, notamment des protéines, des peptides, des acides nucléiques (ADN et ARN), des membranes, des substrats, des solvants, des métaux et des atomes tels que l'oxygène. Le terme de « ligand » est donc utilisé pour définir toute molécule capable de se lier à une protéine avec une spécificité et une affinité élevées.

1.3.1 Le site de liaison

La partie de la protéine à laquelle se lie le ligand est primordiale dans la réaction. Elle est définie comme le site de liaison et il en existe plusieurs définitions. Pour Ehrt et al, ce sont tous les résidus de la protéine à moins de 5 Å des atomes du ligand (Ehrt, Brinkjost, and Koch 2018). L'outil SiteMap (Halgren 2009) définit un site de liaison, en se basant sur les énergies de van der Waals, comme les « points » reliés de la surface de la protéine et protégé du solvant.

Les sites de liaison peuvent aussi être appelés « poches de liaison » ou « cavités ». Ils peuvent être enfouis, à la surface d'une protéine, à l'interface entre deux chaînes d'une même protéine ou à l'interface entre deux protéines différentes. Des bases de données existent pour référencer les sites de liaison. Par exemple, SitesBase (Gold and Jackson 2005) est une base de données de petits sites de liaison spécialement conçus pour la comparaison structurelle des sites de liaison de ligand connus. La base de données sc-PDB (Desaphy et al. 2015) regroupe une collection de sites de liaison sélectionnés spécifiquement pour leur intérêt pharmacologique. Enfin, il existe un jeu de données de sites de liaison, LigASite, répondant à des critères précis : les sites de liaison doivent (i) être représentatifs des données structurelles connues, (ii) être non redondant, (iii) combiner des informations sur les structures protéiques liées (holo) et non liées (apo), (iv) être biologiquement pertinents, (v) prendre en compte l'unité biologique de la protéine, (vi) considérer les données de toutes les structures holo disponibles pour chaque protéine et enfin (vii) être mis à jour automatiquement (Dessailly et al. 2008).

1.3.2 Les différents types de liaisons (non covalentes)

Il existe différents types de liaisons entre une protéine et un ligand, au sein du site de liaison.

La liaison hydrogène est l'interaction attractive non covalente la plus classique. Cette interaction se fait entre un atome donneur d'hydrogène (noté X–H) et un atome accepteur d'hydrogène (noté Y), formant la liaison X–H•••Y–Z où le Z représente les atomes liés à l'accepteur et les trois points représente la liaison (Arunan et al. 2011). L'atome X doit être plus électronégatif que H, et la force de la liaison hydrogène est proportionnelle à cette électronégativité. La distance entre l'accepteur et le donneur lors d'une liaison hydrogène varie en général entre 2,5 et 3,2 Å, avec un angle variant entre 130–180° et une énergie libre (ΔG) de –4 à –30 kJ/mol (Schaeffer 2008). La table 2 détaille les potentiels groupements donneurs et accepteurs classés selon leur force d'interaction.

Table 2 : Potentiels groupements donneurs et accepteur d'hydrogène classés selon la force de leur interaction. X est un atome donneur, Hal est l'un des halogènes les plus légers et M est un métal de transition. Cette table est extraite et traduite du livre The Practice of Medicinal Chemistry (Schaeffer 2008)

	Donneur	**Accepteur**
Très forte (–30 kJ/mol)	N^+H_3, X^+–H, F–H	CO_2^-, O^-, N^-, F^-
Forte (–15 kJ/mol)	O–H, N–H, Hal–H	O=C, O–H, N, S=C, F–H, Hal−
Faible (–4 kJ/mol)	C–H, S–H, P–H, M–H	C=C, Hal–C, π, S–H, M, Hal–M, Hal–H, Se

La liaison halogène est une variante de la liaison hydrogène, c'est l'atome halogène qui est partagé entre le donneur électrophile et l'accepteur nucléophile. Les halogènes les plus connus sont le fluor (F), le chlore (Cl), le brome (Br), l'iode (I) et l'astate (At). La distance entre l'accepteur et le donneur lors d'une liaison halogène varie en général entre 2,5 et 3,5 Å, avec un angle se rapprochant de 180° et une énergie ΔG de –5 à –180 kJ/mol.

Le pont salin est aussi une autre forme de liaison hydrogène. Ce sont des liaisons hydrogène formées entre un groupe fonctionnel chargé négativement (tel que le groupe latéral issu de l'aspartate ou du glutamate) et un donneur chargé positivement (tel que le groupe latéral de la lysine, de l'arginine ou de l'histidine). Cette liaison est considérée comme très forte et se caractérise par une courte distance entre les atomes en interaction (~ 2,8 Å). Les ponts salins sont des interactions qui apportent une stabilité dans la conformation des protéines (Kurczab et al. 2018).

Parmi les interactions non covalentes, il y a les interactions électrostatiques. Elles se forment entre les cations et les anions qui sont chargés. Elles peuvent être répulsives ou attractives selon les charges : des charges opposées s'attirent et des charges identiques se repoussent. Les interactions électrostatiques non covalentes peuvent être fortes et agir à longue distance bien qu'elles diminuent progressivement avec la distance. Lorsque des interactions électrostatiques se font entre dipôles électriques, ce sont des interactions de van der Waals. Les interactions ioniques font partie des interactions charge-charge. L'énergie de la liaison électrostatique dépend de la nature des deux partenaires : ΔG de –4 à –12 kJ/mol pour une interaction dipôle–dipôle, ΔG de –12 à –20 kJ/mol pour une interaction charge–dipôle et ΔG de –20 à –40 kJ/mol pour une interaction charge–charge.

Une autre interaction non covalente, appelée interaction π, se forme entre un système π et un cation, un anion ou un autre système π. Ces trois interactions s'appellent respectivement cation–π, anion–π et π–*stacking*. Pour les protéines, les systèmes π sont portés par les résidus aromatiques : le tryptophane, la phénylalanine, l'histidine et la tyrosine. Ces interactions jouent un rôle important dans la reconnaissance du ligand par la protéine et dans la structure tridimensionnelle de la protéine (Dougherty 2013).

Enfin, dans les interactions protéine–ligand il y a également les interactions hydrophobes, qui sont principalement dues aux molécules d'eau. L'effet peut aussi être lié à la surface d'interaction non-polaire entre le ligand et la protéine. L'énergie de cette interaction varie entre un ΔG de –50 à –200 J/mol et par $Å^2$ de surface de contact. Les interactions hydrophobes forment souvent des agrégats moléculaires. Le repliement protéique en milieu aqueux est conduit par les interactions hydrophobes : Les chaînes latérales hydrophobes s'éloignent de l'eau, elles seront plus souvent enfouies, alors que celles hydrophiles préfèrent interagir avec l'eau et seront à la surface, en contact avec le solvant. Ce phénomène peut être appliqué à l'interaction entre la protéine et le ligand : si le site de liaison et le ligand sont hydrophobes, ils vont s'agréger et écarter les molécules d'eau qui les séparent.

Toutes ces liaisons sont essentielles à la structure de la protéine et à ses interactions avec d'autres molécules.

1.3.3 Le concept clé-serrure

Un chimiste allemand, Emil Fisher – prix Nobel en 1902 – parle pour la première fois du concept clé-serrure lors de son étude sur la liaison des glucosides sur les enzymes. Il illustre l'interaction entre la protéine et le ligand comme une serrure et une clé qui interagissent spécifiquement entre elles pour que l'interaction aboutisse à une fonction biologique. Découlant de ce concept, un médicament est donc développé spécifiquement pour une cible, intervenant dans une voie de signalisation d'une maladie.

Cependant, ce concept a été remis en question par des études récentes montrant qu'une cible peut interagir avec plusieurs ligands et inversement. Le concept de polypharmacologie a donc émergé ces dernières années (expliqué dans la section 1.3.4).

1.3.4 Le modèle d'ajustement de Koshland

Plus de 60 ans après la première formulation du concept clé-serrure, un autre mécanisme d'interaction a été décrit par Daniel Koshland en 1958 (Koshland 1958). En effet, l'hypothèse du concept clé-serrure n'explique pas tous les cas et ce sont les exceptions qui ont conduit à une révision des théories. Nommé aussi « mécanisme d'adaptation induit », ce modèle a d'abord été appliqué aux enzymes, suggère que la réaction entre l'enzyme et le substrat ne peut se produire qu'après un changement de structure protéique induit par le substrat lui-même, impliquant des changements de conformations de la protéine allant de la forme non-liée à liée.

Ce modèle met donc en avant l'importance de la flexibilité lors de l'interaction entre les protéines et les molécules, qui permettrait aux protéines d'accepter plusieurs ligands. Une étude relate plusieurs cas dans lesquels plusieurs substrats interagissent avec une protéine, certains cas étant régit par le mécanisme clé-serrure, d'autre par le mécanisme d'adaptation induit mais aussi plusieurs protéines qui combinent les deux mécanismes, bien que l'interaction diffère par leur dépendance à chaque mécanisme (Kwon and Park 2019). La connaissance des mécanismes qui entrent en jeu lors des interactions est primordiale pour le développement de médicaments pour cibler ces protéines. Ce mécanisme est compatible avec le concept de polypharmacologie évoqué précédemment.

1.3.5 La polypharmacologie

Une définition simple de la polypharmacologie est la suivante : la conception ou l'utilisation d'agents pharmaceutiques agissant sur plusieurs cibles ou voies de signalisation de maladies. Son utilisation a été suggérée au début des années 2000 par Ekins (Ekins 2004) et elle a été prouvée comme plus efficace par une étude inédite pour traiter une maladie complexe (Kumar, Tiwari, and Sharma 2018). Cette approche a déjà fait ses preuves dans le traitement de maladies complexes similaires telles que le cancer, le VIH et l'hypertension, où elle atteint une efficacité maximale en s'attaquant à plusieurs cibles médicamenteuses.

De la même manière, une même cible peut être liée par plusieurs ligands de natures différentes. Ainsi, des médicaments (parfois indiqués pour des pathologies totalement différentes) peuvent cibler la même protéine. C'est le cas par exemple du récepteur histaminique H1. Il est inhibé par le Zyrtec® (cétirizine) indiqué dans le traitement des allergies

(Malhotra et al. 2019), et il est activé par le Serc® (bétahistine) indiqué dans le traitement des vertiges dans la maladie de Ménière (Barak 2008).

La polypharmacologie est une notion primordiale à prendre en compte dans le développement de médicaments dont les bases sont les interactions multiples. Elle peut aussi limiter les inconvénients résultant de l'utilisation d'un médicament à cible unique ou d'une combinaison de plusieurs médicaments (Anighoro, Bajorath, and Rastelli 2014). Ce type de médicament est appelé « MTD » (*Multi-Target Drug*). Il est aussi primordial de tester, grâce à des modèles in silico puis in vivo, d'autres protéines que celle ciblée par le médicament afin d'éviter les effets secondaires (Ekins 2004).

Une définition plus précise de la polypharmacologie est donnée dans l'étude de Meyers et al. (Meyers et al. 2018). Ils distinguent dans leur étude quatre types de polypharmacologie.

Le premier type est souvent rencontré dans les sondes à petites molécules utilisées en chémobiologie. Deux composés distincts ayant généralement des fonctions et des sélectivités différentes sont liés l'un à l'autre via un « lien » flexible, l'objectif étant que chaque motif de la sonde chimique conserve son activité malgré son rattachement.

Le second type est décrit par des interactions faibles et souvent non spécifiques avec des protéines extérieures à la famille de la protéine cible. Elles sont largement caractérisées dans les criblages de pharmacologie de sécurité in vitro, où la liaison à des cibles multipartenaires peut provoquer des effets secondaires indésirables au médicament à forte dose.

Dans le troisième type de polypharmacologie, les ligands se lient souvent avec une forte affinité à de multiples membres d'une même famille de protéines. Cela est dû à la grande similarité de séquence dans les sites actifs au sein d'une même famille protéique.

Le dernier – et le moins évident – type de polypharmacologie joue un rôle important dans le processus de développement de médicament. Des interactions spécifiques de haute affinité peuvent se produire entre un ligand et des protéines de différentes familles malgré l'absence de similarité entre les sites de liaison ou entre les séquences. C'est à ce dernier type de polypharmacologie qui nous nous intéresserons le plus, car les mécanismes le régissant ne sont pas encore tous compris.

Le développement de MTD appartient au troisième ou au quatrième type de polypharmacologie, selon les familles et les caractéristiques des protéines ciblées. Afin de viser des sites de liaison précisément, il faut que le médicament créé des interactions précises avec le site de liaison. Cela souligne l'importance de caractériser conjointement les sites de liaison et les ligands qui s'y lient.

1.4 Étude in silico des partenaires multiples

Dans les processus de découverte de médicaments, il est primordial de commencer par comprendre les mécanismes d'interaction *a fortiori* lorsqu'elles sont multipartenaires. La caractérisation de tous les acteurs de l'interaction s'impose donc. Lorsque cette caractérisation est conjointe entre les sites de liaison et les ligands, elle apporte des informations supplémentaires par rapport à une caractérisation unilatérale (Pérot et al. 2013). Ces approches

conjointes sont appelées « approches protéochémométriques ». Elles sont d'autant plus informatives lorsque les protéines se lient à plusieurs ligands (ou l'inverse).

1.4.1 Caractérisation des poches

La première étape de caractérisation des poches consiste à les détecter. Il existe de nombreuses méthodes qui permettent de les estimer, se basant sur des algorithmes différents et parfois des critères de seuil.

De nombreux outils existent tels que DeepDrug3D (Pu et al. 2019), conçu pour détecter et classifier les sites de liaison aux nucléotides et à l'hème en utilisant le *deep learning*, CASTp (Tian et al. 2018), qui permet de localiser, délimiter et mesurer des régions superficielles concaves sur des structures tridimensionnelles de protéines en se basant sur la géométrie, 3DLigandSite (Wass, Kelley, and Sternberg 2010), qui prédit des poches à partir de structure ou de séquence et se base sur la similarité des structures ou enfin DoGsite (Volkamer et al. 2010) basé sur une grille qui couvre la zone entourant la protéine avec un filtre de différence de Gauss (DoG).

Au sein du laboratoire, un outil a été développé pour estimer les poches de liaison et prédire si elles sont capables de lier un ligand candidat médicament (appelée drugabilité), PockDrug (Borrel et al. 2015). Leur jeu de données est composé de 113 structures liées et non redondantes extraites du jeu de données NRDLD (Krasowski et al. 2011), et 109 protéines non liées, collectées dans le *Druggable Cavity Directory*, qui ont une protéine équivalente liée. Le serveur Web PockDrug (Hussein et al. 2015), accessible à l'adresse http://pockdrug.rpbs.univ-paris-diderot.fr/cgi-bin/index.py?page=home, permet d'utiliser deux types d'estimation des poches, de les caractériser et de prédire leur drugabilité. Les deux types d'estimation sont les suivants :

- une estimation par proximité au ligand (appelée « prox »), pour les protéines qui sont complexées avec un ligand. La poche sera alors composée de tous les atomes qui se trouvent à une distance choisie du ligand. Ce seuil est généralement entre 4 et 6 Å.
- une estimation par recherche de cavité, pour les protéines non liées. Basé sur la géométrie, l'outil Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) calcule tous les rayons des sphères alpha. Chaque sphère est en contact avec quatre atomes à sa surface sans aucun à l'intérieur, leur rayon étant compris entre le rayon de van der Waals et l'infini (si les quatre atomes sont sur le même plan). Seules les sphères de rayon compris entre 3 et 6 Å sont gardées et classées pour former des poches. Les poches proches sont fusionnées, les petites poches sont éliminées et un score est attribué à chaque poche finale afin d'évaluer sa probabilité de lier un ligand.

PockDrug fourni une description des poches grâce à 52 descripteurs, développés ou reprogrammés dans le laboratoire, dont 36 sont physicochimiques et les 16 autres géométriques pour caractériser chaque poche estimée. La liste des descripteurs complète est décrite dans l'article de Borrel et al. (Borrel et al. 2015). Pour n'en lister que certains, on retrouve les descripteurs physicochimiques d'hydrophobicité (Kyte and Doolittle 1982), d'aromaticité, de polarité, ainsi que les fréquences des atomes (Milletti and Vulpetti 2010). Pour la géométrie, le

rayon, le volume et le nombre d'atomes et de résidus sont quelques-uns des descripteurs utilisés (Petitjean 2014).

Ces descripteurs sont un des moyens les plus utilisés pour décrire les poches, les comparer et analyser les interactions. Ils sont notamment utilisés dans les modèles QSAR afin de relier la structure à l'activité. Ils servent aussi de base à la prédiction de la *drugabilité*, anglicisme définir par Hajduk et al. (Hajduk, Huth, and Fesik 2005), comme étant la capacité d'une poche à être liée par un ligand (aux propriétés candidat médicament) de manière assez forte pour modifier son activité biologique.

L'outil PockDrug se base sur ces descripteurs afin de prédire la drugabilité des poches prédite. Dans les processus de développement de médicaments, la drugabilité des protéines est d'un intérêt majeur (Hussein et al. 2017).

Dans ce manuscrit, un site de liaison est défini par les poches qui le composent. Elles correspondent aux atomes des résidus de la protéine qui sont à une distance inférieure à 5,5 Å du ligand.

1.4.2 Caractérisation des ligands

Comme pour les poches, il existe de nombreux descripteurs qui permettent de caractériser les ligands, en termes de géométrie ou de propriétés physicochimiques. De nombreux outils, logiciels et serveurs sont développés encore récemment afin de calculer ces descripteurs.

Parmi ces propriétés, il a été montré que certaines jouaient un rôle important dans leur faculté à être apparentées à des candidats médicaments, appelés *drug-like* par anglicisme. Toutes les molécules ne peuvent pas être apparentées à des médicaments, certaines propriétés avantagent les molécules alors que d'autres en font des molécules toxiques ou dangereuses qui ne peuvent être considérées comme médicament. Le concept de « candidat médicament » est décrit en 2000 par Lipinski (Lipinski et al. 2001).

Lipinski s'est basé sur un jeu de données de 2200 composés et a étudié leurs caractéristiques physico-chimiques, au regard de leurs propriétés ADME-TOX (Lipinski et al. 2001). Il en a déduit 4 règles, appelées « règles des 5 » que doivent suivre les molécules afin d'être compatibles avec des propriétés « candidat médicament » :

- le nombre de donneurs de liaison hydrogène doit être inférieur ou égal à 5
- le nombre d'accepteurs de liaison hydrogène doit être inférieur ou égal à 10.
- la masse moléculaire doit être inférieure ou égale à 500 Daltons
- le coefficient de partition (logP, qui mesure la solubilité) doit être inférieur ou égal à 5.

L'origine du nom vient du fait que toutes les valeurs sont des multiples de 5. Une étude ultérieure de 10 000 molécules en phase II d'essai clinique a confirmé la validité de ces règles pour les molécules candidats médicaments (Lipinski et al. 2001).

Ces règles sont majoritairement basées sur les propriétés physico-chimiques des molécules. En effet, un nombre élevé de liaisons hydrogène augmente la solubilité dans l'eau et réduit la séparation de la phase aqueuse dans la membrane. Un poids moléculaire élevé réduit quant à lui cette solubilité, en plus de la diffusion à travers les membranes. Un coefficient de

partition élevé réduit aussi cette solubilité. Lorsqu'elle est affectée, l'absorption du médicament est réduite, voire impossible. Il devient donc inefficace et il peut être toxique.

Ces règles ont donc été intégrées aux processus de développement de médicaments. Cependant, elles ne sont à appliquer que pour les mécanismes d'absorption passifs et pour les médicaments à diffusion orale. Ces règles ont été « améliorées » et modifiées en fonction de l'utilisation prévue de la molécule. Par exemple, Veber (Veber et al. 2002) définit qu'un composé apte à la biodisponibilité orale devrait avoir (i) un nombre de liaisons rotatives inférieur à dix, (ii) une surface polaire inférieure à 140 $Å^2$ et (iii) un nombre de liaisons hydrogène (liaison H) donneurs/accepteurs inférieur à 12.

La « règle des trois » (Congreve et al. 2003) est utilisée pour construire des fragments pour la génération de leads (appelés « *lead-like* »), composés qui ont un potentiel à être optimisés pour être candidats médicaments. Cette règle stipule que ces fragments répondent aux règles suivantes : (i) un poids moléculaire inférieur à 300 Da, (ii) une surface polaire inférieure à 60 $Å^2$ (iii) un nombre de donneurs de liaison hydrogène inférieur à trois, (iv) un nombre d'accepteurs de liaison hydrogène inférieur à trois, (v) un coefficient de partition inférieur à trois et (vi) un nombre de liaisons rotatrices inférieure à trois. Les fragments seuls ne se lient que faiblement à la cible biologique, mais leur combinaison peut produire un composé avec une plus grande affinité. L'importance de leur utilisation et leur impact dans le développement de médicaments ont été démontrés (D Rognan 2012).

Ces règles semblent cependant un peu différentes pour les médicaments dont l'administration est ophtalmique, par inhalation ou transdermique (Choy and Prausnitz 2011).

1.4.3 Prédiction des interactions

La prédiction de l'affinité d'interaction entre protéines et ligands constitue un défi majeur dans le processus de découverte de médicaments. L'un des intérêts des méthodes in silico est justement de pouvoir développer des modèles de prédiction de ces interactions, avec un coût financier modéré et une durée moindre comparé aux méthodes *in vitro* et *in vivo*.

La prédiction peut se baser sur les données 2D, mais le nombre grandissant des données 3D implique de plus en plus leur utilisation dans les protocoles. Les méthodes peuvent être variées : *machine learning* ou *deep learning*, méthodes statistiques, pharmacophores, « coupable par association » (basé sur la similarité), criblage moléculaire, etc. Il existe de nombreux outils de prédiction, dont les principaux sont détaillés ici afin d'illustrer leur diversité.

Les outils de prédiction d'interactions sont nombreux, répondant chacun à une problématique spécifique et utilisant des méthodes différentes pour y répondre.

Dans le contexte du nombre grandissant de données et de l'efficacité des méthodes de *Deep Learning* (L. Zhang et al. 2017), BindScope (Skalic et al. 2018) a pour objectif de prédire les liaisons protéine–ligand en caractérisant à la fois la poche de liaison et la pose du ligand. C'est donc l'affinité/l'énergie de l'interaction qui est prédite. Pour ce faire, l'outil les voxélise (transformation en une représentation volumétrique) selon différentes propriétés de type pharmacophorique et entraîne un réseau de neurones de convolution en trois dimensions pour

prédire la probabilité de liaison. Ainsi, le choix d'une fonction de score de docking adaptée, souvent difficile voire arbitraire, est évité tout en proposant une interface simple d'utilisation. L'outil a été entraîné sur plusieurs familles différentes de protéines telles que les protéases, des récepteurs nucléaires, les GPCR, un cytochrome (P450) et plusieurs types d'enzymes, ce qui est varié. Cependant, il faut d'avance fournir les ligands et les protéines à tester, nécessitant une pré-sélection en amont, ce qui représente une limite de l'outil. Pour la plupart de leurs protéines, leur modèle est performant (AUC comprise entre 0,5 et 1 avec une moyenne à 0,88). Concernant les performances, ils estiment leur outil au moins aussi bon (pour la plupart de leurs protéines) qu'un autre outil utilisant les réseaux de neurones.

Encore dans le domaine du *Deep Learning*, l'outil DeepDTIs (Wen et al. 2017) est aussi dédié à la prédiction des interactions protéine–ligand. Cette méthode, appelée *Deep Belief Network* (DBN) est basée sur un empilement de réseaux de neurones artificiels non supervisés, ici, une Machine de Boltzmann Restreinte (RBM en anglais).

La puissance de cet outil réside dans la double possibilité de prédiction : d'une protéine à partir d'un ligand ou l'inverse.

Dans l'esprit de la méthodologie « Relation Structure Activité » des interactions multipartenaires, qui consiste à créer des modèles mettant en relation les propriétés structurales et l'activité biologique d'un composé sur une cible, TargetNet (Yao et al. 2016) prédit les interactions de plusieurs cibles pour une molécule donnée. Ce serveur Web, rapide et ne nécessitant pas d'installation préalable utilise sept types différents de *Fingerprints* pour caractériser les structures moléculaires qui servent de base aux modèles bayésiens. En effet, les 623 modèles de haute qualité, correspondant chacun à une protéine humaine différente, ont été préalablement calculés ainsi que les *Fingerprints* des molécules avec lesquelles chaque protéine interagit. Lors de la soumission d'une nouvelle molécule, le serveur prédit son activité sur les 623 protéines à l'aide des modèles, générant ainsi un profil d'interaction pouvant être utilisé comme vecteur de caractéristiques pour de larges applications. Cette méthode a l'avantage d'être plus rapide que le docking, et de ne nécessiter que des informations chimiques de la structure. Cependant, le nombre restreint des cibles semblent être une limite importante de l'étude. Concernant les performances, les valeurs des AUC sont comparées selon les différents types de *Fingerprints* utilisés. Ce sont les *Fingerprints* ECFP qui semblent prédire au mieux les interactions (AUC entre 0,75 et 1). Les autres *Fingerprints* ont des résultats satisfaisants avec une AUC comprise entre 0,65 et 1. Le jeu de données et les modèles sont mis à disposition et peuvent servir à tester d'autres outils.

Les peptides sont aussi importants dans les processus biologiques et dans les interactions avec les protéines, mais leur grande flexibilité est une difficulté majeure dans la modélisation des interactions. Afin de palier à cette flexibilité, un protocole appelé IRDL (Diharce et al. 2019) a été développé. Il permet, grâce à une méthode itérative de docking, de prédire les interactions peptide-protéine en s'affranchissant des difficultés liées à la flexibilité des peptides, en les divisant en segments et en les recomposant itérativement. En testant de nombreux fragments, il est possible de prédire avec efficacité les peptides qui se lient à une protéine

d'intérêt. Les limitations énoncées par l'équipe sont les suivantes : la position du premier fragment du peptide avant l'itération est cruciale et l'outil commence forcément par l'extrémité C-terminale, forçant le sens de la reconstruction. Il faut aussi que ce premier fragment soit très affin pour le site de liaison, conditionnant totalement le reste de la reconstruction. Enfin, l'équipe met tout en œuvre pour réduire les temps de calcul qui sont très longs. L'approche a tout de même été comparée à deux autres modules de docking et montre de très bonnes performances puisque dans tous les cas testés, IRDL permet de récupérer la pose cristallographique de peptides. De plus, dans 10 cas sur 11, IRDL est capable de proposer dans les 5 premières poses les plus probables, des poses dont l'écart quadratique entre la structure issue du docking et la structure cristallisée ne dépasse pas 2 Å, la limite acceptable étant fixée 2 Å pour les atomes du squelette du peptide et à 3 Å pour tous les atomes du peptide.

Enfin, l'équipe du Pr Rognan a développé un outil appelé IChem (Da Silva, Desaphy, and Rognan 2018). Cet outil très complet permet de faire de nombreuses analyses dont les principales sont l'alignement structural de deux molécules, la comparaison de *Fingerprints*, la détection de cavité et la prédiction de la drugabilité, l'extraction des *Fingerprints* des interactions protéine–ligand, la conversion des modèles d'interaction protéine–ligand en graphes et la prédiction des interactions protéine-protéine biologiquement pertinentes. La nouvelle fonctionnalité intégrée à cet outil permet de prédire les interactions protéine–ligand par la combinaison de plusieurs modules. La méthode consiste à détecter automatiquement les cavités de liaison au ligand, puis à prédire leur drugabilité structurelle, et créer un pharmacophore basé sur la structure des cavités de liaison. Ensuite, un autre module intervient pour aligner les ligands sur les caractéristiques pharmacophoriques dérivées de la cavité afin de trouver les plus pertinents à une éventuelle interaction. Cette méthode est démontrée comme aussi efficace que les méthodes de criblage virtuelles à la pointe de la technologie (Surflex-Dock, etc.). Cet outil combine de manière automatique différents modules et propose ainsi un protocole intégré et rapide de prédiction.

Le nombre, en constante augmentation, d'outils développés est la preuve du caractère actuel et nécessaire de l'amélioration de la prédiction des interactions.

Chapitre 2. CARACTERISATION DES INTERACTIONS

Dans ce chapitre, je décris les méthodes mises en place pour optimiser la caractérisation conjointe des poches et des ligands. C'est la première étape dans l'étude des interactions multipartenaires. Le protocole de caractérisation a été testé sur une famille de protéine d'intérêt thérapeutique : les Urokinases. Cette étude a donné lieu à une publication (section 2.2).

2.1 Exemple des urokinases

L'intérêt des approches protéochémométriques a été décrit dans le chapitre précédent. J'ai donc effectué une analyse conjointe du double espace des protéines et des ligands sur une cible d'intérêt thérapeutique : le site de liaison catalytique à l'inhibiteur de l'activateur humain du plasminogène de type urokinase (appelé uPA). Pour ce faire, j'ai réalisé une classification hiérarchique des complexes associés pour mettre en évidence des correspondances privilégiées entre les profils des sites de liaison et ceux de leurs ligands.

L'urokinase joue un rôle essentiel dans les processus de l'inflammation dans divers états physiologiques (par exemple, la cicatrisation des plaies, l'élimination de l'endomètre), où elle contrôle l'activation et l'inhibition de la voie. L'expression et l'inhibition dérégulées sont liées à de multiples états pathologiques, par exemple des cancers invasifs ou des troubles inflammatoires comme l'arthrite rhumatoïde (Buckley et al. 2018). Cette protéine est connue pour être exprimée dans de nombreux organes, son niveau d'expression est le plus élevé dans l'épithélium des bronches, mais aussi dans le tissu synovial des articulations. Elle est l'unique principe actif du médicament Actosolv®, indiqué dans le traitement des complications thrombotiques des cathéters veineux.

La voie de signalisation de l'uPA est résumée dans la figure 3 et ne sera pas détaillée ici.

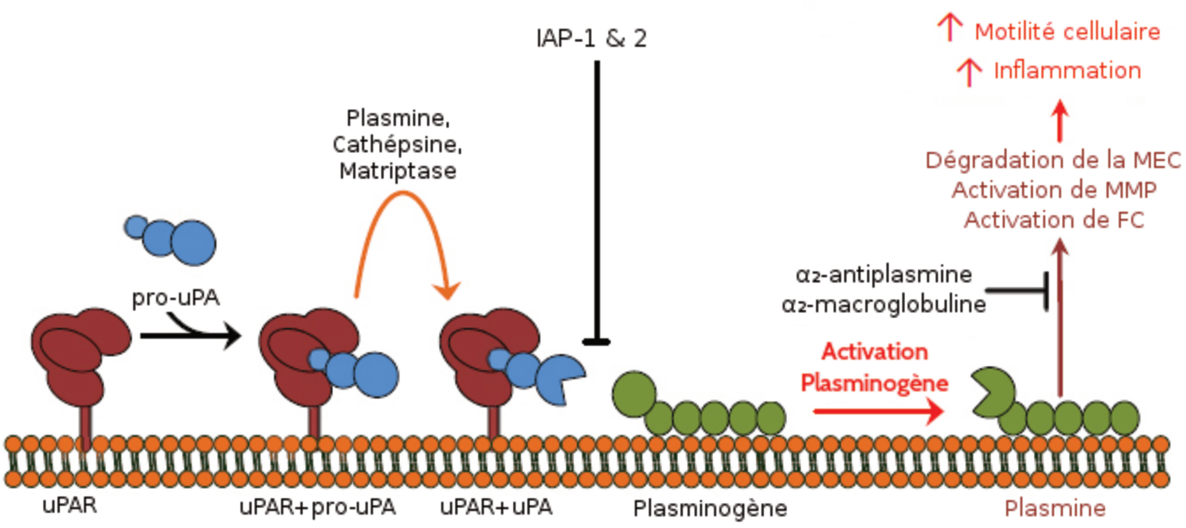


Figure 3: Voie de signalisation de l'activateur humain du plasminogène de type urokinase (uPA). Cette figure est tirée de l'article de Buckley et al 2019 et traduite. IAP : Inhibiteur de l'Activateur du Plasminogène ; MEC : Matrice extra-cellulaire ; MMP : Métalloprotéase matricielle ; FC : Facteurs de croissance.

La partie de la voie de signalisation qui nous intéresse est celle durant laquelle interviennent les inhibiteurs de uPA (IAP-A et IAP-2). Ils forment un complexe covalent avec uPA/uPAR provoquant l'endocytose (et donc la dégradation) de la totalité du complexe, due aux récepteurs spécifiques de l'endocytose. Cette inhibition permet donc la réduction du signal d'inflammation (Croucher et al. 2008) Il est connu que l'inhibition de l'uPA par IAP-1 induit des interactions secondaires de haute affinité avec les membres de la famille des récepteurs de l'endocytose, avec des effets d'activation subséquents sur la migration et la prolifération des cellules, pouvant être la cause de métastases cancéreuses (Cochran et al. 2011).

Cette cible étant d'intérêt thérapeutique, elle est très étudiée. De nombreuses structures tridimensionnelles complexées sont donc disponibles dans les banques de données structurales.

2.1.1 Les données disponibles

Pour cette étude effectuée en début de thèse, 101 structures complexées du domaine catalytique *trypsine-like* de la protéine uPA humaine ont été utilisées dont la plus précise est recensée dans le fichier PDB 4JNI (1,17 Å).

En juillet 2019 (deux ans après l'étude), 129 structures tridimensionnelles de l'uPA humaine sont disponibles, montrant ainsi que le nombre de données structurales ne cesse d'augmenter. La structure du fichier 4JNI est devenue obsolète, elle a été remplacée par la structure 5YC6 de résolution équivalente (1,18 Å).

Les structures de ce domaine de l'urokinase sont composées de 264 acides aminés dont certains sont manquants dans la plupart des structures tridimensionnelles cristallisées (environ quinze). C'est une protéine sphérique de 25 Å de diamètre, repliée en deux sous-domaines constitués de feuillets β antiparallèles. Les sous-domaines sont liés par un pseudo-axe double (Spraggon et al. 1995). D'après les informations extraites de la base de données UniProt (The UniProt Consortium 2019), l'uPA a pour rôle de cliver spécifiquement les liaisons entre l'arginine et la valine dans le plasminogène afin de former la plasmine.

Parmi les résidus de cette protéine, 4 mutations naturelles sont connues aux positions 15 (valine en leucine), 141 (proline en leucine), 214 (isoleucine en méthionine) et 231 (lysine en glutamine). Il y a aussi 4 acides aminés qui ont été modifiés pour les besoins cristallographiques, les résidus 150 (aspartate en glycine), 151 (cystéine en tryptophane), 386 (glycine en cystéine) et 430 (alanine en valine). Des mutations ont été tentées expérimentalement et deux sont rapportées pour avoir des effets sur son action. Les résidus en positions 158 et 323 (deux sérines) sont mutés expérimentalement en glutamate. Cela a pour effet de supprimer la phosphorylation, la fonction pro-adhésive et la capacité d'induire la réponse chimiotactique attendue, lorsque les deux résidus sont mutés (Franco et al. 1997).

Les données de structures tridimensionnelles disponibles pour cette étude sont nombreuses, variées et de bonne qualité. Les 75 ligands en interaction sont aussi nombreux et variés puisqu'on retrouve des petites molécules et des grands peptides. Le nombre important de ligands divers pour un seul site de liaison en fait une protéine promiscuous.

La méthode protéochémométrique développée sera donc adaptée à des protéines variées (voire mutées) liant de nombreux ligands et la prédiction des interactions sera adaptée à des molécules dont les propriétés sont proches de celles des candidats médicaments.

2.1.2 Le site de liaison

Nous avons appliqué plusieurs méthodes d'estimation du site de liaison sur l'urokinase, dont deux de ces méthodes sont proposées par PockDrug.

Une de ces méthodes est expliquée en détail dans l'article publié (section 2.2.), il s'agit de l'estimation par proximité au ligand (Borrel et al. 2015). Elle nécessite le choix d'un seuil de distance en dessous duquel les résidus sont considérés comme appartenant à la poche de liaison au ligand.

Deux autres méthodes ont été utilisées parallèlement à l'étude (résultats non publiés) pour estimer les sites de liaison de cette protéine. La première méthode consiste à rassembler tous les ligands disponibles en un ligand consensus afin d'appliquer la méthode par proximité à ce « super-ligand » (Triki, Billot, et al. 2018). Ainsi les poches extraites de protéines identiques dans des conditions biologiques identiques, mais complexées avec des ligands différents seront similaires. Cette méthode a pour avantage de prendre en compte la diversité et la flexibilité des ligands indépendamment de la déformation de la protéine due à chaque ligand spécifique. Elle peut être aussi utilisée pour des protéines non liées par superposition aux structures liées. Cette méthode nous a permis de définir le site de liaison le plus complet possible des urokinases sans dépendre d'un seul ligand. L'analyse montre qu'il est plus volumineux que ceux déterminés par proximité mais les propriétés physico-chimiques ne sont pas modifiées pour autant. Le super-ligand, quant à lui, n'a pas été étudié plus en détail car il ne correspond pas à une molécule biologique existante.

L'outil Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) a aussi été appliqué sur les 101 structures des urokinases. Comme décrit dans la section 1.4.1, cet outil permet d'estimer les cavités de la protéine qui peuvent probablement accueillir un ligand. L'avantage de cette méthode de détection de cavité est qu'elle ne nécessite pas la présence d'un ligand dans la structure cristallisée mais est basée uniquement sur la géométrie de la protéine. Cette méthode est moins précise mais a pour avantage de pouvoir être utilisée afin de caractériser les poches de liaison d'une structure sans ligand qui peut par exemple être issue de modélisation par homologie ou de dynamique moléculaire. La comparaison des poches de l'urokinase estimées par proximité et par Fpocket montre une superposition tridimensionnelle suffisante (score de superposition $> 0{,}25$) entre les deux types de poches. La caractérisation des poches des urokinases pourrait donc être enrichie avec les caractéristiques des poches des structures tridimensionnelles non liées.

La structure tridimensionnelle ainsi que certains résultats d'estimation de sites de liaison sont représentés en figure 4A. Le site de liaison estimé par la méthode de ligand consensus est en vert sur la figure 4B. Les différentes poches estimées par proximité et les ligands associés sont détaillés en figure 4C et 4D et sont de taille variable. Pourtant, ils se lient tous de manière

spécifique (K_i de 35 nM et 220 nM respectivement pour les ligands) au même site de liaison (en vert) des uPA.

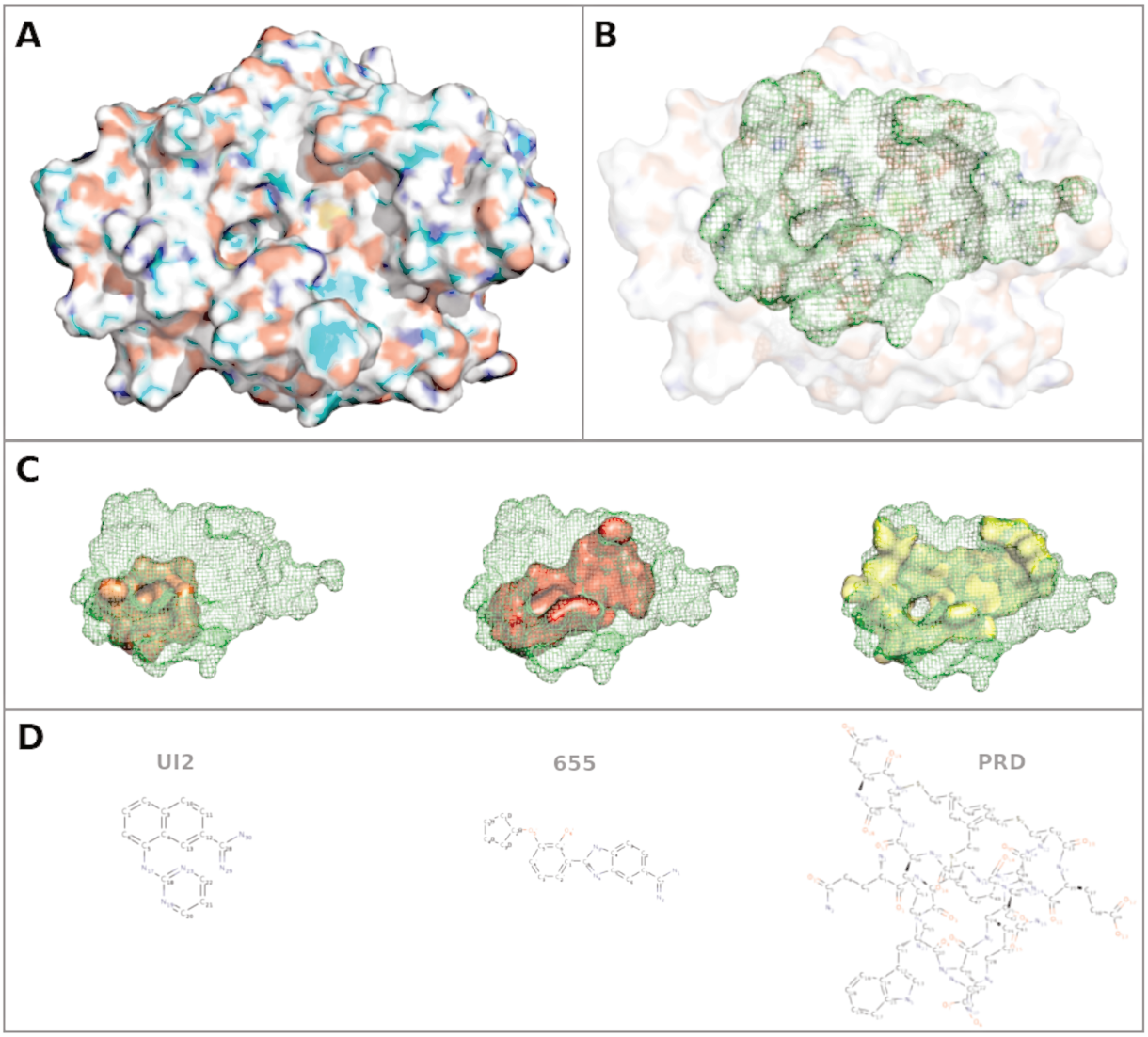


Figure 4 : Représentation de la structure 3D de l'urokinase (PDB 1GJ7) ; A) surface de la protéine colorée selon les propriétés physicochimiques des résidus (carbones en cyan, hydrogènes en blanc, azote en bleu, oxygène en rouge et soufre en jaune) ; B) Le site de liaison de l'uPA représenté en vert ; C) Trois poches extraites, estimées par proximité au ligand en utilisant respectivement les structures PDB suivantes (de gauche à droite) : 1SQO, 1O3P, 4MNW ; D) Les trois ligands correspondants (de gauche à droite) : UI2, 655 et le peptide PRD. Les deux premiers ligands sont drug-like et le dernier est un peptide. La représentation est faite grâce au logiciel PyMOL.

Afin de mieux comprendre la correspondance entre les poches et les ligands de cette protéine, ces derniers ont été caractérisés à l'aide de méthodes décrites dans l'article (section 2.2). Leurs propriétés physicochimiques et géométriques ont été extraites et analysées avec des outils statistiques.

2.1.3 Corrélation entre descripteurs de poche et descripteurs de ligand.

En parallèle de cette étude, une partie du jeu de données des urokinases a été utilisé afin de montrer la corrélation entre certaines caractéristiques des poches et des ligands. Une méthode d'estimation de poche simple et sans seuil arbitraire a été utilisée, elle ne nécessite

que la présence d'un ligand dans la structure. La poche est estimée comme étant les atomes des résidus voisins les plus proches des atomes du ligand.

Pour chaque structure, le rayon de la plus petite sphère englobant la poche et celui de la plus petite sphère englobant le ligand ont été calculés. Il a été montré une corrélation attendue entre ces deux rayons, montrant que cette méthode d'estimation des poches à partir du ligand était convenable, ce qui est une étape indispensable pour la prédiction des partenaires d'interaction. Ensuite, la convexité des poches a été étudiée, plus particulièrement un descripteur, *Distributional Sphericity Coefficient* (DISC), qui indique dans quelle mesure les atomes de la poche sont situés à la surface d'une sphère. Pour les urokinases, il a été montré que la forme des poches se rapproche modérément de celle d'une sphère et que les atomes lourds sont plutôt localisés à la surface de cette sphère. Ce descripteur permet donc de caractériser la sphéricité de la poche et principalement son adaptation à la taille du ligand. Il permet donc de caractériser conjointement la poche et le ligand.

La création de ce descripteur original (disponible à l'adresse http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html) vient ainsi augmenter les possibilités de caractériser les poches et de différencier celles qui acceptent de multiples ligands, des autres. Elle a donné lieu à une publication (cf. article n°1 en annexe).

2.1.4 Les clusters de complexes

La PDB étant le référentiel mondial pour les structures macromoléculaires tridimensionnelles, elle contient un ensemble de données extrêmement redondantes en termes de séquences et de structures. Il est donc fort probable que les complexes d'interactions poches-ligand qui en sont extraits se ressemblent en termes de propriétés physicochimiques et géométriques. Cette redondance nous permet d'étudier la promiscuité puisqu'une structure ne peut capturer qu'une seule interaction dans le même site de liaison, excepté dans le cas où deux ligands (coligands) se lient entre eux pour interagir avec une poche (Tonddast-Navaei, Srinivasan, and Skolnick 2017). Afin de faire ressortir les principaux types d'interaction, les complexes similaires ont été regroupés en utilisant la méthode de classification hiérarchique.

C'est une méthode de classification automatique qui, à partir d'un ensemble d'individus (ici les complexes poches et ligands), vise à regrouper les individus similaires. Pour ce faire, nous avons sélectionné l'espace optimal des complexes, en sélectionnant les composantes principales représentant 95 % de variabilité globale des données (descripteurs des poches et descripteurs des ligands) (Pérot et al. 2013). La distance euclidienne sur l'espace optimal entre tous les complexes a ensuite été calculée, elle sera utilisée comme mesure de dissimilarité entre les complexes. Une méthode d'agrégation intervient aussi dans le regroupement des classes, comme *ward*, *complete*, *single* ou *average* (Ward 1963). Dans cette étude, nous avons utilisé la méthode d'agrégation *ward*, car elle est robuste et donc très utilisée (Kimes et al. 2017). Elle consiste à regrouper les classes en minimisant l'inertie intraclasse (donc en maximisant l'inertie interclasse), associée au centre de gravité de chaque classe. L'inertie interclasse est la moyenne des carrés des distances des centres de gravité des classes, vis-à-vis du centre de gravité total.

Cette méthode ne détermine pas le nombre de classes le plus judicieux, c'est à l'utilisateur de le définir en fonction des données.

Dans cette étude, nous avons déterminé le nombre de classes le plus judicieux à 5, car cela permet la séparation des poches liées aux peptides des autres poches, et la séparation d'un groupe de poches ayant la même mutation (cf. article section 2.2).

2.2 Correspondances entre les poches et les ligands

La caractérisation conjointe des poches et des ligands de 75 complexes nous a permis d'obtenir une description précise des 5 principales interactions observées entre le site de liaison et les inhibiteurs de l'urokinase, et d'établir à partir de ces classes d'interactions les correspondances principales entre les caractéristiques des ligands liés et le site de liaison. Elle permet ainsi de détecter les changements conformationnels qui peuvent être dus aux mutations ou à la liaison (ou non) avec un ligand. Cette approche protéochémométrique est la première étape de l'étude des interactions multipartenaires pour une cible promiscuous. Cette caractérisation conjointe nous a permis d'établir des profils types d'interactions existant pour les urokinases, en tirant parti du grand nombre de données disponibles, qui justifie d'étendre cette double caractérisation dans un but de prédiction des partenaires des interactions. La haute redondance observée dans les structures tridimensionnelles est prometteuse pour étudier le site de liaison promiscuous.

Article n°2 :

Cerisier N*, Regad L*, Triki D, Petitjean M, Flatters D, Camproux AC. Statistical Profiling of One Promiscuous Protein Binding Site: Illustrated by Urokinase Catalytic Domain. *Mol Inform.* 2017 Jul 11. doi: 10.1002/minf.201700040. (*co-premiers auteurs)

Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201700040

Statistical Profiling of One Promiscuous Protein Binding Site: Illustrated by Urokinase Catalytic Domain

Natacha Cerisier[+],[a, b] Leslie Regad[+],[a, b] Dhoha Triki,[a, b] Michel Petitjean,[a, b] Delphine Flatters,[a, b] and Anne-Claude Camproux*[a, b]

Abstract: While recent literature focuses on drug promiscuity, the characterization of promiscuous binding sites (ability to bind several ligands) remains to be explored. Here, we present a proteochemometric modeling approach to analyze diverse ligands and corresponding multiple binding sub-pockets associated with one promiscuous binding site to characterize protein-ligand recognition. We analyze both geometrical and physicochemical profile correspondences. This approach was applied to examine the well-studied druggable urokinase catalytic domain inhibitor binding site, which results in a large number of complex structures bound to various ligands. This approach emphasizes the importance of jointly characterizing pocket and ligand spaces to explore the impact of ligand diversity on sub-pocket properties and to establish their main profile correspondences. This work supports an interest in mining available 3D holo structures associated with a promiscuous binding site to explore its main protein-ligand recognition tendency.

Keywords: Urokinases · proteochemometric modeling · statistical profiling · binding site · protein-ligand recognition

1 Introduction

For a decade, there have been fewer compounds entering clinical phases of study and fewer new compounds submitted to the authorities. However, the number of drugs being developed has not decreased accordingly, indicating a longer time in development. To reduce this time, the companies must identify new drugs with greater efficiency and specificity, using for example computer-aided drug design[1] and target-based drug discovery.[2] Indeed, with the increased speed of modern macromolecular structure determination techniques, structural information is more and more commonly included in "in silico" drug discovery pipelines.

Recent proteochemometric approaches[3,4] were developed relying on the description of multiple ligands along with multiple targets to quantitatively analyze their relations. The advantage of proteochemometric modeling is that it integrates information on both the ligand and the three-dimensional (3D) target with the interaction information simultaneously.[5,6,7,8] Thus, it can capture some information on protein–ligand interactions, such as different binding modes, different binding sites and interaction features between ligand and binding sites. Previously, we demonstrated the interest of proteochemometric protocol in analyzing double binding site and drug spaces and in establishing binding site and drug profile correspondences enlarging the notion of protein family.[9] To analyze the correspondences between the target and the drug profiles using proteochemometric approach, it is necessary to precisely characterize the ligand and the target binding site spaces. For the chemical space, 90% of orally active drugs that achieved Phase II status corresponding to drug-like molecules.[10] These drug-like molecules must follow Lipinski's Rule-of-Five (RO5).[11] Since the early years of Quantitative structure–activity relationship (QSAR), numerous chemoinformatic methods have been developed to describe the structural features of compounds such as small molecules and peptides molecules.[5,12,13,14,15,16,17,18] Considering 3D target space, the number of 3D structures in the Protein Data Bank (PDB) have grown to more than 130 000 protein structures,[19] with high redundancy, i.e., more than half share at least 95% sequence identity. Precisely characterizing a binding site determines its potential in drug design projects. This requires a crucial preliminary step of performing pocket estimation and using relevant pocket descriptors. This pocket estimation step is complicated because of the difficulty in defining the pocket boundaries.[20,21,22] As there is no consensus among the various estimation methods, developing these methods is always in progress.[23,24] Likewise, descriptors for optimal geometrical

[a] *N. Cerisier,[+] L. Regad,[+] D. Triki, M. Petitjean, D. Flatters, A.-C. Camproux*
INSERM, UMRS-973, MTi,35, rue Hélène Brion, 75205 PARIS CEDEX 13
phone/fax: +331 57 27 83 86 / +331 57 27 83 72
E-mail: anne-claude.camproux@univ-paris-diderot.fr
[b] *N. Cerisier,[+] L. Regad,[+] D. Triki, M. Petitjean, D. Flatters, A.-C. Camproux*
University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi
[+] *The two first authors equally contributed.*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.201700040

and physicochemical pocket characterization are ongoing.[18,25,26,27,28,29] Moreover, during earlier stages of drug discovery, an accurate evaluation of the target druggability is an important characteristic for studying the affinities of targets for drug-like molecules and for reducing the high failure rate in drug discovery research.[30] This property corresponds to the target's potential in accommodating small-molecule drugs thus modulating its biological function. Different learning methods based on different pocket estimations and descriptors have been developed to predict pocket druggability.[31] Recently, a high-performance druggability prediction method PockDrug was developed to be robust with respect to pocket boundary and estimation uncertainties.[32,33]

The old drug design paradigm, i.e. drugs selectively interact with one protein, resulting in treatment and prevention of disease has been challenged by several studies.[34,35,36] Nowadays, it is well known that a drug may be involved in different disease functions and interact with more than one target, which defines drug promiscuity.[37] Many studies focus on the promiscuous drugs[38] but few focus on promiscuous targets (i.e. targets that bind several compounds).[21,39]

Moreover, druggable binding sites are known to correspond to rather large regions,[32,40] consequently they can potentially include multiple sub-pockets and be promiscuous, with available complex forms associated with drug-like or non-drug-like ligands. Thus the druggability prediction of ligandable targets, bound with non-drug-like ligands[31] is of interest. In this work, we propose a proteochemometric method to explore the promiscuity of a druggable binding site. In this study, we develop a proteochemometric method to explore the promiscuity of a druggable binding site and to identify its main pocket-ligand profile correspondences by analyzing associated redundant target PDB structures. The considered target is the human urokinase Plasminogen Activator (uPA) and we focus on its druggable catalytic inhibitor-binding site (uPA binding site). This protein was chosen to illustrate this approach because of its role in cancer pathways[41] and is well studied, with numerous available PDB structures. We proposed a protocol to identify and characterize pocket and ligand profile correspondences. It consists of four steps: (i) characterizing diverse ligands associated with this uPA binding site and their corresponding estimated sub-pockets estimated by ligand proximity, (ii) optimally selecting complementary pocket-ligand descriptors for a precise ligand-pocket correspondence characterization, (iii) clustering the pocket-ligand pairs using these descriptors and identifying the main pocket-ligand clusters and, (iv) analyzing correspondences between pocket and ligand profiles. Using this application, we showed that this approach is relevant to exploring the impact of the diversity of ligands associated with one binding site on its corresponding binding sub-pockets and to establish main profile correspondences between them. The use of this protocol with accurate pocket and ligand characterization and the consideration of promiscuous targets could improve our understanding of pocket and ligand recognition.

2 Materials and Multivariate Methods

2.1 uPA Catalytic Domain Target Complex Data

In this study, we analyzed the uPA, as its different roles make it a therapeutic target of interest. Indeed, this enzyme has a proteolytic function, converting plasminogen to plasmin.[42] This conversion triggers a proteolytic cascade and plays a role in both thrombolysis and extracellular matrix degradation.[43] This enzyme also promotes the intracellular signaling by its interaction with transmembrane proteins such as integrin and participates in cellular adhesion mediation, differentiation, proliferation and migration.[44]

The protein uPA is 411-residues long and consists of three domains: the trypsin-like catalytic, kringle and growth factor domains. Here, we focused on the trypsin-like catalytic domain because one way to inhibit this target is to develop a chemical molecule that binds this target in competition with the substrate, i.e., at the inhibitor binding site, called the uPA binding site. We extracted 101 X-ray structures of the uPA trypsin-like catalytic domain complexed with ligands from the PDB.[16] These X-ray structures have quality resolutions ranging from 1.17 Å to 3.10 Å. Most of these 3D structures have up to three of the following mutations: N147, N145, D102, N145, and S190, made to facilitate crystallography.[45,46] All these mutations are located outside the uPA binding site, as shown in Figure 1. For this study, we selected 3D uPA complex structures with complete coordinates and removed redundant (identical SMILE code) or mixture ligands by passing the FAF-Drugs filter.[47,48] This results in the creation of the "uPA set".

2.2 Ligand Space Characterization and Diversity

The uPA-ligand set is composed of co-crystallized ligands: small molecules and peptides, bound to the inhibitor-binding site of the complex structures from the uPA set. The diversity of these ligands was quantified by computing Tanimoto coefficients for all ligand pairs based on MACCS FingerPrints (MACCS Drug Data Report, Release 2000.2, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, 2000), using the OpenBabel software.[49] A Tanimoto coefficient of 1 corresponds to identical MACCS FingerPrints whereas a coefficient of 0 does not share any FingerPrints similarities.

Each ligand of the uPA-ligand set was described using 44 geometrical and physicochemical descriptors: 14 geometrical descriptors used in[50,51] and 30 physicochemical ligand descriptors proposed by the FAF-Drugs3 software

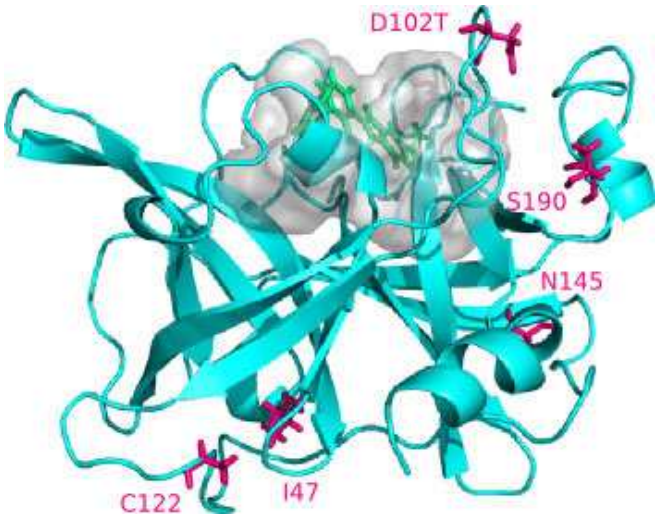


Figure 1. Representation of human uPA catalytic domain (PDB 1GJ7) with PyMOL.[69] The pocket is represented as a surface (in gray), the ligand as a stick model (in green) and the rest of the protein in cartoon form (in cyan). Five possible mutations are represented as a stick model (in magenta) on the structure. They are all located outside the pocket.

[Free ADME-Tox Filtering Tool.[47,48] The Wilcoxon test[52] was performed to compare descriptor values between peptides and small molecules, using the "wilcox.test" function and the R stats package version 3.3.1.[53]

Drug-like molecules correspond to orally bio-available small molecules that have an optimal physicochemical property profile in terms of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME-Tox), as defined by Lipinski in 1997[54] and recently reviewed by Ritchie and Macdonald in 2014.[55] In the literature, there are many rules for detecting "drug-likeness".[54,56] We used the "drug-like soft" filter of FAF-Drugs3 based on Lipinski's rules corresponding for instance to Rings ≤ 6, RotatableB ≤ 11 and RatioH.C included in [0.1, 1.11][47] to identify the drug-like molecules.

2.3 Pocket Set Estimation and Characterization

The uPA binding site can be estimated by extracting the binding pocket of each complex structure from the uPA set. Choosing the most appropriate pocket estimation method is complex[21] because the pocket boundaries are difficult to determine and depend on certain arbitrary selections.[20,31] In this study, the uPA binding pockets were estimated using proximity, an approach guided by ligand proximity.[32] A pocket is defined as the set of atoms that are situated at a threshold distance from the co-crystallized ligands commonly between 4 Å and 6 Å.[57,58,59,60] A threshold of 5.5 Å was used in order to capture the specific part of the binding site (sub-pocket) in contact with the binding ligand and a relatively complete environment of the binding site,[32,33] using the PockDrug-Server pocket estimation (http://pockdrug.rpbs.univ-paris-diderot.fr/cgi-bin/index.py?page=home). All ions such as SO4, K, CL, NA appear as ligands in the PDB files and are excluded from the pocket estimation procedure. The sub-pockets estimated by proximity compose the uPA-pocket set.

Each pocket of the uPA-pocket set was described using 70 various descriptors: 54 are physicochemical and 16 are geometrical, including those from Petitjean.[50] The druggability of each pocket was predicted using the PockDrug web server, which was developed to overcome pocket estimation uncertainties, highly performing in terms of average accuracy (87.9 % ± 4.7 %) on a test set using several pocket estimation methods.[33]

2.4 Proteochemometric Method

The uPA set is split into the uPA-ligand and uPA-pocket sets, which corresponds to the uPA pocket-ligand pair set (also called the uPA complex set). We previously developed a proteochemometric protocol on a large dataset,[9] considering both ligand and pocket descriptors, to allow better exploration of the pocket-ligand pair space instead of using the pocket space or the ligand space alone. In this work, we adapted this protocol[9] to study the impact of ligand diversity on the multiple sub-pockets of one unique binding site, to establish some (physicochemical and geometrical) descriptor profile correspondences between the bound ligands and their associated sub-pockets and to characterize the main observed pocket-ligand recognition observed.

2.4.1 Pocket and Ligand Descriptor Selection and Pocket-Ligand Pair Set Characterization

First, we removed most redundant information and selected the most highly correlated pocket and ligand properties of the uPA pocket-ligand pair set using the following protocol. Pearson correlation coefficients were computed to group redundant ligand descriptors and redundant pocket descriptors separately. Then, we chose one representative descriptor per ligand group that had the highest Pearson correlation with pocket descriptor groups. The same step was performed for the pocket descriptor groups. The aim was to remove redundant information and to choose a representative from each ligand descriptor group and a representative from each pocket descriptor groups that optimized the correlation between the pocket and ligand descriptors. These pocket and ligand representative descriptors optimally characterize the uPA pocket-ligand pairs set.

Based on these representative pocket and ligand descriptors, a Principal Component Analysis (PCA) projected the uPA pocket-ligand pair set into a subspace made of principal components (orthogonal linear combinations of these representative descriptors). The FactorMineR pack-

age[61] from software R[53] was used to carry out the PCA. Following the developed protocol[9] to reduce noise, we retained the first K PCA principal components which capture 95% of data variability. Then a Euclidean similarity distance was calculated between all uPA pocket-ligand pairs on these first K PCA components.

2.4.2 Building and Characterization of Complex Clusters

The next step consisted of clustering complexes of the uPA set and identifying the main sub-pocket and ligand profile correspondences to characterize the observed uPA protein-ligand recognition. Using the Euclidean distance between all members of the uPA pocket-ligand pairs set, a hierarchical classification was performed. The classification process builds embedded partitions by progressively gathering the complex pairs and then the pair clusters according to the Ward metric.[62] Hierarchical classification results in a tree representation, which visualizes the proximity of pocket-ligand pairs, both in terms of both pocket and ligand descriptors. This representation allows C main clusters of pocket-ligand pairs (complexes) to be identified by choosing a threshold. This threshold can be chosen to obtain a minimum number of complexes in each cluster or a minimum distance between two clusters. It can also be adjusted to examine the proximities of complexes at different similarity levels. Thus, this approach could be applied to fewer complexes but requires ligand and/or 3D structure diversity to be informative. The function used to cluster complexes is the "hclust" function from the software R.[53]

These C clusters correspond to C pocket-ligand pair correspondences. The average and standard deviation of each descriptor is illustrated in one star plot of each cluster. This plot illustrates the main profiles of the geometrical and physicochemical pocket and ligand descriptors of each complex cluster and the correspondence between the profiles of the pocket and ligand descriptors associated with each cluster.

Then, a comparison of pocket and ligand descriptor values between the C clusters was performed with a Bonferroni correction applied to take into account the multiple (C) comparisons of clusters. These comparisons determine which representative descriptors are significantly different between the C considered clusters.

3. Results and Discussion

3.1 uPA Set Characterization and Ligand-Pocket Pair Descriptor Selection

The uPA set corresponds to the 75 PDB complexes that match the human uPA catalytic domain, selected using method section 2.1. The corresponding uPA pocket-ligand pairs set is composed of 75 co-crystallized ligands: small molecules and peptides bounded to the uPA inhibitor-binding site and 75 sub-pockets estimated by proximity to the ligands.

The first step consisted of selecting representative descriptors among the 70 pocket and 44 ligand descriptors to optimally characterize the uPA pocket-ligand pairs set. This selection was performed by identifying the most highly correlated pocket-ligand pair properties involved in the pocket-ligand interactions but removing redundant pocket descriptors and ligand descriptors redundancy keeping a maximal correlation threshold of 95%, see Materials and Methods. We selected respectively nine ligand and nine pocket descriptors, described in Supporting Information (S.I.) Table S1, to maintain balance between the pocket and ligand partners of the interactions. Two geometrical descriptors were selected for pockets and ligands: the radius of the smallest enclosing sphere (RADIUS_HULL_P and RADIUS.HULL_L) for both pocket and ligand, the number of pocket residues (C_RES) and the index of sphericity of the ligand (PSI_lig). The other seven physicochemical ligand descriptors were molecular weight (MW), polarity descriptors (LogP, logD), the number of rotatable bonds (RotatableB), the size of the smallest ring (rings), solubility (logSw) and the ratio between non-carbon and carbon atoms (ratioH.C). The seven physicochemical pocket descriptors were hydrophobicity (hydrophobic_kyte and p_hydrophobic_atom); residue proportions in terms of charged, aromatic or positive residues; and atom proportions (main-chain, N-atom).

The average values and standard deviations of these 18 descriptors using the uPA pocket-ligand pairs set are given in Table 1 (first column). Figure 2 presents the correlations between nine pocket and nine ligand representative descriptors. We observed a high correlation between the pocket and ligand geometrical descriptors, (0.81, p-values $<$2.10-16 for RADIUS_HULL_P and RADIUS.HULL_L), which is supported by the fact that the uPA-pocket set is estimated by ligand proximity. We also note that some physicochemical descriptors are highly correlated between pocket and ligand descriptors such as p_main_chain_atom and RotatableB ($-$0.80, p-values $<$2.10-16). PCA performed on the 18 representative pocket and ligand descriptors enables a suitable representation of the uPA pocket-ligand pairs set, as shown in S.I. Figure S1A. As indicated by the projection of the 18 variables near the PCA correlation circle, different selected pocket and ligand descriptors contribute in a balanced way to the variability of the data and are relevant to the study, as shown in S.I. Figure S1B. Therefore, the representative descriptors successfully captured various pocket and ligand information and relevant correspondences between the ligand and pocket spaces. The first K$=$10 PCA principal components accounted for 95% of the global data variability, as shown in S.I. Figure S1C. They were used to quantify similarity between pocket-ligand pairs of the uPA set.

Table 1. Average values and standard deviations for uPA pockets and five groups of complexes according to the complex classification into five complex clusters (C1 to C5), see Figure 3A. The first ten rows correspond to 10 selected ligand descriptors, and the last ten rows are pocket descriptors. In both cases, the first eight are physicochemical descriptors and the last two are geometrical ones.

Descriptor	uPA set average	sd	C1 average	sd	C2 average	sd	C3 average	sd	C4 average	sd	C5 average	sd
MW	510.34	449.78	1550.62	132.32	242.33	73.25	382.82	85.80	332.18	49.81	499.80	71.27
logP	−0.26	3.96	−8.50	3.13	0.63	1.14	2.54	1.57	2.31	0.79	−0.88	1.43
logD	−0.66	4.75	−7.78	9.04	0.18	1.42	1.56	1.13	1.89	1.52	−1.53	1.30
logSw	−2.50	1.44	−2.49	2.00	−1.80	0.85	−3.68	1.35	−3.36	0.59	−1.64	1.34
RotatableB	7.81	8.18	24.73	5.00	2.41	1.50	4.93	2.37	3.85	1.28	13.00	1.83
Rings	1.99	0.85	2.18	0.60	1.37	0.49	2.79	0.89	2.31	0.48	1.90	0.99
ratioH.C	0.43	0.23	0.71	0.06	0.46	0.26	0.30	0.07	0.21	0.06	0.54	0.11
RADIUS.HULL_L	7.16	2.78	11.87	2.60	4.96	1.91	7.82	0.56	7.59	0.86	6.46	0.88
PSI_lig	0.21	0.12	0.30	0.06	0.13	0.09	0.20	0.06	0.16	0.03	0.42	0.05
hydrophobic_kyte	−0.52	0.21	−0.59	0.12	−0.58	0.15	−0.18	0.14	−0.65	0.13	−0.61	0.13
p_aromatic_residues	0.19	0.04	0.24	0.03	0.16	0.02	0.19	0.05	0.22	0.02	0.18	0.02
p_charged_residues	0.23	0.03	0.25	0.02	0.21	0.02	0.20	0.02	0.25	0.02	0.25	0.02
p_hydrophobic_atom	0.07	0.02	0.09	0.01	0.05	0.01	0.07	0.01	0.06	0.01	0.09	0.01
p_main_chain_atom	0.61	0.06	0.52	0.03	0.65	0.03	0.63	0.03	0.59	0.03	0.56	0.02
p_N_atom	0.17	0.03	0.13	0.01	0.19	0.01	0.17	0.01	0.16	0.01	0.16	0.01
p_positive_residues	0.14	0.03	0.15	0.02	0.13	0.02	0.11	0.02	0.17	0.02	0.15	0.02
C_RES	29.39	9.37	49.18	5.47	22.89	2.74	30.93	3.34	24.85	2.54	28.90	1.52
RADIUS_HULL_P	11.02	2.11	14.24	0.99	8.75	0.73	12.04	0.75	11.96	1.01	10.95	0.86

3.2 Ligand-Pocket Pair Characterization

3.2.1 uPA Ligand Space Diversity

The uPA-ligand set is composed of 64 non-peptide ligands and 11 peptide ligands.

First, the chemical diversity of the uPA-ligand set was assessed using the Tanimoto similarity coefficient, resulting in an average and standard deviation of 0.28 (±0.19). The peptide set is less diverse than the small molecule set (average Tanimoto coefficient of 0.56±0.20 versus 0.31±0.20). Only 4.52% of ligand pairs exhibits high similarity with Tanimoto coefficient >0.8 and the majority (85%) exhibit a Tanimoto coefficient less than 0.5 which demonstrates a high diversity of ligands associated with the uPA binding site.

Second, the average and variability of the uPA-ligand set were studied using the nine representative ligand descriptors for peptide and non-peptide ligands, (S.I. Table S2). We observed significant descriptor values for most of the descriptors (7/9) of peptides compared to small molecules, for instance a significantly higher MW value: 1550 (±132) versus 332 (±116), p-values < 10-6.

When we compared variability of the uPA non-peptide ligands with respectively marketed oral drugs [63] and non-peptide ligand values of the FAF-Drugs3 "Drug-like soft filter":[47,48] we obtained similar average values and variability of MW of 332 (±116) versus respectively 337 (±157) and values less than 500 Da. Other descriptor values obtained in the uPA-ligand set are consistent with the literature but exhibit high variability, which is supported by the presence of peptides within uPA-ligand set. For instance, the Rings, RotatableB and RatioH.C descriptors (1.95±0.88, 4.91±4.02, and 0.39±0.21) are coherent in average with the FAF-Drugs3 "Drug-like soft filter" values (see method) but present relatively high variability. Therefore, less than half of the uPA-ligand set ligands (41.33%, 31/75) and of its small molecules (48.44%, 31 of 64) are identified as drug-like molecules using FAF-Drugs3. All these results confirm the high diversity and the high geometrical and physicochemical variability of the uPA-ligand set that can bind the uPA binding site.

3.2.2 uPA Binding Site Characterization and Sub–Pocket Diversity

The uPA-pocket set is composed of the 75 sub-pockets estimated by proximity to the 75 ligands of the uPA-ligand set using the 75 complexed chains of the uPA set. It represents the multiple binding sub-pockets of the uPA binding site.

First, the average and variability of the uPA-pocket set were studied using nine representative pocket descriptors. To study the variability of these sub-pockets obtained by proximity to variable uPA-ligand set, drug-like or no, we compared their descriptors values to those obtained using on the extensive dataset of 113 "NonRedundant dataset of Druggable and Less Druggable binding sites" (NRDLD) of Krasowski et al.[40] The NRDLD set corresponds to diverse and non-redundant complexed proteins (sharing a pairwise sequence identity of less than 60%) and various binding sites, for instance druggable or not druggable, but bound only to small molecules. The uPA-pocket set exhibits

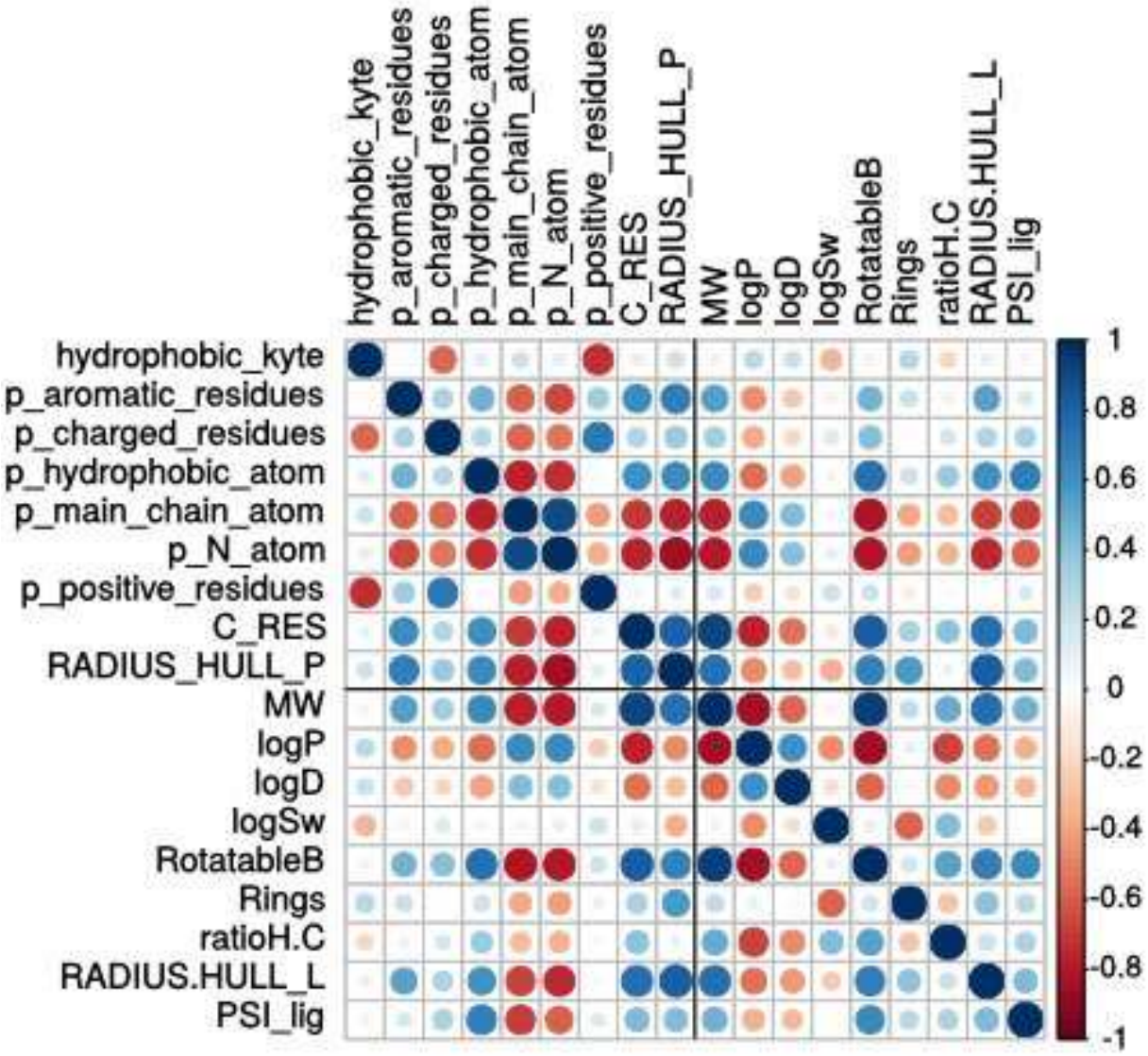


Figure 2. Correlation matrix of the selected 18 descriptors (nine ligands and nine pocket descriptors). Correlation varies between −1 and 1. Blue circles represent positive correlation; red circles represent negative correlation and white circles represent not significant correlation. The size of the circle is proportional to the absolute value of the correlation.

variable properties that cover a broad range of values, close to the ones obtained estimated by ligand proximity (using PockDrug with an identical threshold of 5.5)[33] on NRDLD. The geometrical descriptors: RADIUS_HULL_P and C_RES are close in average and variability for the uPA-pocket set versus NRDLD: 29.39 ± 9.37 versus 21.88 ± 7.26 and 11.02 ± 2.11 versus 10.19 ± 1.76. The uPA-pocket set also presents physicochemical descriptors values that are consistent with but less variable than NRDLD: for instance, for hydrophobicity_kyte and for the proportion of aromatic residues: −0.53 ± 0.21 versus −0.33 ± 1.11 and 0.19 ± 0.04 versus 0.19 ± 0.13, respectively. This weaker physicochemical descriptor variability of the uPA-pocket set related to NRDLD can be explained by the fact that corresponding sub-pockets are estimated from one druggable uPA-binding site. This suggests these multiple sub-pockets capture common physicochemical tendencies of the binding site (even if these descriptors are dependent on the frontiers and shape of each estimated sub-pocket).

Second, an important characteristic of the binding sites concerns their potential druggability,[31] i.e., their ability to bind drug-like molecules. The uPA-pocket set estimates the

known druggable uPA binding sites. Among them, 85% (64) are well predicted as druggable using the PockDrug web server[33] with an average druggability score of 0.67 (±0.13), from 0.42 to 0.87. Moreover, we note that eleven wrongly predicted sub-pockets exhibit druggability scores very close to the druggability threshold of 0.5, from 0.42 to 0.49 (0.46 ±0.02). Therefore, 85% of pockets in the uPA-pocket set are well predicted as druggable whereas only 48% of ligands in the uPA-ligand set are defined as not drug-like. This result indicates that druggability is correctly predicted for 85% of the sub-pockets estimated by proximity, even if 59% of the ligands are not drug-like, according to the binding site physicochemical properties captured using pocket estimation by ligand proximity. However, we can note that there is no direct correspondence between drug-like ligand properties and pocket druggability prediction as 36/44 (81.82%) pockets estimated for a non-drug-like ligand are predicted as druggable and 3/31 pockets estimated for a drug-like ligand are predicted as non-druggable.

Thus, the uPA-pocket set corresponds to geometrically variable sub-pockets of the uPA binding site, in agreement with the diversity of its observed co-crystallized ligands and the estimation guided by the ligand proximity. The uPA-pocket set exhibits relatively weak variability in terms of physicochemical properties. This weakness suggests that even if the estimated sub-pockets depend on their bound ligand, this estimation using a threshold of 5.5 Å is able to capture the global trend of the uPA binding site. This result can explain the good druggability prediction score of the uPA-pocket set. This result is encouraging, as it confirms that the druggability of a binding site can be predicted even when from complexes not bound to drug-like ligands.

These results confirm the pertinence of the proximity estimation method for estimating multiple sub-pockets of a unique binding site. It is able to capture the target druggability property and the specific part of the binding site (sub-pocket) in contact with the binding ligand. It also confirms that uPA sub-pockets are variable enough to explore the impact of ligand diversity on this sub-pocket variability. The great variability of uPA ligands and pockets are well described by representative pocket and ligand descriptors, which makes it possible to establish correspondences between ligand and pocket profiles and to focus on description of the pocket-ligand interaction region. The next step consists of identifying and characterizing the main correspondences between the profiles of ligands and pockets to capture information about pocket-ligand recognition for the uPA binding site.

3.3. uPA Complex Analysis

3.3.1 Pocket-Ligand Pair Cluster Characterization

To map and group together uPA complexes with similar pocket and ligand properties, we built a hierarchical clustering tree using their first K PCA principal components, (here K = 10 were sufficient to capture 95% of the data variability, see S.I. Figure S1C). Hierarchical clustering of uPA pocket-ligand pairs highlights that five main clusters are well differentiated, as shown in Figure 3A1 and 3A2. These five clusters, denoted C1-C5, correspond to 11, 27, 14, 13, and 10 complexes, respectively. Their corresponding profiles of nine ligand and nine pocket descriptors are described in Table 1 in terms of average and standard deviation. They are illustrated on five double star plots, as shown in Figure 3B1 and 3B2: one ligand and one pocket star plots for each considered cluster. It allows us to determine which ligand profile could bind to some sub-pockets and conversely identify the properties of ligand profiles which can bind to some sub-pocket profiles, among those observed in uPA set. Descriptor comparisons of the five clusters were performed using p-values from a Wilcoxon test and Bonferroni correction, as shown in Figure 4. Both the star plots and the Wilcoxon comparison (Figure 3B and Figure 4 respectively) illustrate the main correspondences between pocket and ligand descriptor profiles of one cluster and the significant difference from other clusters. The observed correspondence between pocket and ligand geometrical properties within each pocket-ligand pair cluster is expected because the proximity method estimates the pocket based on ligand proximity, as shown in S.I. Figure S2.

We observe that the five main pocket-ligand pair clusters are well differentiated as shown in Figure 3A1 with at least six significantly different descriptor values, p-value < 0.0125, as shown in Figure 4. The C1 cluster split distinguishes the 11 peptides exactly from the other non-peptide ligands of the uPA set, as shown in Table 1 and S.I. Table S1. As expected these C1 complexes not only exhibit distinct for all ligand descriptors (except logD and logSw), but also have distinct pocket properties, except for hydrophobic-kyte, p_positive_residues and p_charge_residues in some cases. Cluster C1 associated with peptides is characterized by significantly larger geometrical pocket and ligand values. The other four clusters are differentiated as follow: C2 corresponds to rather small pockets and ligands, while C3 and C4 exhibit different pocket profiles: 6/9 significantly different descriptors. The C5 cluster corresponds to the second largest ligand with the more spherical hull that are not peptides. The closest pocket profiles are associated with at least five different pocket descriptors (clusters C4 and C5) and correspond to two different profiles of bound ligands. The most similar ligand profiles (only ratioH.C differs significantly) are associated with clusters C3 and C4 which correspond to two well-differentiated pocket profiles (at least five different descriptors, notably p_charged_residues and p_positive_residues and hydrophobic-kyte), explaining the existence of the two clusters.

Thus, this joint analysis confirms that the uPA ligand diversity corresponds to five significantly different profiles of sub-pockets. It establishes that uPA promiscuous binding site corresponds to five main pocket-ligand recognition

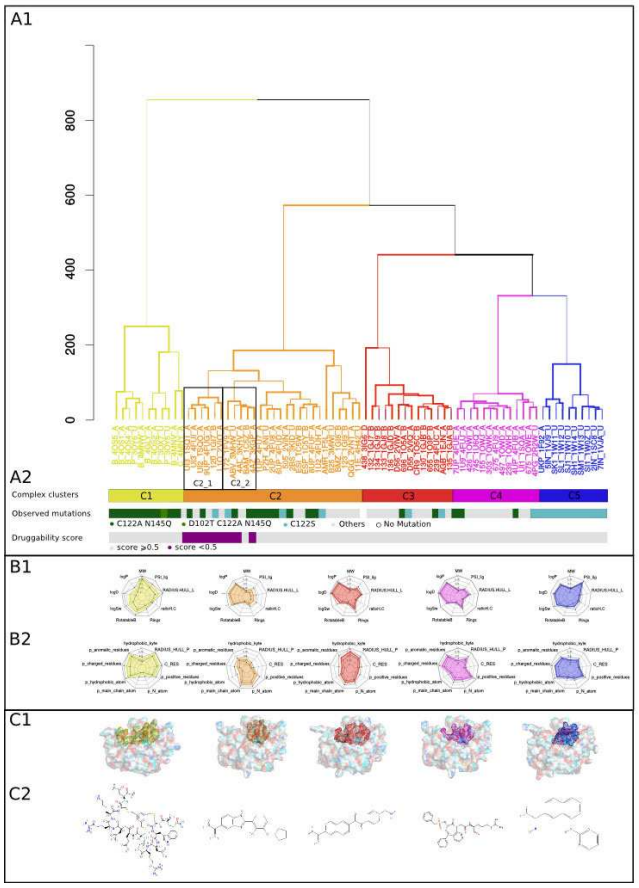


Figure 3. Illustration of pocket-ligand pair classification. A1) Classification trees of pocket-ligand pairs clustering into five clusters in rainbow colors. Sub-clusters C2_1 and C2_2 are framed in black boxes. A2) Three color bars show the five main pocket-ligand pairs classification (noted C1 to C5), observed mutations and druggability score, as predicted using PockDrug.[31,30] The first bar is colored in the same rainbow colors as the classification. The second bar is colored according to considered mutations: C122A–N145Q in dark green, D102T–C122A–N145Q in light green, C122S in blue, the other in grey and no mutation in white (PDB 3IG6). The third bar is colored according to the pocket druggability score: ≥ 0.5 in grey and < 0.5 in purple. B1) Star plots representing the nine ligand descriptors of pocket-ligand pairs colored according to the classification. B2) Star plots representing the nine pocket descriptors of pocket-ligand pairs colored according to the classification. For all star plots, the solid surface is delimited by the normalized average value of descriptors and the black lines correspond to the normalized standard deviation of the considered descriptors. They are colored in yellow, orange, red, magenta and blue, according to the classification. C1) Representation of five uPA structures representative of the the five cluster: PDB 1SQO, 1O3P, 4MNW, 1OWH, 1W12, respectively colored in yellow, orange, red, magenta and blue, according to the classification. PDB are represented as surfaces with PyMOL[69] and their respective pockets estimated by proximity represented in mesh. The protein surface is colored according to the atoms in the residues: carbon in cyan, hydrogen in gray, nitrogen in blue, oxygen in red, sulfur in orange. C2) The five ligands corresponding to the complexed structures are represented in 2D including one peptide (UI2, 655, peptide chain B, 239, SM1).

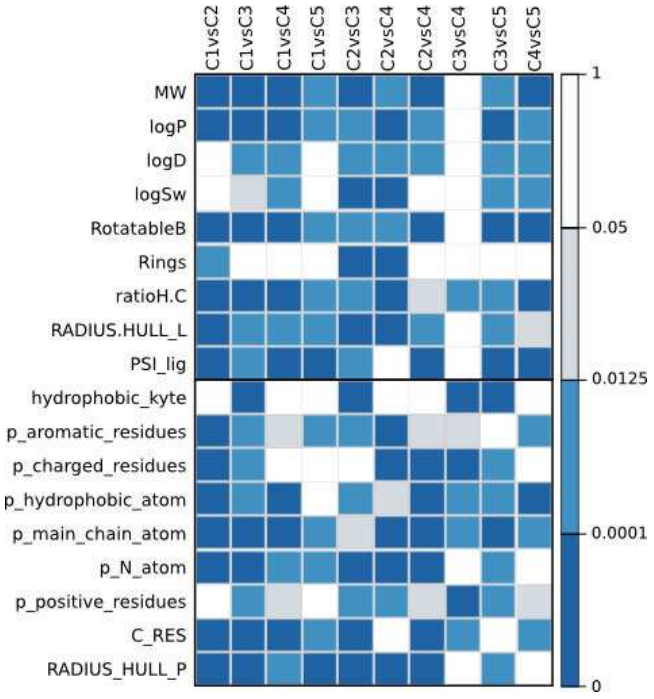


Figure 4. Representation of p-value matrix for average comparisons between all pairs of five complex clusters for 18 descriptors. Due to the multiple tests between the five clusters, we apply a Bonferroni correction, which gives a threshold of significance of 0.0125 for an alpha risk of 5%. Significant differences are in blue, and the darker the box is, the more significant the p-value is, according to the rating scale on the right. White and gray boxes show no significant differences. The first nine descriptors are ligand descriptors and the last nine are pocket descriptors. Column names are compared pair of clusters.

tendencies, related to the available PDB. Figure 3C1 and 3C2 illustrate respectively five sub-pockets and five corresponding ligands from five representative complexes of these five clusters.

3.3.2 Complex Cluster Analysis and Interpretation

By selected appropriate pocket and ligand descriptors, this joint clustering approach provides a useful classification of the uPA complex in five main clusters of pocket-ligand pairs and highlights the main uPA correspondences between ligand and sub-pocket profiles. Thus, the analysis of this complex classification can capture information in terms of pocket-ligand profile correspondence associated with one binding site.

In the case of the uPA set, when examining pocket druggability score information in the complex classification, we observe that 11 sub-pockets wrongly predicted in terms of PockDrug druggability score (see method) are grouped into the C2 cluster corresponding to smaller ligands and pockets. More in details, they are divided into two sub-

clusters C2_1 of four complexes and C2_2 of six complexes, as shown in Figure 3A2. Thus, non-parametric comparison of 18 pocket-ligand descriptors were computed (using non parametric Wilcoxon test) between sub-cluster C2_1 and C2 and sub-cluster C2_2 versus C2, as shown in S.I. Figure S2. This Figure illustrates two particular complexes profiles with pockets wrongly predicted as non-druggable. For instance sub-cluster C2_2 corresponds to significantly smaller ligands and pockets (in terms of RADIUS.HULL_L), with particular physicochemical properties for ligands (significantly lower values of MW, rotatableB) and sub-pockets (significantly lower values of p_aromatic_residues, higher values of p_main_chain_atom, p_N-atom) related to C2 complexes. These results are coherent with their low druggability prediction scores as druggable pockets are expected to be rather large and aromatic.[33] Thus, by finer browsing of the complex classification obtained, this approach directly identifies two profiles of pocket-ligand pairs associated with pockets wrongly predicted in terms of druggability, which can be characterized in terms of pocket-ligand profiles.

This example illustrates analyze of complex similarity can be informative, even in case of a few numbers of pocket-ligand pairs. However, complex classification can only be informative in terms of pocket-ligand profile correspondence and main pocket-ligand recognition, in case of available 3D structure variability or/and ligand diversity.

In case of variable data, the main pocket-ligand recognition clusters can be analyzed, a posteriori to understand relationship between 3D structural or ligand changes and main pocket-ligand recognition tendency. For example, the main pocket-ligand recognition groups can be studied related to structural 3D conformation changes such as mutations or different partners (protein or ARN) bounding. Concerning uPA data, several mutations are made to facilitate crystallography. This complex clustering highlights, the double or triple mutation (C122A–N145Q-(D102T)) and the isolated C122S mutation, all located outside the binding site (Figure 1), are respectively observed in four complex clusters (C1, C2, C3, C4), clusters (C2, C3, C4, C5), Figure 3A2. This confirms that these mutations, located outside the binding site, have no impact on observed ligand binding corresponding to different clusters, as expected for crystallographic mutations that must not significantly modify the 3D structures. However, we can note that peptide complexes clustered in C1 all correspond to chains associated with the double or triple mutation (C122A–N145Q-(D102T)) while cluster C5 consists entirely of chains presenting the isolated C122S mutation. In the uPA data, observed crystallographic mutations do not appear to be associated with particular bound ligands and particular pocket-ligand recognition. However, the analysis of pocket-ligand pair clusters could serve to evaluate the possible impact of some mutations on the bound ligand profile, by studying and comparing clusters of complexed structures obtained with or without mutation(s).

This uPA example illustrates the possible a posteriori double reading of the joint clustering, to detect possible effect of ligand diversity or of pockets or 3D variability on the pocket-ligand recognition tendencies. A significant challenge for modeling drug-target interactions is to take into account both the protein target and the drug involved with target interaction mechanisms and functions.[64,65,66] The flexibility of a target binding site allows its binding to different ligands.[67] Interestingly, this clustering approach can also capture flexibility binding site information related to the ligand binding. If a sub-pocket is flexible but binds only one ligand profile, this classification will identify a unique cluster of pocket-ligand pairs corresponding to one unique sub-pocket cluster and ligand cluster, respectively. If a sub-pocket has little flexibility but binds two different ligands types, this classification will identify two pocket-ligand pair clusters corresponding to two clusters of pockets (weakly different) and to two ligand clusters. In this case, deeper structural analysis is necessary to understand if this pocket flexibility is explained either by its deformation due to the binding to different ligands, as detailed in,[66] or, on the contrary, by the fact that a deformation of the pocket impacts the ligands- type to which it is capable of binding.

We proposed a complex clustering approach to study promiscuous binding sites. Studying multiple ligands and multiple sub-pockets and their recognition can be an important way to identify a particular sub-pocket that can be bound to a specific ligand. This potentially helps select the ligands, from a large number of those that bind to the considered target, that bind particularly to this sub-pocket.

4 Conclusions

In this work, we presented a statistical modeling to characterize and optimally establish correspondences between pocket and ligand spaces of a promiscuous binding site. Taking into account pocket space and ligand space together, this approach successfully investigates multiple sub-pocket profiles associated with one unique binding site and the correspondences that exist with ligand profiles.

As a preliminary step to estimate the multiple sub-pockets, pocket estimation by ligand proximity was used to capture the specific part of the binding site (sub-pocket) in contact with the binding ligand. Interestingly, it is confirmed that binding site druggability can be predicted even for structures complexed with non-drug-like ligands using such pocket estimation (82% in the uPA case). This developed protocol provides a useful and detailed classification of pocket-ligand clusters associated with one promiscuous binding site. It highlights main pocket-ligand pair clusters, which can be interpreted in terms of the main pocket-ligand recognition tendency. Our results confirm that this proteochemometric protocol can establish statistical correspondences between pocket and ligand profiles and high-

lights main pocket-ligand recognition for promiscuous binding site for which there are corresponding 3D complexes that include structure or/and ligand data.

This resulting pocket-ligand profile correspondence can be used to evaluate whether certain clusters are related to a particular ligand or pocket profile. This analysis makes it possible to detect changes in 3D conformations (mutation or conformational differences such as protein bound or not bound to another protein or ARN) or ligand diversity, which would impact potentially the pocket-ligand recognition.

Thus, this approach can take advantage of the increasing number of PDB structures in holo form to improve knowledge of the protein-ligand interactions. We can note that the high redundancy observed in PDB structures is promising to study promiscuous binding site. An interesting perspective could be to develop profile prediction methods of one partner in the interactions, similarly to.[9] For instance, if it could be established a precise correspondence between some sub-pocket and ligand profiles associated with a promiscuous binding site, a possible prediction of ligands susceptible to bind this sub-pocket profile will consist in researching new ligand presenting highly similar profile. This perspective of profile prediction methods could be enriched by the integration of pharmacophore approaches such as[58,68] and available bioactivity data information, as proposed by Qiu et al., 2017.[7]

List of Abbreviations

- 3D (three dimensional)
- QSAR (Quantitative structure–activity relationship)
- PDB (Protein Data Bank)
- uPA (human urokinase-type plasminogen activator)
- ADME-Tox (Absorption, Distribution, Metabolism, Excretion and Toxicity)
- PCA (principal Component Analysis)
- NRDLD (Non Redundant dataset of Druggable and Less Druggable binding sites)
- 2D (two dimensional)

Conflict of Interest

None declared.

Acknowledgements

Funding from INSERM, Paris Diderot University, MTi UMRS-973, recurring funding and the BIP: BIP project, ANR-10-BINF-0003.

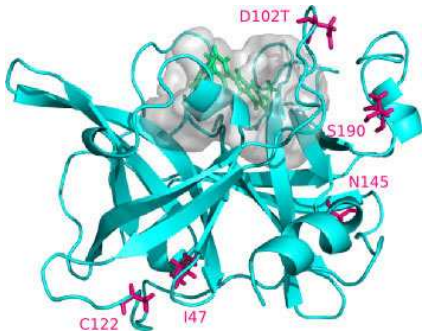
References

[1] S. P. Rohrer, E. T. Birzin, R. T. Mosley, S. C. Berk, S. M. Hutchins, D. M. Shen, Y. Xiong, E. C. Hayes, R. M. Parmar, F. Foor, S. W. Mitra, S. J. Degrado, M. Shu, J. M. Klopp, S. J. Cai, A. Blake, W. W. Chan, A. Pasternak, L. Yang, A. A. Patchett, R. G. Smith, K. T. Chapman, J. M. Schaeffer, *Science.* **1998**, *282*, 737–740.
[2] J. Drews, *Drug Discov. Today* **2003**, *8*, 411–420.
[3] M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt, J. E. Wikberg, *Biochim. Biophys. Acta.* **2001**, *1525*, 180–190.
[4] P. Prusis, R. Muceniece, P. Andersson, C. Post, T. Lundstedt, J. E. Wikberg, *Biochim. Biophys. Acta.* **2001**, *1544*, 350–357.
[5] G. J. van Westen, J. K. Wegner, P. Geluykens, L. Kwanten, I. Vereycken, A. Peeters, A. P. Ijzerman, H. W. van Vlijmen, A. Bender, *PLoS One* **2011**, *6*, e27518 (10.1371/journal.-pone.0027518).
[6] J. Meslamani, D. Rognan, *J. Chem. Inf. Mod.* **2011**, *51*, 1593–603.
[7] T. Qiu, J. Qiu, J. Feng, D. Wu, Y. Yang, K. Tang, Z. Cao, R. Zhu, *Briefings in Bioinformatics* **2017**, *18*, 125–136.
[8] W. Shoombuatong, P. Prathipati, V. Prachayasittikul, N. Schaduangrat, A. A. Malik, R. Pratiwi, S. Wanwimolruk, J. E. S. Wikberg, M. P. Gleeson, O. Spjuth, C. Nantasenamat, *Curr. Drug. Metab.* **2017** (doi: 10.2174/1389200218666170320121932).
[9] S. Pérot, L. Regad, C. Reynès, O. Spérandio, M. A. Miteva, B. O. Villoutreix, A-C. Camproux, *PLoS. One.* **2013**, *8*, e63730 (doi: 10.1371/journal.pone.0063730).
[10] C. A. Lipinski, *Drug Discov. Today. Technol.* **2004**, *1*, 337–341.
[11] X. Jalencas, J. Mestres, *Mol. Inform.* **2013**, *32*, 976–990.
[12] P. Gedeck, B. Rohde, C. Bartels , *J. Chem. Inf. Model.* **2006**, *46*, 1924–1936.
[13] D. Rognan, *Brit. J. Pharmacol.* **2007**, *152*, 38–52.
[14] A. Bender, R. C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
[15] T. I. Oprea, *J. Comput. Aided. Mol. Des.* **2000**, *14*, 251–264.
[16] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166.
[17] R. Todeschini, V. Consonni in *Handbook of Molecular Descriptors* (Eds: Wiley-VCH), New-York, **2008**.
[18] Cerisier et al. **2017 [In publication]**.
[19] H. M. Berman, J. Westbrook, *Nucleic Acids Res.* **2000**, *28*, 235–242.
[20] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, B. O. Villoutreix, *Drug Discov. Today* **2010**, *15*, 656–667.
[21] M. Gao, J. Skolnick, *PLoS Comput. Biol.* **2013**, *9*, e1003302 (doi:10.1371/journal.pcbi.1003302).
[22] G. Caumes, A. Borrel, H. Abi. Hussein, A. C. Camproux, L. Regad, *Mol. Inform.* **2017** (doi: 10.1002/minf.201700025).
[23] L. Benkaidali, F. André, B. Maouche, P. Siregar, M. Benyettou, F. Maurel, M. Petitjean, *Bioinformatics* **2014**, *30*, 792–800.
[24] E. Harigua-Souiai, I. Cortes-Ciriano, N. Desdouits, T. E. Malliavin, I. Guizani, M. Nilges, A. Blondel, G. Bouvier, *BMC Bioinf.* **2015**, *19*, 16–93.
[25] M. Petitjean, In: Zadnik Stirn, L., Žerovnik,J. Povh,J. Drobne,S., Lisec,A. (eds. Proceedings of SOR'13), the 12th International Symposium on Operational Research in Slovenia, Slovenian Society INFORMATIKA, **2013**.
[26] E. Freyhult, P. Prusis, M. Lapinsh, J. E. Wikberg, V. Moulton, M. G. Gustafsson, *BMC. Bioinf.* **2005**, *10*, 6–50.
[27] C. Kramer, P. Gedeck, *J. Chem. Inf. Model.* **2011**, *51*, 707–720.
[28] M. Lapins, A. Worachartcheewan, O. Spjuth, V. Georgiev, V. Prachayasittikul, C. Nantasenamat, J. E. Wikberg, *PLoS One* **2013**, *8*, e66566 (doi: 10.1371/journal.pone.0066566).
[29] J. Meslamani, D. Rognan, *J. Chem. Inf. Model.* **2011**, *51*, 1593–1603.

[30] D. Brown, G. Superti-Furga, *Drug. Discov. Today.* **2003**, *8*, 1067–1077.
[31] H. Abi Hussein, C. Geneix, M. Petitjean, A. Borrel, D. Flatters, A.-C. Camproux, *Drug Discov. Today* **2017**, *22*, 404–415.
[32] A. Borrel, L. Regad, H. Xhaard, M. Petitjean, A.-C. Camproux, *J. Chem. Inf. Model.* **2015**, *55*, 882–895.
[33] H. Abi Hussein, A. Borrel, C. Geneix, M. Petitjean, L. Regad, A–C. Camproux, *Nucleic Acids Res.* **2015**, *43*, W436-42.
[34] G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, A. L. Hopkins, *Nat. Biotechnol.* **2006**, *24*, 805–815.
[35] M. J. Keiser, 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, B. L. Roth, *Nature* **2009**, *12*, 175–181.
[36] S. I. Berge, R. Iyengar, *Bioinformatics* **2009**, *25,* 2466–2472.
[37] A. Lavecchia, C. Cerchia, *Drug Discov. Today* **2016**, *21*, 288–298
[38] V. J. Haupt, S. Daminelli, M. Schroeder, *PLoS One* **2013**, *8*, e65894 (doi: 10.1371/journal.pone.0065894).
[39] J. Bajorath, *Mol. Inform.* **2016**, *35*, 583–587.
[40] A. Krasowski, D. Muthas, A. Sarkar, S. Schmitt S, R. Brenk, *J. Chem. Inf. Mod.* **2011**, *51*, 2829–2842.
[41] B. Degryse, *Curr. Pharm. Des.* **2013**, *17*, 1872–1873.
[42] J. V. Braaten, S. Handt, W. G. Jerome, J. Kirkpatrick, J. C. Lewis, R. R. Hantgan, *Blood.* **1993**, *81*, 1290–1299.
[43] P. Ragno, *Cell. Mol. Life. Sci.* **2006**, *63*, 1028–1037.
[44] F. Blasi, P. Carmeliet, *Nat. Rev. Mol. Cell. Biol.* **2002**, *3*, 932–943.
[45] V. Nienaber, J. Wang, D. Davidson, Jack Henkin, *J. Biol. Chem.* **2000**, *275*, 7239–7248.
[46] L. Gong, V. Proulle, C. Fang, Z. Hong, Z. Lin, M. Liu, G. Xue, C. Yuan, L. Lin, B. Furie, R. Flaumenhaft, P. Andreasen, B. Furie, M. Huang, *Journal of cellular and molecular medicine* **2016**, *20*, 1851–1860.
[47] D. Lagorce, O. Sperandio, J. B. Baell, M. A. Miteva, B. O. Villoutreix, *Nucleic Acids Res.* **2015**, *43*, W200-207.
[48] M. A. Miteva, S. Violas, M. Montes, D. Gomez, P. Tuffery, B. O. Villoutreix, *Nucleic Acids Res.* **2006**, *34*, W738-44.
[49] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.
[50] M. Petitjean, RADI, version 4.0, **2014**. http://petitjeanmichel.-free.fr/itoweb.petitjean.freeware.html.
[51] M. Petitjean. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.
[52] M. Hollander, D. A. Wolfe in *Nonparametric Statistical Methods.* (Eds: John Wiley & Sons.), New York, **1973**, 27–33.
[53] R Development Core Team, in R: A language and environment for statistical computing, Vienna, Austria, **2013**, URL: http://www.R-project.org/
[54] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug. Deliv. Rev.* **1997**, *46*, 3–26.
[55] T. J. Ritchie, S. J. Macdonald, *Drug Discov. Today.* **2014**, *19*, 489–495.
[56] R. P. Sheridan, V. N. Maiorov, M. K. Holloway, W. D. Cornell, Y.-D Gao, *J. Chem. Inf. Model.* **2010**, *50*, 2029–2040.
[57] C. Schalon, J. S. Surgand, E. Kellenberger, D. Rognan, *Proteins Struct. Funct. Bioinf.* **2008**, *71*, 1755–1778.
[58] N. Weill, D. Rognan, *J. Chem. Inf. Model.* **2010**, *50*, 123–135.
[59] K. Yeturu, N. Chandra, *BMC Bioinf.* **2008**, *9*, 543.
[60] H. J. Feldman, P. Labute, *J. Chem. Inf. Mod.* **2010**, *50*, 1466–1475.
[61] S. Le, J. Josse, F. Husson, *Journal of Statistical Software* **2008**, *25*, 1–18.
[62] J. H. Ward, *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
[63] M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis, P. D. Leeson, *J. Med. Chem.* **2003**, *46*, 1250–1256.
[64] C. S. Goh, D. Milburn, M. Gerstein, *Curr. Opin. Struct. Biol.* **2004**, *14*, 104–109.
[65] R. Grünberg, J. Leckner, M. Nilges, *Structure* **2004**, *12*, 2125–2136.
[66] M. F. Lensink, R. Méndez, *Curr. Pharm. Biotechnol.* **2008**, *9*, 77–86.
[67] S. Surade, T. L. Blundell, *Chem. Biol.* **2012**, *19*, 42–50.
[68] F. Bonachera, G. Marcou, N. Kireeva, A. Varnek, D. Horvath, *Bioorg. Med. Chem.* **2012**, *20*, 5396–5409.
[69] The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.
[70] J. S. Delaney, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
[71] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615–2623.
[72] F. Milletti, A. Vulpetti, *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.
[73] J. Kyte, R. F. Doolittle, *J. Mol. Biol.* **1982**, *157*, 105–132.

Received: February 27, 2017
Accepted: June 26, 2017
Published online on ■■ ▮▮, 0000

FULL PAPER



molecular informatics
models – molecules – systems

Supporting Information

Supporting information

Table S1: List of descriptors used in this study to describe uPA-ligand set and uPA-pocket set. Type of descriptors: (*) PC(L) physicochemical (ligand); PC(P) physicochemical (pocket); G(L) geometrical (ligand); G(P), geometrical (pocket).

Name of descriptors (unit)	Description	Type(*)	Reference
MW (Da)	Molecular weight	PC(L)	FAF-Drugs3 web server [47]
logP	The logarithm of the partition coefficient between n-octanol and water, characterizing lipophilicity	PC(L)	FAF-Drugs3 web server [47]
logD	The logP of compounds at physiological pH (7.4)	PC(L)	FAF-Drugs3 web server [47]
logSw	The logarithm of compounds water solubility computed by the ESOL method	PC(L)	FAF-Drugs3 web server [47] et [70]
ratioH.C	The ratio between the number of non-carbon atoms and the number of carbon atoms	PC(L)	FAF-Drugs3 web server [47]
RADIUS.HULL_L (Å)	Radius of the smallest enclosing sphere	G(L)	Petitjean 1992 [5149]
PSI_lig	Equivalent descriptors of the PSI, computed on the ligand. PSI: Pocket Sphericity Index is the ratio of the radius of the largest sphere inscribed in the hull to the radius of the smallest enclosing sphere. Closer PSI_lig is to 1, more spherical the ligand hull is is. A small PSI-lig value indicates that the ligand hull is flat.	G(L)	Borrel et al, 2015 [32]
RotatableB	Number of rotatable bonds corresponds to the number of any single non-ring bond, bounded to nonterminal heavy (i.e., non-hydrogen) atom.	PC(L)	FAF-Drugs3 web server [47] et [71]
p_charged_residues	Frequency of charged residues in pocket (D, E, R, K, H)	PC(P)	Borrel et al, 2015 [32]
p_positive_residues	Frequency of positive residues in pocket (H, K, R)	PC(P)	Borrel et al, 2015 [32]
p_main_chain_atom	Frequency of main chain atoms in pocket	PC(P)	Borrel et al, 2015 [32]
p_N_atom	Frequency of N atoms in pocket	PC(P)	PockDrug web server [33]; Milletti et

Supporting information

			al. 2010 [72]
hydrophobic_kyte	Hydrophobicity based properties of residues	PC(P)	PockDrug web server [33]; Kyte et al. 1982 [73]
p_hydrophobic_atom	Proportion of hydrophobic atoms in pocket (atoms of residues C, G, A, T, V, L, I, M, F, W, Y, H, K)	PC(P)	Borrel et al, 2015 [32]
p_aromatic_residues	Frequency of aromatic residues in pocket (F, Y, H, W)	PC(P)	Borrel et al, 2015 [32]
RADIUS_HULL_P (Å)	Radius of the smallest enclosing sphere of the pocket	G(P)	Petitjean 1992 [51]Petitjean 1992 [49]
C_RES	Number of residues in pocket	G(P)	In-house team

Table S2: Average and standard deviation values of nine ligand descriptors are provided for non-peptide and peptide ligands of uPA ligand set. Comparison between non-peptide and peptide ligands is performed with a Wilcoxon test.

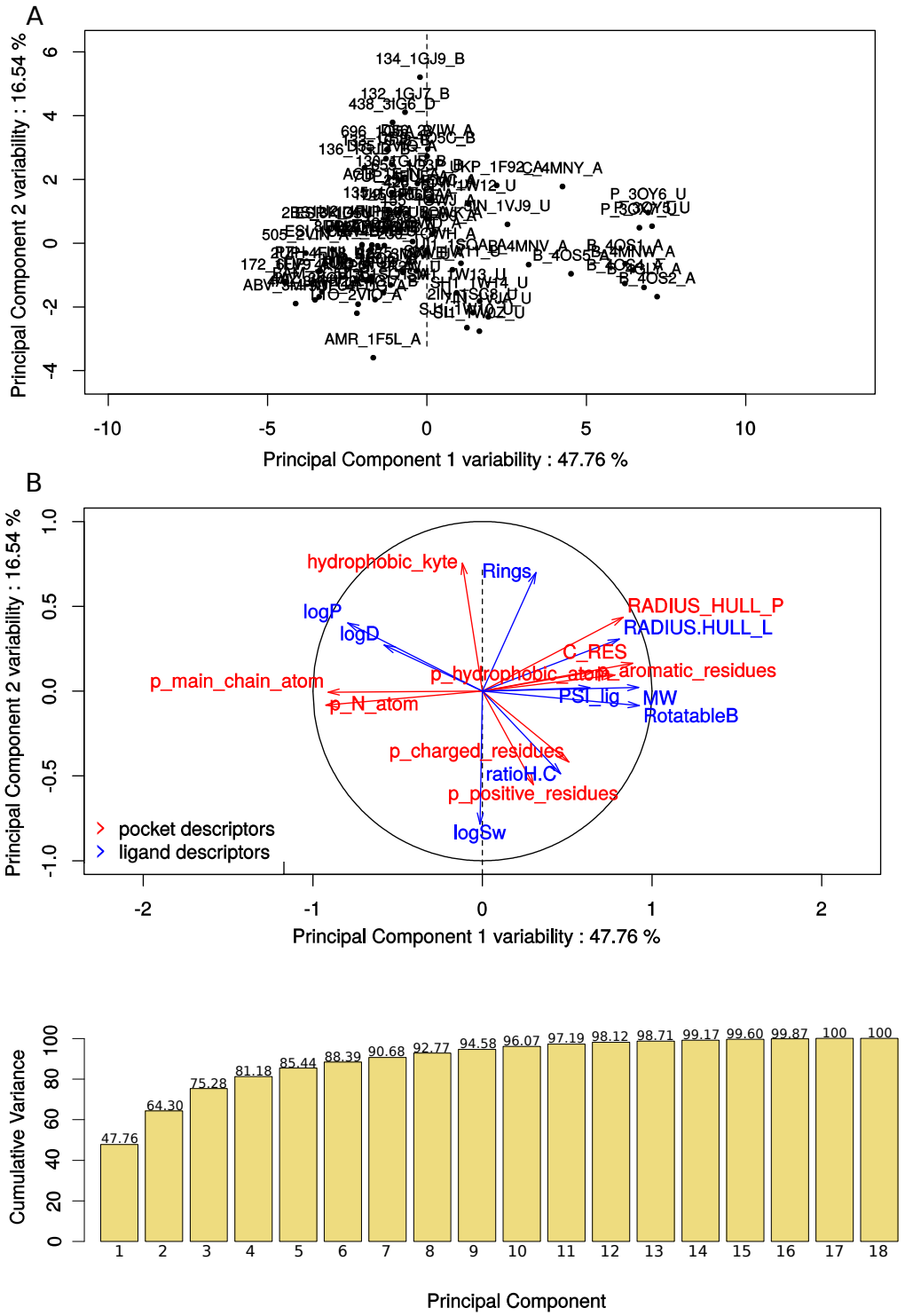
	ligand		peptide		
	mean	sd	mean	sd	P-value
MW	331.54	115.85	1550.62	132.32	<10-6
logP	1.15	1.71	-8.5	3.13	<10-6
logD	0.56	1.77	-7.78	9.04	<2.10-2
logSw	-2.51	1.34	-2.49	2	0.64
RotatableB	4.91	4.02	24.73	5	<10-6
Rings	1.95	0.88	2.18	0.6	0.26
ratioH.C	0.39	0.21	0.71	0.06	<10-5
RADIUS.HULL	6.35	1.86	11.87	2.6	<10-5
PSI_lig	0.2	0.12	0.3	0.06	<10-2

Figure S1: A) Representation of uPA complex space in PCA using our 18 representative descriptors. The first PCA plane captures a total of 64.33% of the variability. Complexes are named as follows: 1 (peptides) or 3 (non-peptides) letters, then the protein PDB code and one letter to indicate the chain of the protein. B) Projection of our 18 representative descriptors on the first PCA plane. The descriptors are colored based on their type: ligand or pocket descriptors. The descriptors are explained in Table Supporting Information S1.

As indicated by the projection of the variables close to the PCA correlation circle, the different selected pocket and ligand descriptors contribute to the variability of the data and are relevant for the study. Pocket p_main_chain_atom and p_N_atom are strongly negatively correlated to the pocket geometrical descriptors and the p_aromatic_residues, p_hydrophobic_atom descriptors are related to the Principal Component 1 whereas the p_positive_residues and p_charged_residues descriptors are negatively correlated with hydrophobic_kyte and are related to Principal Component 2. The ligand logP and logD descriptors are strongly correlated. Peptide ligands seem to be separated by pocket

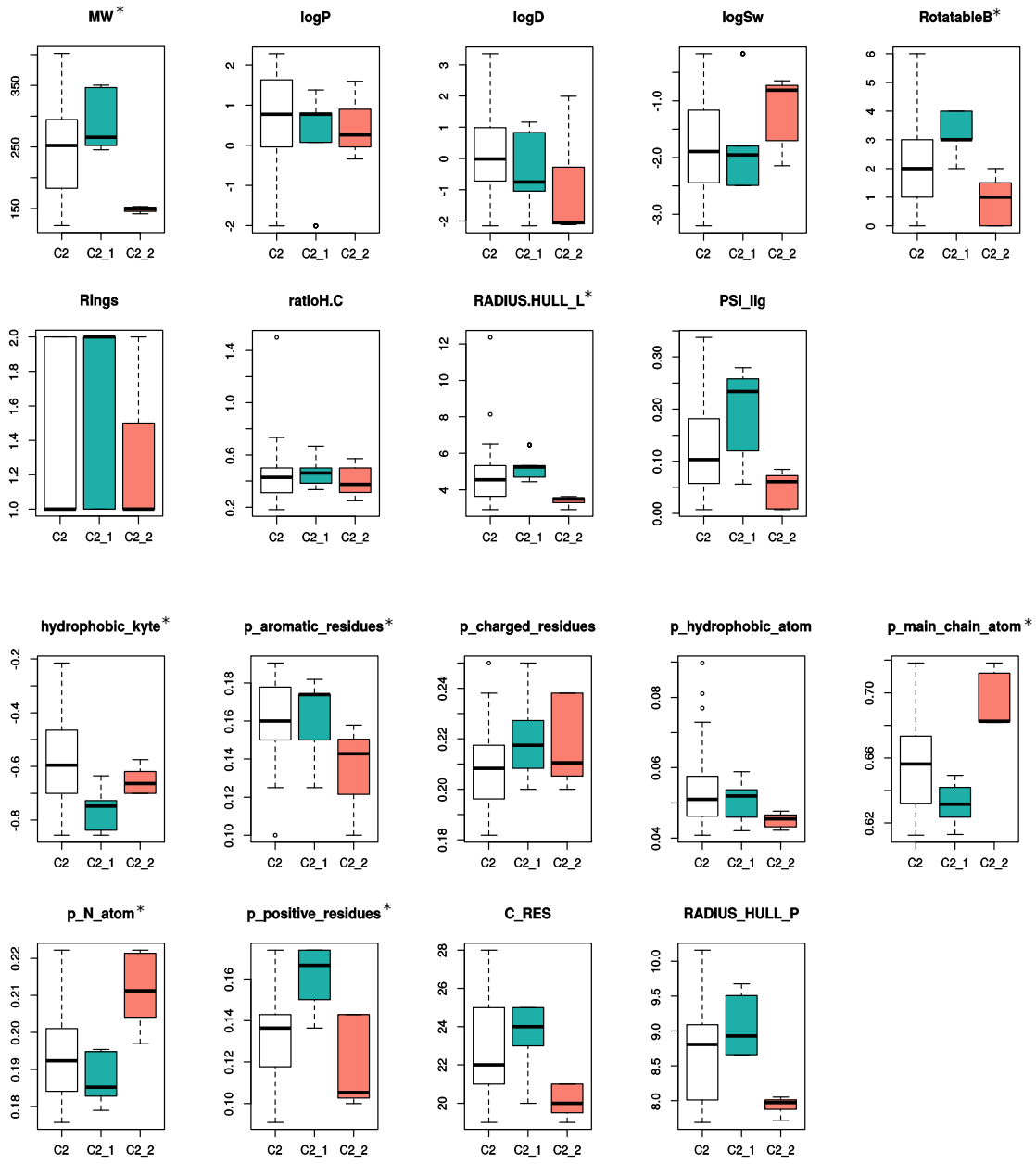
Supporting information

size of their pockets and ligand size. C) Barplot showing the cumulative percentage of variance for the 18 first Principal components. The percentage is explicitly marked at the top of the bars. The first bar with more than 95% of cumulative variance is the 10th.



Supporting information

Figure S2: Boxplot of the 18 descriptors computed on 27 complexes of cluster C2 and respectively six and four complexes predicted as non-druggable by PockDrug of sub-clusters C2_1 and C2_2. The horizontal black bar represents the median, the colored rectangle (white, green or orange) represents 50% of the data, and the dotted lines represent the first and fourth 25% of the data. Black circles are outlier individuals. Significant descriptors are indicated by *. It illustrates four complexes of sub-cluster C2_1 corresponds to rather high MW, rotatableB and PSI_lig values, associated with significantly lower hydrophobicity_kyte and higher p_positive_residues pockets, (p-values < 0.01, 0.02). Six complexes of C2_2 sub-cluster corresponds to significantly lower MW, rotatableB and RADIUS.HULL_L values (p-values < 0.02, 0.04, 0.02) and to sub-pockets with significantly lower p_aromatic_residues and RADIUS_HULL_P, higher p_main_chain_atom, p_N-atom values (p-values< 0.05, 0.05, 0.02, 0.03, results not shown) than C2.



Chapitre 3. ÉTUDE DE LA PROMISCUITE

Dans ce chapitre, je présente l'étude développée dans le but de déterminer la promiscuité des sites de liaison (leur capacité à se lier à un seul, ou plusieurs ligands différents) et la caractérisation de leur propriétés, associées à leur niveau de promiscuité. Ces travaux ont donné lieu à une publication, section 3.4. Cette étude a pour but d'améliorer la compréhension des mécanismes d'interactions entre protéines et ligands (candidats médicaments) afin de développer un outil capable de prédire ces interactions et, à plus long terme, de comprendre l'impact sur la polypharmacologie.

3.1 Intérêt de l'étude de la promiscuité

La caractérisation des deux acteurs d'une interaction entre une protéine et un ligand nous permet d'en savoir plus sur les mécanismes de cette interaction, mais aussi la manière d'empêcher ou de renforcer cette interaction. Comme démontré dans la section précédente, il existe des profils de ligands qui se lient à des profils spécifiques de poches (et inversement). Ces correspondances se compliquent lorsque l'interaction est multipartenaire, comme c'est le cas pour l'urokinase. Cependant, de plus en plus de protéines sont identifiées comme capables de lier des ligands différents (cf. section 1.3.3, le concept clé-serrure).

Comme cette promiscuité semble à l'origine des effets secondaires, elle reçoit de plus en plus d'attention de la part de la communauté scientifique.

Notre caractérisation conjointe des poches et des ligands a donc été utilisée afin de mieux comprendre la différence, en termes de propriétés physicochimiques et géométriques, entre les poches promiscuous et les non-promiscuous (appelées aussi « sélectives ») ainsi qu'entre leurs ligands (ligands liés aux sites de liaison promiscuous vs ligands liés aux sites de liaison sélectifs), mais aussi les différences de propriétés entre les ligands promiscuous et non-promiscuous (aussi appelés « sélectifs »).

Peu d'études fournissent les caractéristiques de propriétés physicochimiques et géométriques détaillées des poches ainsi que des ligands, récemment (Desaphy et al. 2013; Bosc et al. 2015), et aucune ne montre à notre connaissance les différences de ces caractéristiques détaillées en fonction de la promiscuité ou la sélectivité des partenaires. Cette étude apporte donc de nombreuses informations, primordiales à la connaissance des mécanismes des interactions multipartenaires. Elle est la seconde étape (après la caractérisation conjointe) dans notre processus de prédiction des interactions et, à plus long terme, la compréhension des mécanismes régissant les effets secondaires.

3.2 Importance du jeu de données

La croissance récente des données structurales et d'interactions dans les bases de données publiques nous permettent d'étudier à grande échelle les interactions entre les protéines et les ligands (candidats)-médicaments et détecter des tendances fiables et des conclusions

significatives (Y. Hu and Bajorath 2013). Cependant il existe des limites à cet apprentissage massif à partir des données.

Le premier à mentionner est le caractère incomplet des données (Mestres et al. 2009). En effet, nous pouvons conclure sur les données dont nous disposons, *c-à-d* les interactions qui ont été testées et résolues. En, revanche, de nombreuses interactions ne sont pas référencées. Cela peut être dû au fait que les deux composés (protéine et ligand) n'interagissent pas ensemble, ou bien au fait qu'ils n'ont pas été testés expérimentalement puis cristallisés l'un contre l'autre, par manque de temps et de moyens. Dans les deux cas, l'information n'apparaît pas dans la base de données et peut biaiser les analyses en générant des « faux négatifs ».

Une autre limite, qui est plutôt un biais, est due à la redondance des données (mentionnée en section 2.1.4). Certaines familles de protéines sont plus étudiées que d'autres, dû à leur intérêt thérapeutique ou à la moindre difficulté à résoudre sa structure tridimensionnelle. Les enzymes les plus représentées de la PDB (juillet 2019) sont les hydrolases (~35 %) et les transférases (~32 %) alors que les ligases et les translocases sont les moins représentées (respectivement 2 % et 0,9 % des fichiers).

Enfin, une autre limite peut être abordée lors de l'étude des données tridimensionnelles de la PDB : la qualité des fichiers. Puisque la plupart des structures protéiques sont résolues par cristallographie (section 1.2.2), la qualité peut être évaluée par la résolution. On dénombre dans la PDB environ 40 000 structures protéiques ayant une résolution supérieure à 2,5 Å, représentant environ 30 % des protéines cristallisées. Cette incertitude sur la position des atomes peut engendrer des différences d'estimation de poche ou d'interaction.

Le choix du jeu de données est donc primordial lors d'une étude à grande échelle de structures tridimensionnelles.

3.2.1 MOAD

Pour l'étude de la promiscuité des protéines, une base de données nous a paru très pertinente pour y extraire notre jeu de données : la base de données MOAD *(Mother Of All Databases)*. Les structures qu'elle contient sont extraites de la PDB et filtrées selon certains critères (section 1.2.1) :

- La nature des structures. Ce sont des protéines liées à un ou plusieurs ligands.
- La résolution des structures. Elle doit être inférieure à 2,5 Å, justifiant ainsi d'une bonne qualité.
- La nature des ligands de chaque structure. Ils doivent être « biologiquement valides » comme par exemple des peptides de 10 acides aminés ou moins, des oligonucléotides de 4 nucléotides ou moins, des petites molécules organiques, des cofacteurs, etc. et ne peuvent être des additifs cristallographiques, des sels, des tampons ou des solvants, ni des métaux.
- Les informations d'interaction. Si l'interaction entre la protéine et le ligand est annotée avec des données de liaison déterminées de manière expérimentale, l'information est extraite de la littérature (IC_{50}, K_i ou K_d). C'est le cas pour environ 35 % des données (Ahmed et al. 2015).

Lors du début de l'étude, la MOAD était composée de 25 769 structures tridimensionnelles, comprenant 12 440 ligands uniques et un total de 44 675 interactions protéine–ligand. Actuellement (version 2017, en juillet 2019), elle compte 32 747 structures et 16 044 ligands uniques, soit une augmentation d'environ 30 % du nombre de structures en 3 ans.

De la même manière que la PDB, la MOAD propose une classification en « famille », basée sur la classification enzymatique (numéro E.C.) et les annotations de la PDB. Cela permet donc de tirer des conclusions à un plus haut niveau de classification que celui de la protéine elle-même. Cependant, la classification pour les protéines non-enzymatiques diffère un peu et la comparaison avec la PDB ou d'autres jeux de données est parfois impossible.

La MOAD contient donc un grand nombre d'interactions et de structures de haute résolution, c'est donc une base pertinente pour extraire notre jeu de données.

3.2.2 Autres études de la promiscuité

De nombreuses études traitent de la promiscuité des ligands et de plus en plus traitent de la promiscuité des protéines. Leurs jeux de données sont divers.

Une étude menée par Mestres et al. (Mestres et al. 2009) a permis de constituer un jeu de données extrait de 7 bases de données différentes. Ils ont rassemblé un total de 4 767 interactions uniques entre 802 médicaments et 480 protéines cibles. Les interactions ont été étudiées et visualisées grâce à des réseaux. Il a été mis en avant que la topologie des réseaux médicaments-cibles dépend implicitement de l'exhaustivité des données, des propriétés des médicaments et des familles des cibles. Il a été démontré qu'en moyenne, chaque médicament du jeu de données étudié dans cet article interagit avec 6 cibles.

Une autre étude portée sur la promiscuité des ligands a créé son jeu de données à partir de différentes bases de données (Haupt, Daminelli, and Schroeder 2013). Ils ont utilisé 706 médicaments (dont 164 qui se lient à plus de trois cibles) et 712 protéines non redondantes (95 % d'identité de séquence). Leurs résultats montrent que la promiscuité des ligands ne corrèle pas avec leur poids moléculaire ou leur hydrophobicité. En revanche elle corrèle avec la similarité locale (et structurale) entre les sites de liaison qu'ils lient.

Une analyse de Barelier et al. (Barelier et al. 2015) traite 59 ligands dans 116 complexes (62 paires au total) et concerne des ligands capables de se lier à des protéines de différentes familles. Ils ont montré que la majorité des sites de liaison liés par le même ligand ne partageaient pas de résidus communs et ils ont conclu qu'il n'existait pas de « code » d'appariement de motifs unique permettant d'identifier les sites de liaison qui liaient des ligands identiques dans des protéines non apparentées. Ils soulignent donc la difficulté de prédire les interactions *« off-target »* des molécules pour des protéines de différentes familles.

Les premières études à grande échelle de la promiscuité des protéines surviennent la même année avec l'étude de Gao et Skolnick (Gao and Skolnick 2013). À partir de 20 000 complexes protéine–ligand extraits de la PDB, ils se sont penchés sur plusieurs questions pertinentes dans l'étude de la promiscuité dont les suivantes :

- quels sont les principaux types de poches qui accueillent les ligands ?
- dans quelle mesure la similarité des sites de liaison permet de prédire les interactions protéine–ligand ?

Chacune de ces questions s'est vu apporter une réponse. Premièrement, les 20 000 complexes ont permis de déterminer le nombre plateau de « types de poches » qui avoisinerait les 1 300. L'espace des poches est donc beaucoup plus réduit que celui des ligands.

Concernant la similarité des sites, l'étude montre qu'il est possible de détecter les poches similaires qui lient des ligands similaires en utilisant l'identité de séquence pour les poches et la similarité de Tanimoto pour les ligands. Entre 50 % et 60 % des complexes protéiques ont une poche dont il existe une poche similaire liée par un ligand similaire (coefficient de Tanimoto supérieur à 0,7). Cependant, l'étude montre aussi que les poches promiscuous sont très nombreuses et qu'elles lient de nombreux ligands aux propriétés chimiques différentes. La prédiction des interactions est donc possible, mais elle n'est pas optimale par la seule utilisation de la similarité des sites de liaison. La conclusion de promiscuité dans cette étude est tout de même un peu surestimée puisque les 20 000 complexes ont été regroupées en 1 300 groupes de poches.

Les auteurs se penchent aussi sur les différences entre les ligands qui se lient à une même poche ou, à l'inverse, les poches qui sont liées par le même ligand, mais ne fournissent pas les caractéristiques précises qui les différencient. La géométrie (plasticité) serait la cause de l'adaptabilité des poches à différents ligands, tandis que les ligands garderaient les mêmes types d'interactions pour se lier à différentes poches. La similarité des poches, même faible, serait aussi la raison pour laquelle un même ligand peut se lier à plusieurs poches de protéines non-homologues.

La dernière étude détaillée dans ce manuscrit a été réalisée en 2015 par Hu et Bajorath (Y. Hu and Bajorath 2015). Leur étude porte sur la promiscuité des molécules, mais aussi des cibles. Leur jeu de données comporte environ 118 000 composés et 1 600 cibles, pour un total d'environ 160 000 interactions, toutes annotées avec des données d'activité comme l'IC_{50} ou le K_i. L'analyse de ces nombreuses interactions a mis en avant une préférence marquée des cibles promiscuous pour les ligands sélectifs alors que le contraire était attendu. Ils soulignent aussi le fait que les caractéristiques structurelles des cibles qui sont en corrélation avec leur capacité à interagir avec des composés promiscuous ou sélectifs sont actuellement inconnues.

Ces travaux ont été la base de nos connaissances préliminaires sur la promiscuité des poches et des ligands. Notre étude de la promiscuité, basée sur un jeu de données de haute qualité, devait donc porter conjointement sur les ligands et les sites de liaison en fournissant les caractéristiques significatives entre les entités sélectives ou promiscuous (poches et ligands).

3.3 Protocole

3.3.1 Obtention des données

L'étape préliminaire de l'analyse de la promiscuité des poches et des ligands est d'obtenir un jeu de données pertinent et complet. Comme mentionné dans la section 2.4.1, nous avons

extrait nos données de la MOAD. En effet, tous les complexes de MOAD ne sont pas pertinents à l'étude puisque les ligands doivent avoir des propriétés médicamenteuses. La première sélection a donc été opérée sur les ligands disponibles. Le filtrage a été réalisé en utilisant l'outil FaF-Drugs3 (Miteva et al. 2006) qui sélectionne les ligands aux propriétés compatibles avec la « règle des 5 » de Lipinski (Lipinski et al. 2001) (cf. chapitre 1 et méthode détaillée dans l'article ultérieurement, section 2.6).

Une sélection a aussi été faite pour les poches estimées par proximité au ligand. Nous avons fait le choix de ne pas traiter les poches qui se trouvent à l'interface entre deux chaînes d'une même protéine. Bien que ce ne soit pas un phénomène rare (par exemple pour la protéase du VIH-2 (Triki, Fartek, et al. 2018)), les algorithmes d'alignement et de comparaison pour les séquences et les structures ne sont pas en mesure d'exécuter cela sur plusieurs chaînes. De plus, le nombre de chaînes dans la structure cristallographique ne correspond pas forcément au nombre de chaînes de l'unité biologique de la protéine, c'est-à-dire le nombre de chaînes qu'elle possède dans sa forme fonctionnelle. Un ligand qui est à la surface d'une chaîne de la protéine peut se retrouver à l'interface de deux chaînes dont l'une a été dupliquée pour les besoins cristallographiques (appelée « unité symétrique »). Il nous est très difficile de différencier de manière simple, efficace et automatique les ligands qui sont à l'interface des unités biologiques de ceux à l'interface des unités symétriques. Les poches aux interfaces ont donc été exclues du jeu de données. En conséquence, chaque fichier de structure a été divisé en autant de chaînes protéiques qui le composent, afin de faciliter la suite du protocole. Le jeu de données peut ainsi être amélioré sur cet aspect.

Pour chacun des complexes, si l'un des deux partenaires n'est pas retenu pour le jeu de données, l'autre n'est pas pris en compte pour l'étude. En effet, la caractérisation doit être conjointe afin de pouvoir prédire les interactions.

Aucune sélection particulière n'a été appliquée sur les fichiers de structures tridimensionnelles de la MOAD, ils sont juste le résultat de la sélection des poches et des ligands selon les critères expliqués précédemment. Nous sommes partis des 25 769 complexes extraits de MOAD et le filtrage nous a permis de garder 4 749 complexes entre 481 sites de liaison et 1 969 ligands différents.

3.3.2 Analyse de la promiscuité

La difficulté principale dans l'analyse de la promiscuité est de déterminer les critères à partir desquels deux structures sont considérées comme identiques. En effet, une structure résolue ne peut figer qu'une seule interaction (au même site de liaison) entre une protéine et un ligand. C'est la principale cause de la redondance des structures tridimensionnelles dans la PDB (et dans la MOAD). Par conséquent si un site de liaison d'une protéine est multipartenaire, il faut la résoudre plusieurs fois avec ses différents partenaires. Une poche, définie comme les atomes proches d'un ligand, ne peut donc pas être caractérisée comme promiscuous. Il convient donc de parler de promiscuité de son site de liaison (décrit par un ensemble de poches).

Les ligands sont aussi redondants dans les bases de données structurales. Dans la PDB, ils sont différents dès lors que leur nom (HET code) est différent. Cependant, deux ligands aux

noms différents peuvent être des isomères, ils seront donc très proches en termes de structure tridimensionnelle et de propriétés chimiques.

L'étape préliminaire de l'analyse a donc été le regroupement des structures homologues en « clusters » afin qu'un cluster ne représente qu'un seul site de liaison puis de regrouper les ligands similaires en termes de coefficient de Tanimoto afin de ne pas surestimer la promiscuité des cibles.

Nous avons choisi de baser le regroupement de structures protéiques en utilisant deux méthodes :

- Regroupement basé sur la similarité de séquence, avec H-CD-HIT (Huang et al. 2010). Il permet de déterminer les protéines similaires. Toutes les chaînes protéiques similaires, définies comme partageant au moins 80 % de séquence, sont regroupées. Pour chaque groupe, une structure représentante est déterminée par l'outil.
- Regroupement basé sur l'alignement de structure tridimensionnelle avec TM-Align (Y. Zhang and Skolnick 2005). Il permet de déterminer les sites de liaison de chaque protéine. Une fois les protéines similaires regroupées, elles sont alignées dans l'espace 3D, par rapport à la structure représentante déterminée précédemment. Cet alignement permet de comparer l'emplacement de toutes les poches sur la structure d'une même protéine. Les poches qui se superposent (critère expliqué dans l'article, section 2.6) décrivent un unique site de liaison. C'est la promiscuité de ce site de liaison qui est analysée. Un site de liaison est lié par autant de ligands que de poches qui le composent.

Le regroupement des ligands similaires s'est fait à l'aide du coefficient de Tanimoto (critère expliqué dans l'article, section 2.6). Les ligands dont le Tanimoto est supérieur ou égal à 0,8 sont considérés comme similaires et formeront un cluster de ligands. Ce critère permet d'éviter de surestimer la promiscuité des protéines. En effet, une protéine dont le site de liaison est lié par plusieurs ligands très proches (par exemple deux anomères ou deux énantiomères) ne peut pas être considérée comme multipartenaire.

Dans cette partie de l'analyse, nous avons été confronté à une des limites énoncées en section 2.4. En effet, le caractère incomplet des données est un phénomène à prendre en compte. Un site de liaison qui n'a qu'une poche pour le décrire (et donc qu'un seul ligand), ne peut pas être qualifié de promiscuous. Nous avons fait le choix de réduire notre jeu de données aux sites de liaison qui sont décrits par au moins 4 poches. Le choix du nombre de poches par site de liaison est un compromis entre le nombre de sites de liaison disponibles ayant plus de 4 poches et le fait que plus il y a de poches par site de liaison, plus la conclusion de promiscuité ou sélectivité est fiable. Cela ne pallie pas totalement au caractère incomplet des données, mais nous permet d'atténuer ce phénomène.

Un sous jeu de données (appelé DBS4 dans l'article) a donc été créé, il comporte :

- 3 488 ligands regroupés en 1 969 clusters de ligands,
- 7 267 poches regroupées en 481 sites de liaison,

- 4 749 fichiers de structures tridimensionnelles, regroupés en 459 clusters de chaînes homologues.

La promiscuité des deux acteurs des nombreuses interactions a donc pu être mesurée et caractérisée.

3.4 La promiscuité joue un rôle majeur dans la polypharmacologie

Dans ce travail, la caractérisation conjointe des poches et des ligands par la méthode originale décrite précédemment nous a permis de mettre en évidence les propriétés qui jouent un rôle important pour la promiscuité des poches et des ligands. L'originalité de cette étude réside dans les critères de sélection du jeu de données, la caractérisation précise des acteurs et l'analyse poussée de la promiscuité à différents degrés (non promiscuous, promiscuous et très promiscuous) en intégrant les données concernant la famille de protéine.

Les résultats confirment que la promiscuité des sites de liaison n'est pas un phénomène exceptionnel (80 % des sites de liaison de notre jeu de données sont promiscuous), indépendamment de la famille de protéine. L'analyse des caractéristiques des sites de liaison promiscuous confirme que leurs poches de liaison sont grandes, hydrophobes et compatibles avec une flexibilité qui a été montrée comme jouant un rôle dans la promiscuité (Pabon and Camacho 2017).

Les poches des sites de liaison sélectifs sont moins favorables aux interactions, car elles ont tendance à être petites, avec une forte proportion d'atomes dans la chaîne latérale, une faible proportion d'atomes de soufre et de résidus aliphatiques ; ou bien elles sont grandes mais faiblement hydrophobes.

Aussi, les sites de liaison sélectifs interagissent avec une petite partie de l'espace des ligands (4 %), ce qui correspond à de petits ligands très adaptables et hydrophiles présentant des proportions faibles en carbone et en atomes lourds.

Concernant la promiscuité des ligands, la tendance opposée a été mise en évidence. Seulement 18 % des clusters de ligands sont promiscuous dans la MOAD. La sélectivité des ligands est en partie due au fait qu'ils ne sont pas testés contre toutes les cibles possibles et au fait que MOAD, bien que de haute qualité, est une base de données fragmentée.

Cette analyse confirme les nouvelles tendances observées lors de l'étude de la promiscuité : le modèle d'interaction « clé-serrure » n'est plus la norme et ne représente donc pas la majorité des interactions. L'étude met en valeur l'importance grandissante de la prise en compte des sites de liaison, de leur caractérisation et de leur promiscuité dans la polypharmacologie.

Ceci est conforme aux conclusion de Meyers et al, (Meyers et al. 2018) qui conclut que l'analyse de la similarité des sites de liaison n'expliquait que partiellement la promiscuité des ligands se liant à des protéines de différentes familles.

Cette étude précède la prédiction des interactions. Elle permet, avec du recul, de proposer des hypothèses pour la réutilisation de médicaments déjà existants et ainsi mieux détecter les possibles interactions *« off-target »* engendrées par la promiscuité et engendrant les effets secondaires néfastes, dans le contexte du développement de médicaments. Il a été montré que

l'utilisation des approches computationnelles est bénéfique pour les études de la polypharmacologie et des interactions mais que les mécanismes qui les régissent sont très peu connus (Chaudhari et al. 2017).

Article n°3 :

Cerisier N, Petitjean M, Regad L, Bayard Q, Réau M, Badel A and Camproux A-C, High Impact: The Role of Promiscuous Binding Site in Polypharmacology, *Molecules*. 2019 Jul 10;24(14). pii: E2529. doi: 10.3390/molecules24142529.

molecules

MDPI

Article

High Impact: The Role of Promiscuous Binding Sites in Polypharmacology

Natacha Cerisier [1], Michel Petitjean [1], Leslie Regad [1], Quentin Bayard [2], Manon Réau [3], Anne Badel [1] and Anne-Claude Camproux [1,*]

1 Université de Paris, Biologie Fonctionnelle et Adaptative, UMR 8251, CNRS, ERL U1133, INSERM, Computational Modeling of Protein Ligand Interactions, F-75013 Paris, France
2 Centre de Recherche des Cordeliers, Sorbonne Universités, INSERM, USPC, Université Paris Descartes, Université Paris Diderot, Université Paris 13, Functional Genomics of Solid Tumors Laboratory, F-75006 Paris, France
3 Laboratoire Génomique Bioinformatique et Chimie Moléculaire, EA 7528, Conservatoire National des Arts et Métiers, F-75003 Paris, France
* Correspondence: anne-claude.camproux@univ-paris-diderot.fr; Tel.: +33-(0)1-57-27-83-77

Academic Editor: J.B. Brown
Received: 3 June 2019; Accepted: 27 June 2019; Published: 10 July 2019

Abstract: The literature focuses on drug promiscuity, which is a drug's ability to bind to several targets, because it plays an essential role in polypharmacology. However, little work has been completed regarding binding site promiscuity, even though its properties are now recognized among the key factors that impact drug promiscuity. Here, we quantified and characterized the promiscuity of druggable binding sites from protein-ligand complexes in the high quality Mother Of All Databases while using statistical methods. Most of the sites (80%) exhibited promiscuity, irrespective of the protein class. Nearly half were highly promiscuous and able to interact with various types of ligands. The corresponding pockets were rather large and hydrophobic, with high sulfur atom and aliphatic residue frequencies, but few side chain atoms. Consequently, their interacting ligands can be large, rigid, and weakly hydrophilic. The selective sites that interacted with one ligand type presented less favorable pocket properties for establishing ligand contacts. Thus, their ligands were highly adaptable, small, and hydrophilic. In the dataset, the promiscuity of the site rather than the drug mainly explains the multiple interactions between the drug and target, as most ligand types are dedicated to one site. This underlines the essential contribution of binding site promiscuity to drug promiscuity between different protein classes.

Keywords: proteochemometrics; multi-targets; drug promiscuity; druggable binding site; descriptors; polypharmacology

Academic Editor: J.B. Brown

1. Introduction

Since the late-1960s, drug development processes have been based on the "one drug, one target" paradigm. The primary goal of drug discovery has been the design and delivery of selective compounds against individual biological targets [1]. A drug may be involved in different disease functions and might interact with more than one target, which is defined as drug promiscuity [1–3]. These multi-target drugs are commonly referred to as promiscuous drugs and they are the origin of polypharmacology, where drugs bind to more than one protein target at concentrations that are relevant to their therapeutic free exposure [4]. Drug promiscuity is accepted as a general phenomenon, with the growing number of compounds associated with multiple target annotations [5]. Jalencas et al. [6] reported that only 15% of drugs are known to interact with a single target, but over 50% interact with more than five targets;

thus, drug promiscuity is a key element in drug discovery and development. Its consequences can be either beneficial or undesirable. Amongst the beneficial outcomes, the drug may be applicable against new diseases, avoiding time and money being spent on preclinical tests or repurposing [7,8]. Amongst the undesirable outcomes, promiscuous drugs can interact with off-targets, which results in adverse drug reactions, harmful side effects, and adverse polypharmacology [1,7,9].

Consequently, drug promiscuity has received considerable attention. It has been analyzed while using the drug physicochemical properties, fragment composition, the protein family to which they bind, and the binding site similarities by taking advantage of the increasing number of three-dimensional structures in a complexed form. The protein binding site similarity confers similar binding interactions with a common ligand. For instance, Haupt et al. [2] demonstrated that the global structure and the binding site similarity within the structures of the Protein Data Bank (PDB) [10] had greater influence on drug promiscuity than physicochemical drug properties, such as hydrophobicity, molecular weight, or flexibility.

The field of binding site comparison has emerged as a powerful approach for investigating drug promiscuity. Naderi et al. [11] surveyed 12 widely used computational tools to match the pockets and underlined the important pharmacological applications of pocket matching. For related targets, small molecules often bind with a high affinity to multiple members of a protein family due to the high sequence similarity in its active sites [2,12]. Duran-Frigola et al. [13] and Meyers et al. [14] confirmed that binding site pocket similarity analysis is a useful tool in promiscuity detection within the same family, being named intrafamily promiscuity. However, the performance of the binding site comparison has not been demonstrated concerning interfamily promiscuity detection, when the promiscuous drug acts across different target families. Interfamily promiscuity is rarely observed; only ~2% of bioactive compounds are considered to be promiscuous across different unique target families on the basis of high-confidence activity data [15]. Their importance should not be neglected for drug repurposing, the prediction of side effects and drug–target interactions, leading optimization, and prioritizing the study of secondary targets for their importance in efficacy or toxicity [14,16]. Similar pockets may occur within unrelated protein structures [17] and one-third of similar cavities from experimental human protein structures can be found among the proteins with no apparent relationship, as related in [13]. They identified 181,500 pairs of similar cavities: 68.8% corresponded to pockets in different structural instances of the same protein and 31.2% to cavities in distinct proteins. High-affinity selective interactions can occur for a ligand when binding to different protein families, despite the absence of an apparent binding site or sequence similarity [18,19]. An analysis by Barelier et al. [20] concerning ligands that are able to bind proteins from different families showed that the majority of examples of binding sites in unrelated proteins were not a result of matched residues in the ligand binding site. They concluded that there was no single pattern-matching "code" to identify the binding sites that bound identical ligands in unrelated proteins. Meyers et al. [14] concluded that the pocket similarity analysis performance for the promiscuity detection has not been demonstrated to be relevant when searching for proteins outside the family of the target protein in some chemotype cases. They highlighted the importance of understanding the factors that control the underlying similarity in protein binding sites. Numerous studies have confirmed that the plasticity of the binding site is important in protein–ligand interactions [21–23]. Gao et al. [24] concluded that more than one-third of their representative pockets from the PDB were promiscuous and they interacted with multiple different ligands. Hu et al. [25] analyzed both the compound and target promiscuity and the target's ability to interact with an increasing number of structurally diverse compounds and concluded that most of the targets bind to varying numbers of promiscuous compounds.

To the best of our knowledge, no exhaustive work has been dedicated to quantifying and characterizing the structural features of target binding sites to understand their propensity to be promiscuous versus selective (that is, able to interact with only one type of ligand). It is possible to explore the target binding site space and multiple binding site–ligand interactions while taking advantage of the increasing number of three-dimensional (3D) structures and their redundancy. In

this study, we used the interaction database, Mother Of All Databases (MOAD), which is one of the largest databases that provides more than 25,000 protein–ligand complexes with only high-quality resolution structures extracted from the PDB that are suitable for studying the promiscuity of binding sites [26]. As druggability is crucial in drug design [4,27], we focused on druggable binding sites (DBS). A DBS is defined by its corresponding pocket set estimated by proximity to a drug-like ligand from homologous PDB chain clusters (chain with 90% sequence identity) in the MOAD. Here, we investigated the promiscuity of DBS that were observed several times in the MOAD (at least four pockets) and their frequency. Next, we studied the characteristics of pockets and ligands corresponding to the DBS of different promiscuity in terms of their geometrical and physicochemical properties and their correspondence. Subsequently, we studied the DBS promiscuity frequency within different MOAD protein classes and their contribution to explain the drug promiscuity between the MOAD protein classes. A network study allowed for the visualization of the multiple connections between DBS and different ligands and confirmed the high impact of DBS promiscuity on these multiple connections. This work demonstrates the importance of DBS promiscuity analysis for understanding and modeling multiple protein–ligand interactions, drug promiscuity, polypharmacology, drug repositioning, or off-target detection.

2. Results

2.1. MOAD Druggable Binding Site Identification

2.1.1. MOAD Protein and Ligand Space

The MOAD provides 25,769 high-resolution protein–ligand complexed structures, which include 12,440 unique ligands and a total of 44,675 protein–ligand interactions. First, we extracted 10,500 different valid ligands (see Section 4) from 9,556 PDB complex structures that were involved in 18,317 pocket–ligand interactions (Figure 1). Selected molecules were filtered based on their drug-likeliness, and 5,824 PDB complex structures containing 4,058 valid and drug-like ligands were retained. From these PDB files, we aligned 10,271 protein mono-chains while using TM-Align, clustering them into 1,137 homologous chain clusters using H-CD-HIT. The final filtered dataset included 8,669 protein–ligand interactions.

Next, similar ligands were clustered while using a Tanimoto coefficient threshold of 0.8 in the "Ligand–Cluster". The 4,058 ligands were clustered into 2,306 Ligand–Clusters, which indicated that more than half of the ligands (56.8%) were diverse. The average Tanimoto coefficient between the pairs of representative ligands from all Ligand–Clusters (see Section 4) was weak at 0.32 (±0.13), confirming that the MOAD includes large valid and drug-like ligand diversity. On average, these 2,306 Ligand–Clusters included 1.75 (±2.23) similar ligands. Most of them (71%) were only observed once in the MOAD, but some were also highly observed in the MOAD. For example, the largest Ligand–Cluster included 46 similar ligands (meaning different Het molecules as defined in the PDB but with a Tanimoto > 0.8) that corresponded to a derivate of nucleosides.

Thus, the MOAD provides a number of proteins, ligands, and protein–ligand interactions that are sufficiently diverse to study the promiscuity of DBS obtained from homologous chain clusters in different MOAD protein classes.

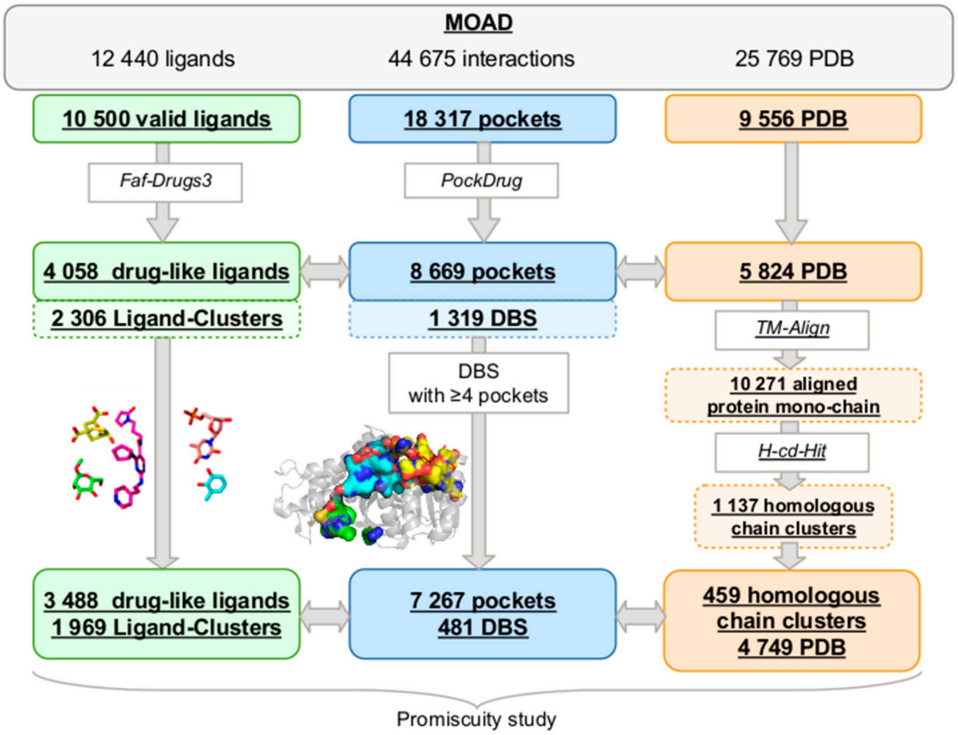


Figure 1. The protocol developed to extract Druggable Binding Sites (DBS) from the Mother Of All Databases (MOAD) data and to determine their promiscuity. We extracted 3,488 valid and drug-like ligands that were regrouped into 1,969 clusters while using the Tanimoto coefficient (green), 459 homologous chain clusters among the 4,749 Protein Data Bank (PDB) files selected using TM-Align [28] to calculate the identity between all the protein chains, and H-CD-HIT [29] clustering algorithms to group the homologous protein chains (orange). The 7,267 pockets clustered in 481 druggable binding sites (DBS) (blue) correspond to the ligands previously selected and all of the homologous protein chains. The promiscuity was calculated from these data.

2.1.2. Druggable Binding Site Extraction

From the selected protein–ligand interactions, 8,669 binding pockets were estimated by proximity to the 4,058 drug-like ligands while using a threshold of 5.5 Å (Figure 1). We applied our Pocket-Clustering algorithm to these pockets to identify those describing an identical druggable binding site (see Section 4). It located 8,669 binding pockets from the 1,137 different homologous chain clusters and clustered them into 1,319 sets of overlapping pockets, named the "Pocket-Cluster", which described 1,319 distinct DBS. These 1,319 DBS are described, on average, by 6.57 (±13.43) pockets, with a minimum of 1 and a maximum of 164 pockets per DBS. The average score of overlap within the Pocket-Cluster was 0.65 (±0.13), with a minimum of 0.25 and a maximum at 0.95. This minimum confirmed that our Pocket-Clustering algorithm split the non-overlapping pockets that corresponded to disconnected regions of a binding site in a different Pocket-Cluster, which is coherent for studying the DBS promiscuity.

To determine the promiscuity of a DBS, we had to quantify its number of different Ligand–Clusters in interaction with its Pocket-Cluster. We defined a DBS in interaction with only one Ligand–Cluster as a "selective DBS". A qualification of DBS selectivity can be due to the limited availability of three-dimensional (3D) crystal structures complexed in the MOAD. One-third of obtained DBS (35%) were observed only once in the MOAD (and described by a unique binding pocket); consequently, they cannot be technically characterized in terms of promiscuity. The frequency of selective DBS directly decreased with DBS occurrence in the MOAD (results not shown). Thus, the higher the occurrence of a DBS in the MOAD, the higher the possibility of observing its interactions with different Ligand–Clusters, and the higher the reliability in its qualification of promiscuous or selective DBS. We selected a compromise between having enough occurrences of each DBS (important enough to

limit the characterization of DBS as "selective" when not true) to analyze the promiscuity of each DBS, but also to retain a high number of different DBS to obtain reliable promiscuity information from the MOAD. In this study, we chose to analyze the promiscuity of the set of 481 DBS observed four or more times, which is referred to as the DBS4 dataset (the complete dataset is available at the following DOI: 10.6084/m9.figshare.8313185). These corresponded to 37% of the DBS that were described by 7,267 pocket–ligand interactions from 459 homologous chain clusters and a large part of the binding pockets (83%) in interactions with 86% of the ligands. These 481 DBS interacted with 3,488 ligands, which were clustered into 1,969 Ligand–Clusters. These DBS are described by 15.1 (±19.5) pockets, on average, with a Pocket-Cluster size between four and 164 pockets and interacted with 8.0 (±14.0) ligands.

2.2. DBS Promiscuity Characterization

2.2.1. Druggable Binding Site Promiscuity Quantification

Complementary to selective DBS, we defined a DBS interacting with several Ligand–Clusters as a "promiscuous DBS". We quantified the DBS interacting with only one type of ligand versus several types of ligands in the DBS4 dataset to determine the promiscuity of the DBS. The DBS was associated with 5.6 (±8.9) Ligand–Clusters on average, from 1 to 90 Ligand–Clusters (Figure 2). Almost one-quarter of the DBS (21%, 100/481) were selective, whereas 79% were promiscuous DBS (Table 1). Amongst the promiscuous DBS, we distinguished 34% (166) as moderately promiscuous (MP) DBS, which were associated with two or three Ligand–Clusters, from 45% (215) highly promiscuous (HP) DBS, which were associated with four or more Ligand–Clusters. The most promiscuous DBS corresponded to the Human Cyclin-dependent kinase 2 (CDK2) transferase, which has been well studied due to its role in the regulation of the acid–base balance in the organism [30]. Its high pocket (164) and Ligand–Cluster (90) numbers were consistent with the high occurrence of complexed structures, the protein high sequence homology of CDK families, and the numerous inhibitors that were developed against CDK2 [31].

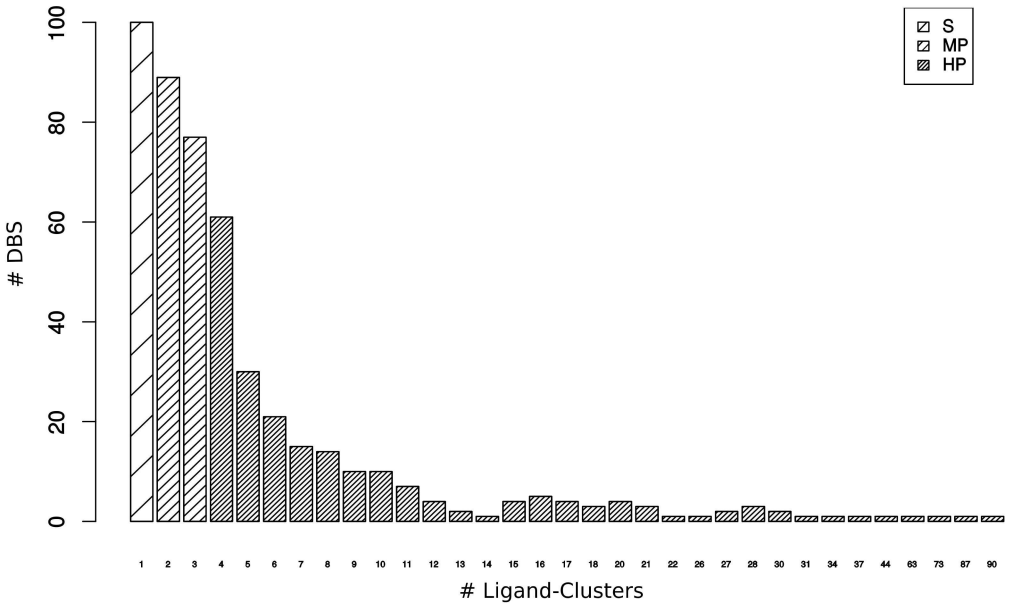


Figure 2. Number of DBS interacting with 1 to 90 Ligand–Clusters from the DBS4 dataset. Among the 481 DBS, 100 were selective (S), 166 were moderately promiscuous (MP), and 215 were highly promiscuous (HP).

Table 1. The number (and proportion) of DBS in the DBS4 dataset according to their promiscuity level and correspondence with the number (and proportion) of binding pockets: selective (S, in interaction with one Ligand–Cluster), moderately promiscuous (MP, two or three Ligand–Clusters), and highly promiscuous (HP, four or more Ligand–Clusters).

	DBS		Pocket	
S	**100**	(20.8%)	791	(10.9%)
MP	166	(34.5%)	1447	(19.9%)
HP	215	(44.7%)	5029	(69.2%)
Total	481	(100.0%)	7267	(100.0%)

Eight and seven pockets, on average, describe Selective and MP DBS, respectively, whereas 23 pockets describe HP DBS. The 45% HP DBS corresponded to 69% (5,029) of the 7,267 pockets (Table 1). However, it was difficult to interpret whether these high pocket occurrences of HP DBS are due to these DBS that belong to targets of therapeutic interest and they are highly crystallized (which contributes to revealing their high promiscuity), or a consequence of their known promiscuity, which explains that they have been intensively studied and crystallized with different ligands. As expected, the 100 selective DBS that interacted with only one Ligand–Cluster each corresponded to a small part of the Ligand–Clusters (less than 4%). The 166 MP DBS interacted with 15.3% of the Ligand–Clusters. The 215 HP DBS corresponded to 89% of the Ligand–Clusters (and 5,029 pockets), according to their high ability to interact with various types of ligands, which resulted in 7.31 Ligand–Clusters on average per DBS. Thus, we conclude that DBS promiscuity is a common phenomenon in proteins.

2.2.2. Binding Pocket Characteristics of DBS with Different Promiscuity Levels

We studied the physicochemical and geometrical pocket properties that discriminated the DBS of the three different levels of promiscuity. To do so, 72 physicochemical and geometrical pocket descriptors described each pocket. First, the results of ANOVA tests when comparing the pocket descriptors that were associated with the DBS of different promiscuity levels showed many significant differences (68/72 descriptors, Table S1). The most different descriptors corresponded to geometry, frequency of certain atoms, or residues and hydrophobicity. Some of them (with a p value $< 2 \times 10^{-5}$) are illustrated on a boxplot (Figure 3A). In terms of geometrical descriptors, the volume and number of residues increased with DBS promiscuity, which indicated that selective DBS correspond to rather small pockets than the HP DBS pockets (16.6 ±4.3 versus 20.1 ±5.1 in the number of residues). Selective DBS also corresponded to significantly smaller Pocket Convexity Index (PCI) values (Table S1), which thus supports more convex pockets. In terms of physicochemical descriptors, the proportion of side chain atoms was noticeably weaker and the proportion of sulfur atoms was significantly higher for pockets that are associated with HP relative to selective DBS. Sulfur atoms are known to form hydrogen bonds that are essential in many enzymatic reactions and they should contribute to DBS promiscuity [32]. The HP DBS pockets exhibited higher values of hydrophobicity (based on the hydrophobicity property of residue) for HP than the selective or MP DBS pockets. The frequency of aliphatic residues increased, while the frequency of oxygen atoms in tyrosine residue atoms (Otyr) decreased for the HP DBS pockets. The MP DBS pockets tended to present intermediate values between the selective and HP DBS pockets, except in some descriptors, such as the high value of frequency of nitrogen atoms in tryptophan residue atoms (Ntrp) and charged residues (Table S1). Pockets were associated with a high average druggability score in terms of the druggability score (78% ±0.26), which was in agreement with the drug-like ligand selection. The HP DBS tended to be the most druggable (81%), which agrees with the druggability score increasing with the size and hydrophobicity of the pockets [33]. The relatively small size of the selective pockets appeared to be compensated for by physicochemical properties, such as high Otyr atom frequency (which characterizes a hydrogen bond donor group by a hydroxyl group in the tyrosine side chain), which plays a key role in binding the drug-like molecules to explain their druggability score [34].

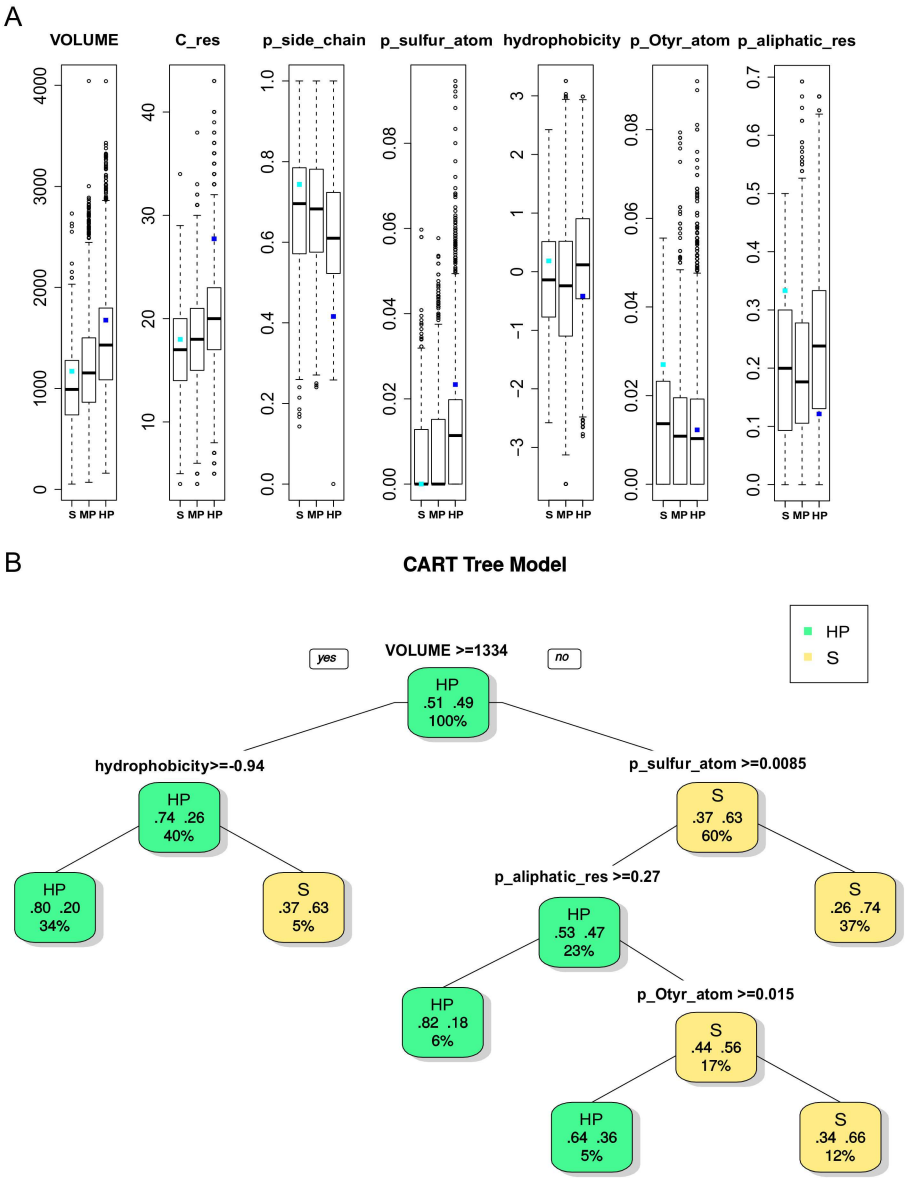


Figure 3. (A) Boxplot of seven pocket properties that differentiated DBS with different levels of promiscuity (p value < 2×10–5). For each boxplot, the black bar in the rectangle represents the medium, the upper side is the third quartile, and the lower side is the first quartile of data. The dashed lines represent values below the first quartile limit and values above the third quartile limit. Circles represent atypical values (determined by the boxplot function). The cyan dot and the blue dot represent the values of the ligand in interaction with a lyase selective DBS (PDB 4KVI) and an urokinase HP DBS (PDB 1F5K), respectively, as illustrated in Figure 5. (B) Representation of the Classification And Regression Tree (CART) for pockets based on the pocket descriptors between the S (yellow) and HP (green) pockets. Each node is described by the class (promiscuity), the probability per class (right = HP and left = S), and the percentage of observations in the node. The CART model performances obtained for distinguishing pockets associated with the selective DBS from the HP DBS resulted in an average accuracy, sensitivity, and selectivity of 68%, 75%, and 67% for the test set (1,252 pockets), respectively, and 64% ±2%, 60% ±5%, and 64% ±3% while using a five-fold cross validation on the training set (500 runs), respectively.

Second, we illustrated the pocket properties that discriminated the two most informative classes of DBS promiscuity: selective and HP DBS, while using the Cart And Regression Tree (CART) method. The CART model performance for distinguishing pockets that are associated with selective from HP DBS resulted in an average accuracy, sensitivity, and selectivity of 68%, 75%, and 67%, respectively, for

the test set. Figure 3B presents a tree that illustrates the optimal combinations of pocket descriptors that discriminate HP from selective DBS, where different combinations explain different DBS promiscuity. Aligned with the ANOVA results, the first discriminating pocket property was the volume, and then different physicochemical properties were involved. HP DBS mainly corresponded to the largest pockets, except for those that were weakly hydrophobic. However, a few small pockets with high sulfur, aliphatic, or Otyr atom frequency were associated with HP DBS. Conversely, selective DBS were mainly associated with small pockets (78.6%) with weak sulfur atom frequency; otherwise, they had weak aliphatic residue and Otyr atom frequency.

2.2.3. Ligand Characteristics interacting with DBS with Different Promiscuity Levels

Of the Ligand–Clusters, 93.6% were associated with DBS presenting the same promiscuity level. Most of them (82.8%) only interacted with HP DBS according to the ability of these DBS to interact with a high number of Ligand–Clusters. We observed that these Ligand–Clusters interacting with DBS of the same level of promiscuity exhibited high diversity in terms of the Tanimoto coefficient (0.33), with the lowest Tanimoto value (0.24) for those that only interacted with selective DBS (Table 2). The ability of these Ligand–Clusters to interact with only the DBS of the same promiscuity level suggests that their ligand properties could correspond to the DBS pocket tendencies of corresponding promiscuity level. First, we characterized the ligands that were only clustered within these 93.6% Ligand–Clusters dedicated to selective or MP or HP DBS. The results of the ANOVA tests showed significant differences between all 21 ligand descriptors (Table S2), where seven are illustrated on a boxplot in Figure 4A. The ligand ability to bind to selective DBS decreased with its number of rings (Rings) and rigid bonds (RigidB) values, the carbon and heavy atom frequency, the molecular weight, and the lipophilicity values. Ligands that only interact with selective DBS were more flexible, as they exhibited a lower number of rings and RigidB values (the more important the number of rings and the value of RigidB, the less flexible the molecule). They tended to be lower in molecular weight, with less carbon and heavy atoms, and exhibited a lower value of lipophilicity (logD and logP), which means that they are more hydrophilic. The ratio between hydrogen and carbon atoms (ratioH/C) was particularly variable for selective DBS ligands, but it tended to be small for HP DBS ligands. Ligands that were bound to MP DBS mainly presented intermediate values between those that were bound to selective or HP DBS. Second, we studied the ligand properties that discriminated the Ligand–Clusters interacting with selective and HP DBS while using the CART model. The CART model performance produced an average accuracy, sensitivity, and specificity of 77%, 67%, and 77% for the test set, respectively. The tree that was obtained for the 53 ligands from the Ligand–Clusters interacting with selective DBS and the 2,621 ligands from those interacting with HP DBS is illustrated in Figure 4B, which confirmed that flexibility is the most important property for discriminating ligands that bind to HP versus selective DBS. Ligands with few rings or with low lipophilicity, or with a low ratio between the hydrogen and carbon atom values, mainly interacted with selective DBS. Thus, we conclude that ligands from Ligand–Clusters that were dedicated to selective DBS tend to be smaller, less rigid, with less carbon and heavy atoms, and tend to be more hydrophilic. This is coherent with their interaction with pockets that are difficult to bind (given their physicochemical and geometrical properties).

Only 6.6% of the 1,969 Ligand–Clusters interacted with DBS of different promiscuity levels. These we called the "mixed" Ligand–Clusters, and they corresponded to 16% of the 3,488 ligands (Table 2). They were split into 39 (2%) Ligand–Clusters interacting with 39 both selective and some promiscuous DBS, and 88 (4.4%) interacted with both MP and HP DBS. ANOVA tests between 2,931 ligands from the dedicated Ligand–Clusters versus 557 ligands from the mixed Ligand–Clusters showed that ligands were able to interact with DBS presenting different promiscuity (mixed Ligand–Clusters) and they tended to be more hydrophilic (low logP values) with a low number of rotatable bonds (RototableB) (Table S3).

Table 2. Description of the Ligand–Clusters according to the different promiscuity levels using the DBS4 dataset. Ligands associated with selective (S), MP, and HP DBS are considered as dedicated, and those associated with DBS of different promiscuity levels are considered mixed. Second column gives the occurrence (and frequency) of Ligand–Clusters according to the promiscuity of DBS to which they bind: 29 were S, 182 were MP, the majority (1,631) were HP, and 127 were mixed, where they bind to DBS of different promiscuities. From the 127 mixed Ligand–Clusters, eight bound to selective and HP DBS; 26 bound to S, MP, and HP DBS; 88 bound to MP and HP DBS; and, five bound to selective and MP DBS. The third column is the average Tanimoto coefficient between the Ligand–Clusters. Ligands associated with S, MP, and HP were considered dedicated, and those associated with Pocket-Clusters with different promiscuity levels were considered as mixed. The last column is the number of corresponding ligands.

Ligand–Cluster		Occurrence (Frequency)		Tanimoto coefficient average (std. dev.)	Corresponding Ligands: Occurrence (Frequency)	
Dedicated to	**S DBS**	**29**	(1.5%)	0.24 (0.11)	53	(1.5%)
	MP DBS	182	(9.2%)	0.29 (0.12)	257	(7.4%)
	HP DBS	1631	(82.8%)	0.33 (0.13)	2621	(75.1%)
Mixed		127	(6.4%)	0.28 (0.13)	557	(16.0%)
ALL		1969	(100.0%)	0.33 (0.13)	3488	(100.0%)

2.2.4. Pocket and Ligand Property Correspondence

We observed significant physicochemical and geometrical tendencies for both pockets and ligands that are associated with DBS of different promiscuity and therefore studied their correspondences. Both the pockets and ligands from selective DBS were significantly smaller than those that were associated with HP DBS. This pocket–ligand size correspondence was expected, as the pocket was estimated by its proximity to its ligand.

Most selective DBS corresponded to small pockets with a high side chain atom frequency, a low proportion of sulfur, Otyr atoms, and aliphatic residues. A few of them were larger, but they presented low hydrophobicity values. These pocket properties are less favorable for establishing contact with various ligands. Accordingly, to be able to bind to these difficult pockets, their interacting ligands tend to be flexible and small with low carbon and heavy atom proportions and they are hydrophilic. Figure 5A illustrates one selective DBS, the aristolochene synthase from *Aspergillus terreus* from the lyase protein class. It is associated with 19 structure chains in the MOAD and five different ligands (clustered into only one Ligand–Cluster and corresponding to a Tanimoto coefficient of 0.88 ±0.07). As shown in Figures 3A and 4A, its properties matched well with the selective tendencies for both the pocket and ligands.

The HP DBS corresponded to pockets that were well adapted to bind several and diverse ligands—these were mainly large and hydrophobic. For relatively small pockets, they presented high frequencies of sulfur atoms and aliphatic residue (or Otyr atoms), but low side chain atom frequencies. These HP DBS pocket properties are favorable for establishing interaction with different ligands. Therefore, their interacting ligands do not need a particular flexibility and adaptability to bind. Accordingly, we observed that their ligands tended to be large, rigid, and have high carbon and heavy atom frequencies. They were also less hydrophilic than those that are associated with selective DBS. Figure 5C illustrates one HP DBS, which is the catalytic inhibitor-binding site from the human urokinase Plasminogen Activator target from the hydrolase protein class. This target has been extensively studied due to its role in cancer pathways and for its high promiscuity [35,36]. The corresponding urokinase DBS presented pocket and ligand property values that matched well with the HP tendencies, as shown in Figure 3A, Figure 4A. The 22 corresponding pockets in the MOAD were large and hydrophobic, with a small proportion of side chain atoms and a high proportion of sulfur atoms. The ligands from its 17 associated Ligand–Clusters had high rigid bonds values, meaning that they were more rigid.

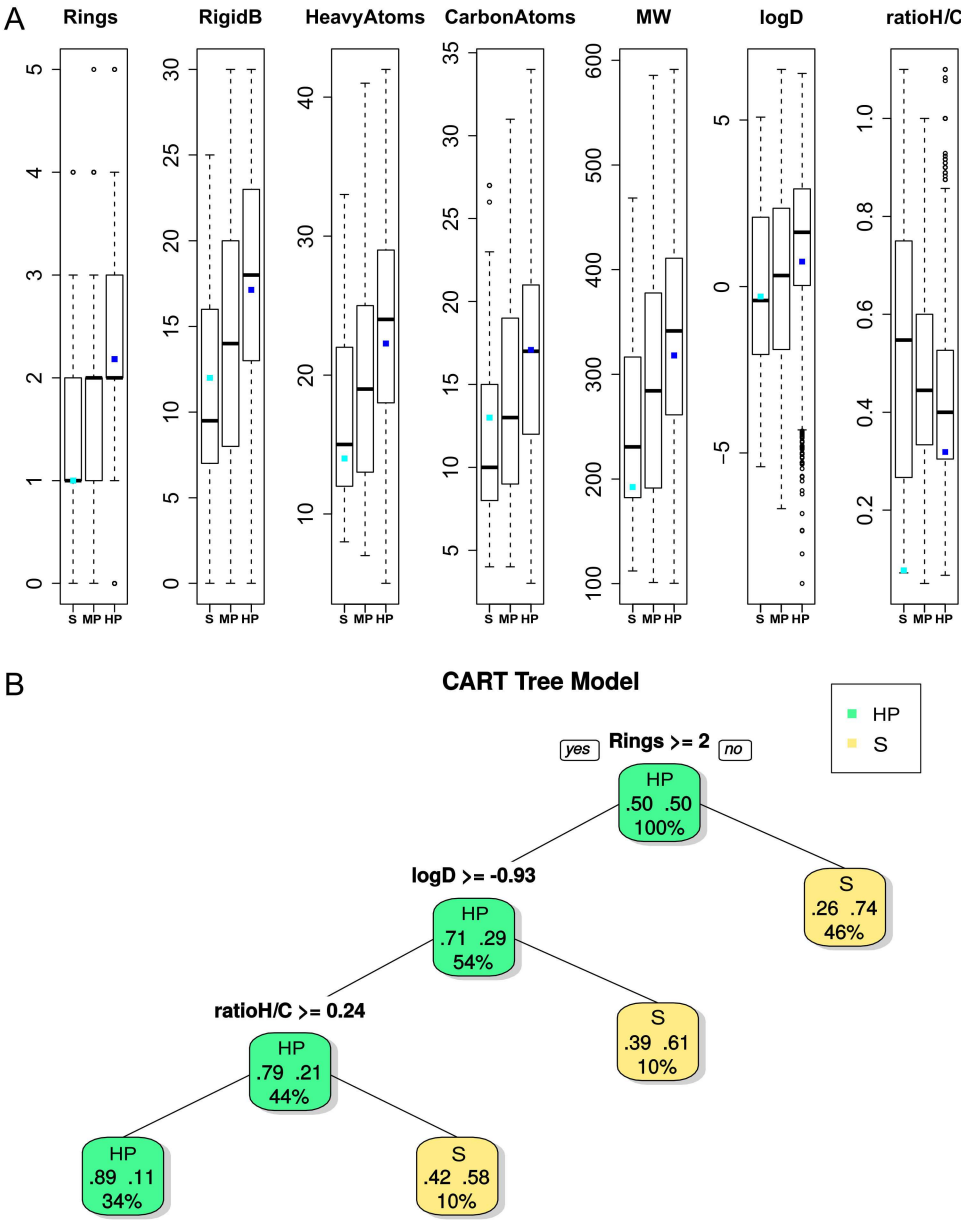


Figure 4. (**A**) Boxplot of seven ligand properties (p-value < 3×10^{-10} except for the ratio between hydrogen and carbon atoms called ratioH/C, p-value = 0.1) from ligands associated with the Ligand–Cluster dedicated to S or HP DBS. For each boxplot, the black bar in the rectangle represents the medium, and the upper side is the third quartile and the lower side is the first quartile of data. The dashed lines represent values below the first quartile limit and values above the third quartile limit. Circles represent atypical values; the cyan dot and the blue dot represent the values of the ligand in interaction with a lyase selective DBS (PDB 4KVI) and an Urokinase HP DBS (PDB 1F5K), respectively, as illustrated in Figure 5. (**B**) Representation of the CART tree for ligands based on ligand descriptors between S (yellow) and HP (green) ligands. Each node is described by the class (promiscuity), the probability per class (right = HP and left = S), and the percentage of observations in the node. The CART model provided an average accuracy, sensitivity, and specificity of 77%, 67%, and 77% for the test set (53 versus 2,621 ligands), respectively, and 72% ±3%, 69% ±7%, and 73% ±3% while using a five-fold cross validation on the training set (500 runs), respectively.

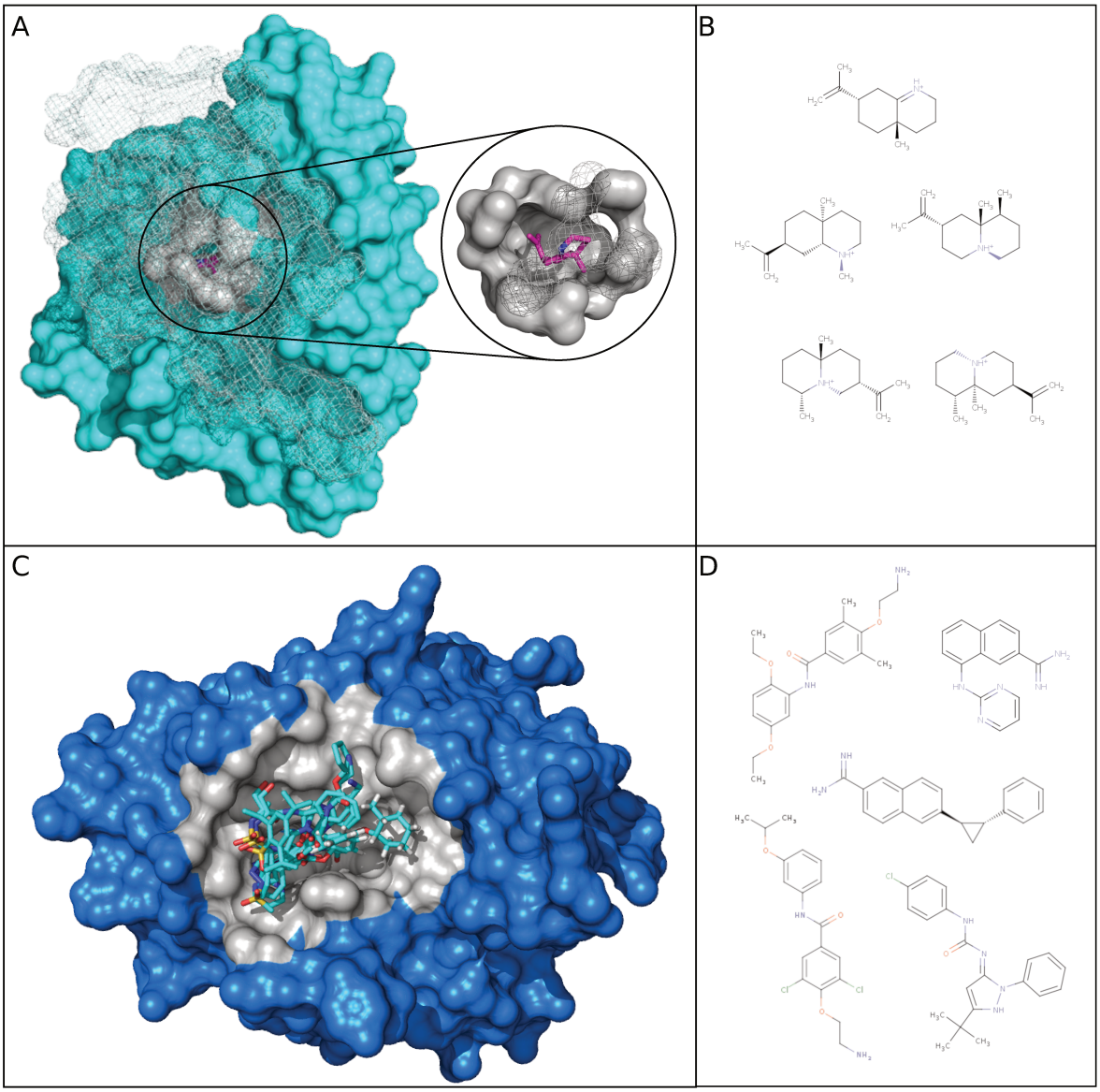


Figure 5. Illustration of one selective and one HP DBS and some of their associated ligands. (**A**) The aristolochene synthase protein with a selective DBS is represented in cyan with its representative structure: a lyase from *Aspergillus terreus* (PDB 4KVI, Pocket-Cluster no. 141_1). The grey part (in the surface and mesh) is the pocket that interacts with the ligand, where the carbon is represented by pink and the nitrogen in blue (PDB 1SV, Ligand–Cluster 300). Part of the protein and pocket are represented in mesh instead of the surface to facilitate the visualization of the pocket buried in the three-dimensional (3D) protein. (**B**) Five ligand two-dimensional (2D) structures (Tanimoto > 0.8) from the Ligand–Cluster associated with the selective DBS. (**C**) The human urokinase plasminogen activator protein with a HP DBS is represented in blue with its representative structure: a hydrolase from *Homo sapiens* (PDB 1F5K, Pocket-Cluster no. 94_2). Its pockets are represented in grey and five representatives of the 17 different Ligand–Clusters that can bind to the DBS (carbon in cyan, nitrogen in blue, oxygen in red, and sulfur in orange). (**D**) Five ligand structures in two-dimensional (2D) (Tanimoto < 0.8) among the 17 different Ligand–Clusters that can bind to the HP DBS. Structures were visualized using PyMOL [37].

2.3. DBS Promiscuity Contribution to Multiple Interactions of Ligands with Different MOAD Protein Classes

2.3.1. DBS Promiscuity Frequency Related to MOAD Protein Classes

Another question of interest was: "Can DBS of different promiscuous levels be observed within all MOAD protein classes or are they dependent on the protein class?"

We observed that the 481 DBS were distributed among the 17 MOAD protein classes (Figure 6). The three most frequently observed protein classes were the enzymes transferase, hydrolase, and oxidoreductase, which together composed 71.9% of the DBS. We observed a high promiscuity with 4.64 (±2.49) Ligand–Clusters per DBS, on average, per protein class. For instance, there were 1.2 Ligand–Clusters per DBS on average for the binding protein class and 5.9 for the lyase protein class. We observed different promiscuous levels of DBS within different MOAD protein classes, independent of their frequency. The three promiscuous levels of DBS were observed in the nine most observed protein classes. Selective DBS and HP DBS were observed in the 10 and 15 most frequent MOAD protein classes, respectively. Moreover, there is a wide diversity of Ligand–Clusters that are associated with each HP DBS of different MOAD protein classes. The Tanimoto coefficient of Ligand–Clusters that was associated with the HP DBS of different MOAD protein classes was, on average, 0.44 (±0.10) per DBS per protein class, with a minimum value of 0.34 (±0.11) for the oxydoreductase and a maximum of 0.52 (±0.09) for the isomerase HP DBS (Table S4). This confirms that DBS with different promiscuous levels can be observed in diverse protein classes with the majority of HP DBS that interact with diverse ligands.

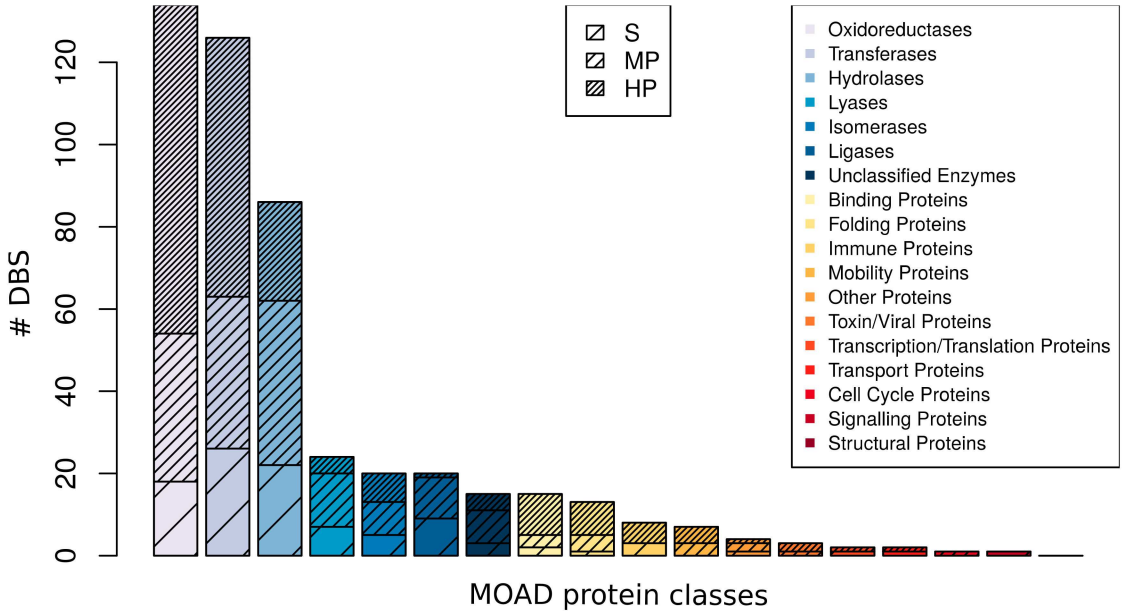


Figure 6. Distribution of the occurrence of DBS observed in the MOAD protein classes and their level of promiscuity in the DBS4 dataset. For instance, the lyase family corresponded to 20 DBS, with 5 S, 8 MP, and 7 HP DBS.

2.3.2. Complementary Study of the Ligand–Cluster Promiscuity

Subsequently, we studied the contribution of the DBS promiscuity and the Ligand–Cluster promiscuity on the multiple interactions between the different protein classes. First, we conducted a complementary study of the promiscuity of the Ligand–Clusters in the DBS4 dataset. We observed a limited Ligand–Cluster promiscuity: 82.2% were selective (in interaction with only one DBS) and 17.8% were promiscuous. This weak rate of promiscuous Ligand–Clusters can be partly explained by 71% (1,407/1,969) of the Ligand–Clusters being composed of only one ligand (where 59% (832/1,407)

only had one occurrence) in the DBS4 dataset. These promiscuous Ligand–Clusters were associated, on average, with 3.11 (±2.54) DBS extracted from 1.76 (±1.02) different MOAD protein classes.

We observed that most of the Ligand–Clusters were associated with DBS of the same promiscuous level. Thus, we studied the frequency of promiscuity of Ligand–Clusters relative to that of their interacting DBS. Only 18.9% the 1,753 Ligand–Clusters in interaction with at least one HP DBS were promiscuous; these interacted with 3.15 DBS, on average, corresponding to 1.76 MOAD protein classes. This could be due to HP DBS pocket properties being favorable for establishing interaction with different ligands and, consequently, their interacting ligands do not need to be particularly flexible, which does not support their ability to interact with several DBS. A total of 44.2% and 57.3% of Ligand–Clusters interacting with at least one MP DBS and at least one selective DBS were promiscuous, respectively (Table 3). This is coherent with their tendency to be flexible and small for interacting with more difficult pockets that are associated with selective DBS. These promiscuous Ligand–Clusters interacted with a high number of DBS: 4.15 DBS corresponding to 2.13 MOAD protein classes and 6.71 DBS corresponding to 2.95 MOAD protein classes, respectively.

Table 3. Distribution of the Ligand–Clusters relative to their ability to interact with DBS of different levels of promiscuity (rows). The last row, "All DBS", indicates the values of all the Ligand–Clusters in the DBS4 dataset. The columns indicate the total number of Ligand–Clusters, and those that interacted with one DBS (selective) or more (promiscuous). For the promiscuous Ligand–Clusters, we detailed the average number of DBS bound and the average number of protein class bound (not indicated for the selective Ligand–Clusters, as they bind to one DBS).

Ligand–Cluster	Total	Selective	Promiscuous	Promiscuous	Promiscuous
	Occurrence	Occurrence	Occurrence	Number of DBS: Average (sd)	Number of Protein Class: Average (sd)
Selective DBS	68	29	39	6.7 (5.37)	2.9 (1.45)
MP DBS	301	168	133	4.2 (3.68)	2.1 (1.32)
HP DBS	1753	1421	332	3.2 (2.60)	1.8 (1.02)
All DBS[1]	1969	1618	351	3.1 (2.54)	1.8 (1.02)

[1] The row "All DBS" is not the sum of the other rows because some Ligand–Cluster bind several types of DBS

Next, we analyzed the 68 Ligand–Clusters that interacted with selective DBS of different MOAD protein classes in further detail. These were split in Ligand–Clusters different in terms of promiscuity. We studied the characteristics and frequency of the 29 selective and the 39 promiscuous Ligand–Clusters. The 29 selective Ligand–Clusters interacted with 29 selective DBS that belonged to nine different MOAD protein classes. This indicated that only 1.4% (29/1969) of the Ligand–Clusters and 6% (29/481) of the DBS supported the "one drug, one target" concept in the DBS4 dataset. The 39 promiscuous Ligand–Clusters were highly promiscuous—these interacted with 6.71 DBS on average. The high promiscuity of these 39 Ligand–Clusters can be explained by their ligands having to be rather small and adaptable to be able to interact with selective DBS pockets. This was not the case for the 29 selective Ligand–Clusters interacting with selective DBS. Subsequently, we then explored the pocket and ligand properties discriminating the 29 selective Ligand–Clusters from the 39 promiscuous Ligand–Clusters while using an ANOVA test. These two types of Ligand–Clusters tended to be small and flexible. However, the selective DBS pockets tended tend to have less tiny and more aliphatic residues, and to be more spherical. The corresponding ligands were very flexible and they presented a weak proportion of weak ratio between hydrogen donor and acceptor (HBD/HBA) and values of topological Polar Surface Area (tPSA). This suggests that these pockets can be more buried and more difficult to bind, and that only dedicated ligands with adapted properties can interact with these pockets.

2.3.3. Ligand–Cluster-DBS Interaction Network Examples

We studied two examples of interaction networks that integrate certain DBS and the Ligand–Clusters to which they interacted.

Firstly, we visualized the 39 promiscuous Ligand–Clusters interacting with 71 selective DBS. These Ligand–Clusters additionally interacted with 142 promiscuous DBS (66 MP and 76 HP DBS) (Figure S5). This network resulted in 262 interactions between these 39 Ligand–Clusters and 213 DBS belonging to 14 different MOAD protein classes. The MOAD protein class and promiscuous level are indicated for each DBS. Some of these Ligand–Clusters are highly promiscuous. For instance, Ligand–Cluster numbers 1, 5, and 17 interacted with 27, 16, and 10 DBS, respectively, which belonged to 5, 7, and 4 different MOAD protein classes, respectively. Accordingly, the Ligand–Cluster numbers 1 and 5 assimilated to nucleotide derivatives or sugar, respectively, which are known to be well adapted to different binding sites. Ligand–Cluster number 17 clustered seven ligands, including triclosan (Drugbank ID: DB08604) and similar molecules. Triclosan is used as a preservative and antimicrobial agent in personal care products, cosmetics, kitchenware, toys, sports equipment, and footwear, and a review has reported that this molecule can have adverse effects on immune responses and cardiovascular functions [38]. Accordingly, we observed its high promiscuity and interaction with 10 DBS from four protein classes: oxidoreductases, transferases, transcription/translation, and transport proteins. These 39 mixed Ligand–Clusters interacted with both selective and promiscuous DBS. Finally, we observed 39 promiscuous Ligand–Clusters and 142 promiscuous DBS. This resulted in an interaction sub-network of 162 DBS sharing 24 Ligand–Clusters. This strong interconnection is mostly due to the 66% (142/213) of promiscuous DBS from different protein classes (Figure S5).

Secondly, we illustrated the impact of the DBS promiscuity by studying one protein class. We represented all of the Ligand–Clusters that interacted with the DBS from the lyase protein class and all of the DBS from other protein classes interacting with some of these Ligand–Clusters (Figure 7). This resulted in a network of 20 lyase DBS (five selective, eight MP, and seven HP DBS) and their 117 interacting Ligand–Clusters. These latter also interact with 84 DBS from 12 other protein classes, resulting in a total of 104 DBS and 220 interactions.

From the Ligand–Cluster point of view, we observed 85 (72%) dedicated and 32 (28%) promiscuous Ligand–Clusters according to the small portion of promiscuous Ligand–Clusters in the DBS4 dataset (Table 3). However, some of the promiscuous Ligand–Clusters were highly promiscuous, as illustrated by Ligand–Cluster number 1 in interaction with 27 DBS (Figure 7).

From the DBS point of view, only a small number (five) of the lyase DBS were not connected to another one DBS by some common Ligand–Clusters (two selective and three HP lyase DBS). Most of the lyase DBS shared at least one Ligand–Cluster with other DBS, and mainly with DBS from other protein classes. We observed one lyase DBS that was particularly promiscuous: DBS 1_1, corresponding to a human carbonic anhydrase protein, interacting with 73 Ligand–Clusters whose 55 were dedicated to this DBS. Mainly, we observed a high frequency of the promiscuous DBS (87.5% = 91/104) in this network. Consequently, a large sub-network, including most of the DBS (82.4%), were interconnected by at least one common Ligand–Cluster. This sub-network only included 27 promiscuous Ligand–Clusters, but 75 promiscuous DBS. This network illustrates the complexity of the interactions between DBS of different protein classes in terms of common Ligand–Cluster. It confirms that the high frequency of promiscuous DBS of different protein classes explained the interconnection in terms of Ligand–Cluster, even in this case of a weak frequency of promiscuous Ligand–Clusters.

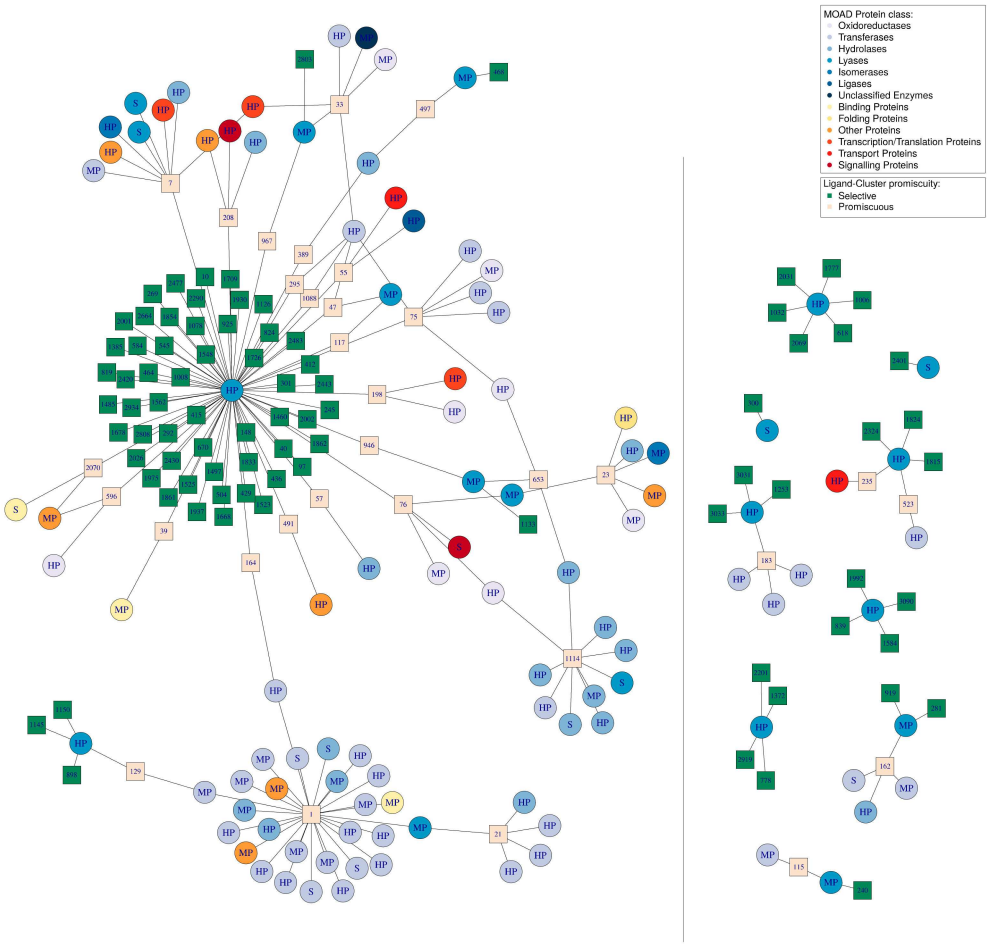


Figure 7. Network of the 20 DBS belonging to the lyase MOAD protein class in the DBS4 dataset, their 117 Ligand–Clusters, and the 84 other DBS from other protein classes that interact with these ligands, resulting in a total of 104 DBS and 220 interactions. Circles represent the DBS; these are colored according to the MOAD protein class to which they belong and are named according to their promiscuity (S for selective, MP for moderately promiscuous, and HP for highly promiscuous). The squares represent the Ligand–Clusters and are colored according to their level of promiscuity (beige for selective, binding to one DBS, or green for promiscuous) and their name is written inside the square. The grey line represents the boundary between the Pocket-Clusters and the Ligand–Clusters. This network visualization was created using the igraph package (see Section 4).

3. Discussion

In this study, we used the high-quality protein-ligand complex database, the MOAD, to study the promiscuity of the druggable binding sites. MOAD PDB chains were clustered according to a sequence identity threshold of 90% in a homologous chain cluster using H-CD-HIT to identify DBS promiscuity by taking advantage of the increasing protein redundancy information. The high quality of the MOAD's complexes is an advantage for characterizing binding site properties, but it can result in a limited number of available chains. Consequently, we studied the promiscuity of DBS bound to at least four drug-like valid ligands in the MOAD to reduce any false conclusions regarding selectivity due to the limited complex structures that are available. The resulting 481 DBS and 3,488 drug-like ligands, corresponding to 7,267 pocket-ligand interactions, confirmed that the MOAD can provide a valuable and large amount of data to study binding site promiscuity, for instance, through a comparison with datasets that Mestres et al. and Haupt et al. presented [2,9].

Our study highlights that promiscuous DBS are not an exceptional phenomenon, but they are common in most MOAD proteins. We mainly observed promiscuous DBS (80%) and a high frequency of HP DBS (45%) in interaction with at least four different Ligand–Clusters. This high occurrence of promiscuous DBS is in agreement with the findings by Gao and Skolnick (2013), who concluded that more than one-third of their representative pockets from the PDB were promiscuous and they interacted with multiple chemically different ligands [39]. However, these authors gathered 20,000 ligand-bound sites into only 1,000 representative pockets, which provided limited information regarding the promiscuity pocket characteristics.

Using the high-precision MOAD allowed for the characterization of both the pockets and ligands that are associated with DBS of different promiscuities in terms of their geometrical and physicochemical properties. Specifically, we tested for the detectable tendencies of the pockets that are associated with DBS of different promiscuity, regardless of the protein class and their impact on the ability to interact with diverse ligands. We tested for the detectable tendencies of the ligands in interaction with DBS of different promiscuity. We found that 20% of the DBS were selective and their pockets were less favorable for establishing contact: they tended to be small, have a high proportion of side chain atoms, a low proportion of sulfur atoms and aliphatic residues, or had a few that were large, but weakly hydrophobic. Therefore, selective DBS interacted with a small portion of the Ligand–Cluster space (4%), which corresponded to small, highly adaptable, and hydrophilic ligands that presented low carbon and heavy atom proportions. As expected, few of these selective DBS exemplify the old concept of "one drug, one target", as they belonged to nine different MOAD protein classes [40]. Complementarily, other selective DBS interacted with highly promiscuous Ligand–Clusters in interaction with, on average, nearly three MOAD protein classes. The high promiscuity of their Ligand–Clusters was in accordance with their ligand properties (high flexibility, small size, and presenting a particularly weak proportion of rotatable bonds), which not only allows them to interact with selective DBS, but also with various DBS.

The 45% of pockets that were associated with HP DBS mainly tended to be large and hydrophobic, or where a few of them were small, they presented high sulfur atom and aliphatic residue frequencies. These pocket properties were compatible with the flexibility of the DBS, which are known to contribute to promiscuity [22]. These pocket properties are expected to facilitate drug-like ligand interactions, as they are well adapted to bind to several and diverse ligands (their interacting Ligand–Clusters were associated, on average, with a Tanimoto coefficient of 0.44 (±0.10) per HP DBS). Accordingly, their ligands do not need a particular flexibility and they tended to be large with high carbon and heavy atom frequencies, and a few were hydrophilic.

The opposite promiscuity trend was observed in the MOAD for DBS and Ligand–Clusters: only 18% of Ligand–Clusters were promiscuous, whereas 80% of the DBS were promiscuous. The high frequency of selective Ligand–Clusters in the DBS4 dataset can be partly explained by the high frequency of ligands (24%) only observed once and by the clustering of homologous chains in one DBS, which thus reduces the number of DBS. However, this surprising opposite promiscuity tendency for DBS and Ligand–Clusters is in accordance with the analysis of compound and target promiscuity by Hu et al. [25]. These authors addressed the question of whether detectable tendencies for targets might exist to either recognize promiscuous or selective compounds, and how such tendencies might relate to the ability of targets to interact with increasing numbers of structurally diverse compounds. They concluded that the majority of compounds were only active against a single target, whereas most of the targets bound to varying numbers of promiscuous compounds. Next, they confirmed that less than 20% of their targets exclusively interacted with one selective compound, so the majority of the targets interacted with structurally diverse and promiscuous compounds [41]. This weak frequency of promiscuous Ligand that was observed both in our analysis and in a previous study [25] may be due to the fact that most of the ligands are not usually tested against other proteins beyond their target family, while proteins are usually tested against many different ligands, thus there is much more room for the observed promiscuity. Moreover, the known protein DBS universe is significantly

smaller than the universe of possible ligand scaffolds. Rifaioglu et al. [42] recently underlined that there are tens of millions of compounds that are available in compound and bioactivity databases, (about 9,000 FDA-approved small molecule drugs approved by experimental), while there are roughly 550,000 reviewed protein records available (20,244 of which are human proteins) in protein sequence and annotations resources (e.g., UniProtKB/Swiss-Prot). Therefore, it could be quite expected that ligands are observed to be more selective.

We conclude the existence of a high, but variable, promiscuity of DBS related to the MOAD protein classes. This is in accordance with the conclusion by Mestres at al. [9] that the promiscuity depends on the protein families. We observed different promiscuity levels of DBS within different protein classes: numerous HP DBS and a few selective ones were present in most of samples. Finally, the analyses of the Ligand–Clusters and DBS interactions confirmed that DBS are highly interconnected by common Ligand–Clusters, regardless of the protein class to which they belong. Whereas, only 20% of the Ligand–Clusters were promiscuous, we conclude that these high interaction numbers between the DBS of different protein classes and Ligand–Clusters are a consequence of the high promiscuity of the DBS. The weak frequency of selective Ligand–Cluster could be due to the choice of a high quality but sparse database, such as MOAD. Nevertheless, this highlights that the high promiscuity of a large number of DBS compensates for a low number of promiscuous Ligand–Clusters. This analysis confirms that the DBS promiscuity strongly contributes to multiple interactions between Ligand–Clusters and DBS belonging to different protein classes. This is in accordance with Meyers et al. [14], who concluded that pocket similarity analysis only partly explains the drug promiscuity across different target families.

Computational approaches have been demonstrated to be very beneficial and promising for polypharmacological studies, but also drug off-target studies, whose mechanisms are poorly understood or largely unknown, in most cases [43]. Our high DBS promiscuity frequency results demonstrate the importance of simultaneously integrating the DBS and the ligand promiscuity information in protocols of multiple drug–target interactions. Taking into account the DBS promiscuity should contribute to explaining the drug promiscuity, the interfamily polypharmacology, presenting hypotheses for drug repositioning or off-target detection, accelerating drug development, and uncovering causes for adverse drug reactions.

4. Materials and Methods

4.1. MOAD Mining

The MOAD (Mother of All Databases) was used, as it is one of the largest databases that provides protein–ligand complexes. It corresponds to high-quality resolution structures that were extracted from the PDB (X-ray structures less than 2.5 Å): 25,769 high-resolution complexed structures were associated with 9,142 binding affinities [26,44].

4.1.1. Drug-Like Ligand Space Analysis and Clustering

Ligand Selection

The MOAD includes a total of 12,440 different ligands. In this study, we focused on valid ligands defined as peptides of fewer than 11 amino acids, oligonucleotides of fewer than four nucleotides, and biologically relevant small molecules. The latter can include agonists, antagonists, cofactors, inhibitors, allosteric regulators, and enzymatic products, but it excludes covalently bonded molecules, crystallographic additives, salts, metals, and solvents.

We focused on drug-like ligands, as druggability is an important aspect of drug design [27]. Drug-like ligands correspond to orally bioavailable small drugs that have an optimal profile of physicochemical properties in terms of absorption, distribution, metabolism, excretion, and toxicity (ADME-Tox), as defined by Lipinski in 1997 [45] and reviewed by Perez-Nueno et al. in 2011 [46]. The FAF-Drugs3 server was used with the “drug-like soft” filter to select drug-like compatible ligands

according to the ADME-Tox properties [47] and based on Lipinski's rules, corresponding, for instance, to a Rings ≤ 6, a RotatableB ≤ 11, and the RatioH/C included between 0.1 and 1.11 values.

Drug-Like Ligand Clustering and Description

The diversity of this drug-like ligand set was analyzed while using Tanimoto similarity, based on MACCS fingerprints [48]. A Tanimoto coefficient that is equal to 1 corresponds to identical MACCS fingerprints, whereas a null coefficient does not share any fingerprint similarity. The Tanimoto coefficients were computed on all ligand pairs using OpenBabel software [49]. Similar ligands were gathered by hierarchical clustering while using the Butina algorithm [50] and Ward method aggregation criterion. Similar ligands with a Tanimoto coefficient threshold greater than or equal to 0.8, as in [29], were clustered and defined as a Ligand–Cluster.

The resulting ligands and Ligand–Clusters were described while using 21 geometrical and physicochemical descriptors that were proposed by FAF-Drugs3 software [47,51] to characterize their main tendencies relative to the promiscuity of their associated DBS. The ligand closest to the average properties of the ligands of each Ligand–Cluster in terms of weighted Euclidean distance was selected as the representative ligand.

4.1.2. Protein Space Analysis and Clustering

The MOAD proposes a protein classification in 18 disjoint classes: seven enzyme classes and 11 others, such as "Binding" (lectin, streptavidin, agglutinins, etc.); "Immune" (antibodies, immunoglobulins, cytokines, etc.); "Transport" (amino acid transporters, electron transport, etc.); and, "Structural" (actin, myosin, etc.). This was used as a reference, called the MOAD protein classes in our study, and corresponded to a large diversity of proteins and enzymes that was consistent for studying various protein classes in comparison to those of marketed small-molecule drug targets (MSMDT) [52].

Homologous protein chains were clustered in "homologous chain clusters" to quantify the different ligand partners of the considered binding sites in the MOAD.

Each protein structure file may contain one or several chains. All of the mono-chains were hierarchically clustered based on their sequence identity while using the H-CD-HIT web server [29], which is a widely used sequence clustering tool for clustering that is based on sequence identity. A first clustering was performed to gather protein chains sharing more than 90% of sequence identity into families, and a second was performed on non-redundant family representatives sharing more than 80% of sequence identity. This two-step CD-HIT classification improves the protein-clustering quality, as it ensures that two similar chains (>90% sequence identity) are clustered into a common homologous chain cluster. Uddin et al. [53] used a sequence identity threshold of 80% to cluster orthologous proteins. Resulting homologous chain clusters may include none or several DBS. For instance, the cluster containing the *Homo sapiens* transferase encompasses a total of 72 mono-chains and three distinct DBS, as illustrated on protein glycogen phosphorylase from *Homo sapiens* (3DDS PDB code chain) in Figure 8.

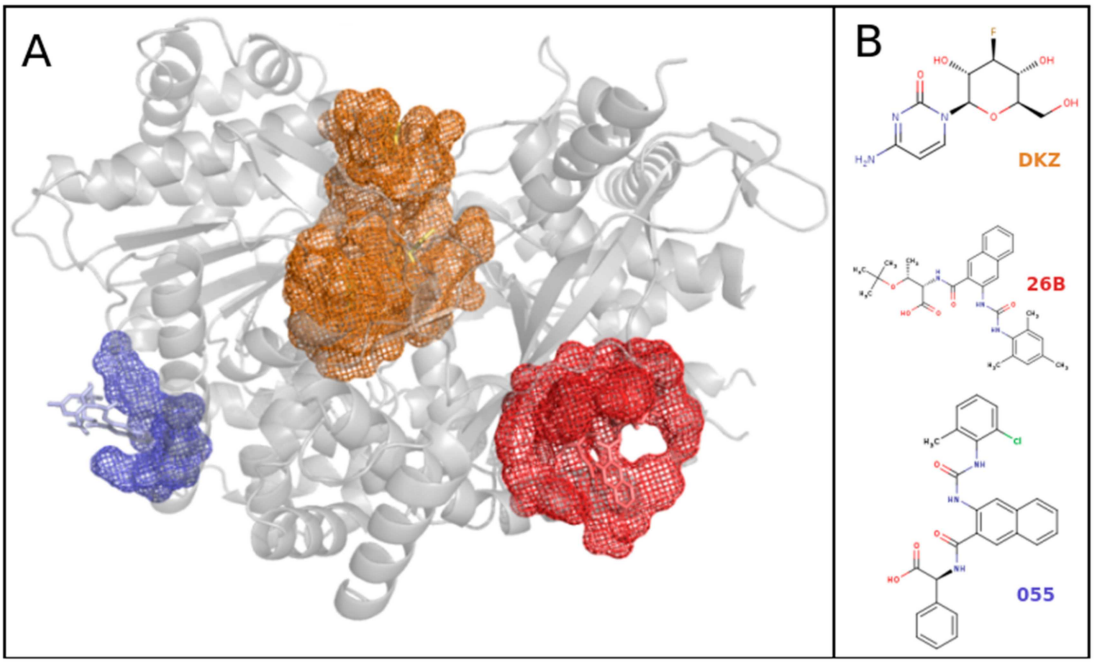


Figure 8. (**A**) Representation of the three pocket clusters (blue, orange, and red) observed on the 72 homologous protein chains from 67 PDB structures (five PDB structures had two chains each) of a family of transferase from *Homo sapiens*. These are illustrated on the PDB file as "glycogen phosphorylase from *Homo sapiens*" (PDB code: 3DDS) in grey. A total of 59 pockets were estimated on this homologous protein chain cluster: one pocket describes the blue binding site, four pockets describe the red binding site, and 54 pockets describe the orange site. (**B**) Representation of one representative ligand binding pocket (DZK, 26B, 055) from these three binding sites. The blue binding site fixed only one ligand, so it could not be determined in terms of promiscuity. The structure visualization was performed using PyMOL.

4.2. MOAD Druggable Binding Sites Extraction Protocol

Each DBS is described by a cluster of pockets estimated by proximity to valid drug-like ligands that are located at an identical region on superimposed homologous chains. DBS were extracted while using the following protocol: (1) we estimated all of the ligand-binding pockets through their proximity to valid drug-like ligands, (2) we superimposed homologous protein mono-chains, including pockets of each homologous chain cluster, and (3) we clustered pockets that were located at the same binding site into a Pocket-Cluster. Each Pocket-Cluster obtained from one homologous chain cluster corresponded to a DBS.

4.2.1. Ligand-Binding Pocket Estimation

The pockets are estimated by proximity to the co-crystallized ligand, as a set of atoms in close contact with a ligand. This close contact is defined by a given threshold to the ligand, commonly between 4 Å and 6 Å [54–57]. Here, the pockets were estimated by a proximity threshold of 5.5 Å, as used by Borrel et al. [34] using PockDrug-Server [33]. In this study, pockets that were located at the interface of several chains were omitted. Estimated pockets were characterized while using 72 physicochemical and geometrical descriptors, such as hydrophobicity, charge, atom, and residue composition described in [34], and the other 20 corresponded to the frequency of each residue to characterize the main tendencies of the pockets.

4.2.2. Superimposition of Mono-Chains from Each Homologous Chain Cluster

All of the mono-chains within each homologous chain cluster were superimposed on the same reference to be able to locate the binding sites. TM-Align software [28] was used to generate the optimal superimposition of structures that are based on the structure similarity and sequence alignment.

4.2.3. Cluster of Pockets Associated with a Druggable Binding Site

A DBS was described by the set of overlapping pockets observed within the superimposed homologous chains, which are referred to as the Pocket-Cluster. It is known that binding sites can be difficult to precisely locate [34]. Here, a criterion was developed to cluster the overlapping pockets corresponding to one DBS.

The steps of the Pocket-Clustering criterion are as follows, for each p pockets of one homologous chain cluster:

- let *n(i)* be the number of atoms of pocket i,
- let *g(i)* be the barycenter of the *n(i)* atoms of the pocket i,
- let d(i,j) be the Euclidean distance between *g(i)* and atom j of pocket i and $d_{max}(i) = \max_j d(i,j)$ the maximum value among the *n(i)* distances between the *n(i)* atoms and the barycenter *g(i)* of pocket i,
- let D and σ be the average value and standard deviation respectively of the population of the p observed values of $d_{max}(i)$ and
- the cutoff value based on D and σ from the p pockets from the homologous chain cluster used is (1):

$$C = D - k \times \sigma \quad (1)$$

where k is a constant parameter.

- Two pockets i_1 and i_2 are fall in the same Pocket-Cluster when $d(g(i_1), g(i_2)) < C$, meaning that the distance between i_1 barycenter and i_2 barycenter is lower than the cutoff value.

For our dataset, we experimentally found that $k = 2$ was an optimal value ensuring the transitivity property of our clustering criterion (i.e., when pockets i_1 and i_2 are in the same cluster and pockets i_2 and i_3 are in the same cluster, therefore i_1 and i_3 must be in the same cluster). Indeed, we do not obtain any non-overlapping pockets when we calculated the score of overlap (SO) between the p pockets of each DBS. It means that the disconnected pockets are clustered in two different Pocket-Clusters. Here, the score of overlap of [34,58] corresponds to the proportion of atoms that are common between two pockets that are associated with one binding site (2):

$$SO = \frac{n_{common}}{n_{i_1} + n_{i_2} - n_{common}} \quad (2)$$

where n_{common} is the number of atoms belonging to both the pocket i_1 and pocket i_2, n_{i_1} and n_{i_2} are the numbers of atoms in pockets i_1 and i_2, respectively. The score of overlap values range from 0 to 1, and a score equal to 1 indicates the maximum overlap.

4.3. Promiscuity Characterization of Druggable Binding Site

4.3.1. Determination of DBS Promiscuity

The promiscuity of a DBS is quantified by the number of different Ligand–Clusters in interaction with its Pocket-Cluster. Selective DBS was defined as DBS in interaction with only one Ligand–Cluster, similar to drug promiscuity and selectivity, as defined by Schneider et al. [59]. Promiscuous DBS was defined as DBS in interaction with more than one Ligand–Cluster. The higher the number of different

Ligand–Clusters in interaction with a DBS, the higher the promiscuity of these DBS. Promiscuous DBS were split into two categories: moderately promiscuous DBS (MP DBS), which were those DBS in interactions with two or three Ligand–Clusters, and the highly promiscuous DBS (HP DBS), which were those observed in interactions with at least four Ligand–Clusters.

Only the DBS observed in more than one complexed chain (Pocket-Cluster size > 1) could be detected as promiscuous, and more than four complexed chains (Pocket-Cluster size ≥ 4) could be detected as HP in the MOAD.

4.3.2. Analysis of DBS Promiscuity in Terms of Pocket and Ligand Properties

We studied the geometrical and physicochemical properties of the pockets that were associated with DBS of different promiscuities to highlight the trends explaining DBS promiscuity. We also studied the geometrical and physicochemical properties of the ligands to understand whether some of the properties explained their ability to interact with DBS of different promiscuity. We performed Student's *t*-tests with a Bonferroni correction, ANOVA, and χ^2-tests to compare the properties of the pockets or ligands that are associated with different DBS promiscuity.

We used Classification and Regression Trees (CART) to select a combination of descriptors that was able to discriminate pockets or ligands associated with selective versus HP DBS [60]. Model performance was evaluated using a five cross-validation approach to prevent overestimation [61]. We balanced the ratio between different selective DBS and HP DBS to train the CART due to the disequilibrium between occurrences of selective and HP pockets or ligands. For pocket prediction, all of the pockets from one Pocket-Cluster were assigned either to the training sample set or to the validation sample set to conduct the cross-validation. This avoided pockets that were associated with an identical DBS (expected to exhibit some similarity) in both the training and validation sets. The quality of these prediction models was evaluated by criteria, such as sensitivity (ability to predict true positive), specificity (ability to predict truly negative), and accuracy, as shown in Equations (3)–(5), respectively:

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$Accucary = \frac{TP}{TP + FP} \quad (5)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

These analyses were performed using the "Stats" package (v3.6.0) implemented in R [62] and the library and function "rpart" in R [63].

4.4. Ligand–Cluster–DBS Interaction Network Illustration

We used a network approach through the igraph library of R software to visualize the interaction between DBS and Ligand–Clusters that are associated with DBS [64]. We drew some interactions between Ligand–Clusters and DBS in a network colored relative to the DBS promiscuity or MOAD protein classification. A DBS and a Ligand–Cluster was linked by an edge if one ligand of the Ligand–Clusters established an interaction with one pocket of the Pocket-Cluster that was associated with the considered DBS. Thus, these networks allowed for the visualization of the main interactions between the DBS and Ligand–Cluster.

Supplementary Materials: The following are available online. Table S1: ANOVA of selective, MP and HP DBS pockets, Table S2: ANOVA of Ligand–Clusters associated to selective, MP and HP DBS, Table S3: ANOVA of Ligand–Clusters associated to selective DBS and Mixed Ligand–Clusters, Table S4: Repartition of DBS among the 17 MOAD protein classes, Figure S5: Network of 39 promiscuous Ligand–Clusters.

Author Contributions: Conceptualization, AC.C; Methodology, ALL; Validation, AC.C., N.C.; Formal analysis, N.C., M.P.; Data curation, N.C., Q.B., M.R.; Writing—original draft preparation, AC.C., N.C.; Writing—review and

editing, ALL; Visualization, N.C.; Supervision, AC.C.; Project administration, AC.C; Funding acquisition: AC.C., N.C.;

Funding: N.C is supported by a fellowship from the Ministère de l'Education Nationale de la Recherche et de Technologie (MENRT)

Acknowledgments: We would like to thank Colette Geneix for her technical help, Laurence Le Gall for her administrative support and Alexandre Borrel for helpful discussion.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Druggable Binding Site (DBS), Selective (S), Moderately Promiscuous (MP), Highly Promiscuous (HP), Mother Of All Databases (MOAD), Classification And Regression Trees (CART), 2 dimensions (2D), 3 Dimensions (3D), Protein Data Bank (PDB), Marketed Small-Molecule Drug Targets (MSMDT); Cyclin-dependent kinase (CDK); Pocket Convexity Index (PCI)

References

1. Lavecchia, A.; Cerchia, C. In silico methods to address polypharmacology: Current status, applications and future perspectives. *Drug Discov. Today* **2016**, *21*, 288–298. [CrossRef] [PubMed]
2. Haupt, V.J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* **2013**, *8*, e65894. [CrossRef]
3. Zhou, H.; Gao, M.; Skolnick, J.; Gao, M.; Skolnick, J.; Skolnick, J.; Gao, M.; von Eichborn, J.; Paolini, G.; Shapland, R.; et al. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* **2015**, *5*, 11090. [CrossRef] [PubMed]
4. Mei, Y.; Yang, B. Rational application of drug promiscuity in medicinal chemistry. *Future Med. Chem.* **2018**, *10*, 1835–1851. [CrossRef] [PubMed]
5. Hu, Y.; Bajorath, J. What is the likelihood of an active compound to be promiscuous? Systematic assessment of compound promiscuity on the basis of PubChem confirmatory bioassay data. *AAPS J.* **2013**, *15*, 808–815. [CrossRef] [PubMed]
6. Jalencas, X.; Mestres, J. Identification of Similar Binding Sites to Detect Distant Polypharmacology. *Mol. Inform.* **2013**, *32*, 976–990. [CrossRef] [PubMed]
7. Paolini, G.V.; Shapland, R.H.B.; Van Hoorn, W.P.; Mason, J.S.; Hopkins, A.L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805–815. [CrossRef] [PubMed]
8. Govindaraj, R.G.; Brylinski, M. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinform.* **2018**, *19*, 1–17. [CrossRef]
9. Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R.V. The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. *Mol. Biosyst.* **2009**, *5*, 1051–1057. [CrossRef]
10. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef]
11. Naderi, M.; Lemoine, J.M.; Govindaraj, R.G.; Kana, O.Z.; Feinstein, W.P.; Brylinski, M. Binding site matching in rational drug design: Algorithms and applications. *Brief. Bioinform.* **2018**, 1–18. [CrossRef] [PubMed]
12. Barnash, K.D.; James, L.I.; Frye, S.V. Target class drug discovery. *Nat. Chem. Biol.* **2017**, *13*, 1053–1056. [CrossRef] [PubMed]
13. Duran-Frigola, M.; Siragusa, L.; Ruppin, E.; Barril, X.; Cruciani, G.; Aloy, P. Detecting similar binding pockets to enable systems polypharmacology. *PLoS Comput. Biol.* **2017**, *13*, e1005522. [CrossRef] [PubMed]
14. Meyers, J.; Chessum, N.E.A.; Ali, S.; Mok, N.Y.; Wilding, B.; Pasqua, A.E.; Rowlands, M.; Tucker, M.J.; Evans, L.E.; Rye, C.S.; et al. Privileged Structures and Polypharmacology within and between Protein Families. *ACS Med. Chem. Lett.* **2018**, *9*, 1199–1204. [CrossRef] [PubMed]
15. Hu, Y.; Bajorath, J. How Promiscuous Are Pharmaceutically Relevant Compounds? A Data-Driven Assessment. *AAPS J.* **2013**, *15*, 104–111. [CrossRef] [PubMed]
16. Kufareva, I.; Ilatovskiy, A.V.; Abagyan, R. Pocketome: An encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res.* **2012**, *40*, 535–540. [CrossRef] [PubMed]

17. Skolnick, J.; Gao, M.; Roy, A.; Srinivasan, B.; Zhou, H. Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function. *Bioorgan. Med. Chem. Lett.* **2015**, *25*, 1163–1170. [CrossRef]
18. Ember, S.W.J.; Zhu, J.-Y.; Olesen, S.H.; Martin, M.P.; Becker, A.; Berndt, N.; Georg, G.I.; Schönbrunn, E. Acetyl-lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors. *ACS Chem. Biol.* **2014**, *9*, 1160–1171. [CrossRef]
19. Antolín, A.A.; Jalencas, X.; Yélamos, J.; Mestres, J. Identification of Pim Kinases as Novel Targets for PJ34 with Confounding Effects in PARP Biology. *ACS Chem. Biol.* **2012**, *7*, 1962–1967. [CrossRef]
20. Barelier, S.; Sterling, T.; O'Meara, M.J.; Shoichet, B.K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS Chem. Biol.* **2015**, *10*, 2772–2784. [CrossRef]
21. Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J.A. Exploring the Role of Receptor Flexibility in Structure-Based Drug Discovery. *Biophys. Chem.* **2014**, *186*, 31–45. [CrossRef] [PubMed]
22. Pabon, N.A.; Camacho, C.J. Probing protein flexibility reveals a mechanism for selective promiscuity. *eLife* **2017**, *6*, e22889. [CrossRef]
23. Stank, A.; Kokh, D.B.; Fuller, J.C.; Wade, R.C. Protein Binding Pocket Dynamics. *Acc. Chem. Res.* **2016**, *49*, 809–815. [CrossRef] [PubMed]
24. Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput. Biol.* **2013**, *9*, e1003302. [CrossRef] [PubMed]
25. Hu, Y.; Bajorath, J. Systematic Assessment of Molecular Selectivity at the Level of Targets, Bioactive Compounds, and Structural Analogues. *ChemMedChem* **2016**, *11*, 1362–1370. [CrossRef]
26. Ahmed, A.; Smith, R.D.; Clark, J.J.; Dunbar, J.B., Jr.; Carlson, H.A. Recent improvements to Binding MOAD: A resource for protein–ligand binding affinities and structure. *Nucleic Acids Res.* **2015**, *43*, D465–D469. [CrossRef] [PubMed]
27. Abi Hussein, H.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A.-C. Global vision of druggability issues: Applications and perspectives. *Drug Discov. Today* **2017**, *22*, 404–415. [CrossRef]
28. Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309. [CrossRef]
29. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef]
30. Mentese, A.; Erkut, N.; Demir, S.; Yaman, S.O.; Sumer, A.; Erdem, M.; Alver, A.; Sönmez, M.G. Serum carbonic anhydrase I and II autoantibodies in patients with chronic lymphocytic leukaemia. *Cent. J. Immunol.* **2018**, *43*, 276–280. [CrossRef]
31. Chohan, T.A.; Chen, J.-Z.J.-J.; Qian, H.-Y.; Pan, Y.-L.; Chen, J.-Z.J.-J. Molecular modeling studies to characterize N-phenylpyrimidin-2-amine selectivity for CDK2 and CDK4 through 3D-QSAR and molecular dynamics simulations. *Mol. BioSyst.* **2016**, *12*, 1250–1268. [CrossRef] [PubMed]
32. Van Bergen, L.A.H.; Alonso, M.; Palló, A.; Nilsson, L.; De Proft, F.; Messens, J. Revisiting sulfur H-bonds in proteins: The example of peroxiredoxin AhpE. *Sci. Rep.* **2016**, *6*, 30369. [CrossRef] [PubMed]
33. Hussein, H.A.; Borrel, A.; Geneix, C.; Petitjean, M.; Regad, L.; Camproux, A.-C.C. PockDrug-Server: A new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res.* **2015**, *43*, W436–W442. [CrossRef] [PubMed]
34. Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.-C.C. PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties. *J. Chem. Inf. Model.* **2015**, *55*, 882–895. [CrossRef] [PubMed]
35. Degryse, B. The urokinase receptor system as strategic therapeutic target: Challenges for the 21st century. *Curr. Pharm. Des.* **2011**, *17*, 1872–1873. [CrossRef] [PubMed]
36. Cerisier, N.; Regad, L.; Triki, D.; Petitjean, M.; Flatters, D.; Camproux, A.-C.A.C. Statistical Profiling of One Promiscuous Protein Binding Site: Illustrated by Urokinase Catalytic Domain. *Mol. Inform.* **2017**, *36*, 1700040. [CrossRef] [PubMed]
37. Schrödinger, LLC. *The PyMOL Molecular Graphics System*; Version 2.0; Schrödinger, LLC.: New York, NY, USA, 2015.
38. Cooney, C.M. Triclosan comes under scrutiny. *Environ. Health Perspect.* **2010**, *118*, A242. [CrossRef]
39. Gao, M.; Skolnick, J. APoc: Large-scale identification of similar protein pockets. *Bioinformatics* **2013**, *29*, 597–604. [CrossRef]

40. Liargkova, T.; Eleftheriadis, N.; Dekker, F.; Voulgari, E.; Avgoustakis, C.; Sagnou, M.; Mavroidi, B.; Pelecanou, M.; Hadjipavlou-Litina, D. Small Multitarget Molecules Incorporating the Enone Moiety. *Molecules* **2019**, *24*, 199. [CrossRef]
41. Bajorath, J. Analyzing Promiscuity at the Level of Active Compounds and Targets. *Mol. Inform.* **2016**, *35*, 583–587. [CrossRef] [PubMed]
42. Rifaioglu, A.S.; Atas, H.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. *Brief. Bioinform.* **2018**, 1–35. [CrossRef] [PubMed]
43. Chaudhari, R.; Tan, Z.; Huang, B.; Zhang, S. Computational polypharmacology: A new paradigm for drug discovery. *Expert Opin. Drug Discov.* **2017**, *12*, 279–291. [CrossRef] [PubMed]
44. Hu, L.; Benson, M.L.; Smith, R.D.; Lerner, M.G.; Carlson, H.A. Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinform.* **2005**, *60*, 333–340. [CrossRef] [PubMed]
45. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. [CrossRef]
46. Ritchie, D.W.; Nancy, I.; Botanique, J.; Peérez-Nueno, V.I.; Ritchie, D.W. 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.* **2011**, *51*, 1233–1248.
47. Miteva, M.A.; Violas, S.; Montes, M.; Gomez, D.; Tuffery, P.; Villoutreix, B.O. FAF-Drugs: Free ADME/tox filtering of compound collections. *Nucleic Acids Res.* **2006**, *34*, W738–W744. [CrossRef]
48. MDL Information Systems, Inc. *MACCS Drug Data Report, Release 2000.2*; MDL Information Systems, Inc.: San Leandro, CA, USA, 2000.
49. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef]
50. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750. [CrossRef]
51. Lagorce, D.; Sperandio, O.; Baell, J.B.; Miteva, M.A.; Villoutreix, B.O. FAF-Drugs3: A web server for compound property calculation and chemical library design. *Nucleic Acids Res.* **2015**, *43*, W200–W207. [CrossRef]
52. Hopkins, A.L.; Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.* **2002**, *1*, 727–730. [CrossRef] [PubMed]
53. Uddin, R.; Jamil, F. Prioritization of potential drug targets against P. aeruginosa by core proteomic analysis using computational subtractive genomics and Protein-Protein interaction network. *Comput. Biol. Chem.* **2018**, *74*, 115–122. [CrossRef] [PubMed]
54. Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins Struct. Funct. Bioinform.* **2008**, *71*, 1755–1778. [CrossRef] [PubMed]
55. Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135. [CrossRef] [PubMed]
56. Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinform.* **2008**, *9*, 543. [CrossRef] [PubMed]
57. Feldman, H.J.; Labute, P. Pocket similarity: Are alpha carbons enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466–1475. [CrossRef] [PubMed]
58. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform.* **2009**, *10*, 168. [CrossRef]
59. Schneider, P.; Schneider, G. A Computational Method for Unveiling the Target Promiscuity of Pharmacologically Active Compounds. *Angew. Chem. Int. Ed.* **2017**, *56*, 11520–11524. [CrossRef]
60. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984; Volume 1, ISBN 978-0412048418.
61. Zhang, P. Model Selection via Multifold Cross Validation. *Ann. Stat.* **1993**, *21*, 299–313. [CrossRef]
62. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.

63. Therneau, T.M.; Atkinson, E.J. *An Introduction to Recursive Partitioning Using the RPART Routine*; Mayo Clinic, Division Of Biomedical Statistics And Informatics: Rochester, MN, USA, 1997.
64. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst*. **2006**, *1695*, 1–9.

Table S1. All the pocket descriptors (except the frequency of each atoms) sorted according to the average comparison test (ANOVA non parametric) between the different degree of promiscuity. The *p*-value range is: $<1x10^{-20}$ (***), $<1x10^{-10}$ (**), $<5x10^{-2}$ (*) and $>5x10^{-2}$(-). The threshold of significance has been determined with the Bonferroni correction (alpha risk divided by the number of tests). Other columns are the average and the standard deviation of the 791, 1447, 5029 and 7267 pockets from the selective, MP, HP and all the DBS from the DBS4 dataset.

Pocket Descriptor	*p*-value	S DBS		MP DBS		HP DBS		Total	
		Av[1]	Sd[2]	Av	Sd	Av	Sd	Av	Sd
SURFACE_HULL	***	538.3	144.1	596.5	175.9	679.7	166.6	647.7	173.6
RADIUS_HULL	***	8.66	1.31	9.11	1.48	9.84	1.47	9.57	1.52
DIAMETER_HULL	***	17.14	2.68	18.00	2.97	19.48	2.97	18.93	3.06
VOLUME_HULL	***	1031.6	408.5	1214.6	538.3	1459.1	530.6	1363.9	541.6
RADIUS_CYLINDER	***	8.48	1.35	8.90	1.49	9.62	1.49	9.35	1.54
SMALLEST_SIZE	***	9.76	1.67	10.32	1.96	11.03	1.66	10.75	1.78
C_res	***	16.62	4.30	18.14	5.34	20.12	5.10	19.34	5.21
C_ATOM	***	72.30	23.08	79.79	28.23	91.02	29.46	86.75	29.36
FACE	***	61.30	13.85	65.43	14.81	70.62	14.42	68.57	14.80
p_main_chain_atom	***	0.328	0.145	0.313	0.140	0.375	0.136	0.358	0.140
p_side_chain_atom	***	0.672	0.145	0.687	0.140	0.625	0.136	0.642	0.140
X._ATOM_CONVEXE	***	0.470	0.079	0.457	0.080	0.430	0.079	0.440	0.081
p_sulfur_atom	***	0.006	0.010	0.009	0.012	0.013	0.014	0.011	0.013
p_O_atom	***	0.076	0.035	0.071	0.035	0.086	0.031	0.082	0.033
hydrophobic_kyte	***	-0.158	1.021	-0.248	1.121	0.182	0.960	0.060	1.018
p_Ntrp_atom	***	0.004	0.009	0.007	0.011	0.003	0.007	0.004	0.009
polarity	***	0.146	0.084	0.138	0.077	0.171	0.079	0.161	0.080
p_positive_res	***	0.122	0.094	0.137	0.093	0.104	0.075	0.112	0.082
p_polar_res	***	0.536	0.174	0.549	0.178	0.487	0.162	0.504	0.169
p_charged_res	***	0.224	0.135	0.255	0.148	0.208	0.112	0.219	0.124
p_aliphatic_res	***	0.206	0.122	0.194	0.124	0.237	0.126	0.225	0.127
p_Car_atom	***	0.200	0.111	0.209	0.129	0.170	0.118	0.181	0.121
p_hydrophobic_res	***	0.689	0.133	0.692	0.158	0.729	0.117	0.717	0.129
p_ND1_atom	***	0.009	0.013	0.011	0.015	0.007	0.011	0.008	0.012
p_S_atom	***	0.002	0.005	0.002	0.006	0.004	0.008	0.003	0.007
p_NE2_atom	***	0.010	0.014	0.012	0.016	0.008	0.012	0.009	0.013
p_nitrogen_atom	***	0.146	0.049	0.147	0.046	0.135	0.041	0.138	0.043
p_aromatic_res	***	0.232	0.119	0.257	0.153	0.218	0.128	0.227	0.134
p_Ocoo_atom	***	0.006	0.009	0.009	0.013	0.007	0.010	0.007	0.011
p_Cgln_atom	**	0.012	0.013	0.011	0.014	0.009	0.010	0.009	0.011
p_N_atom	**	0.092	0.042	0.091	0.038	0.100	0.039	0.097	0.039
hydrophobicity	**	0.033	0.031	0.030	0.031	0.038	0.028	0.036	0.029
INERTIA_2	**	0.308	0.054	0.301	0.045	0.293	0.049	0.296	0.049
charge	**	0.314	1.738	0.431	2.134	-0.012	1.826	0.111	1.892
p_Ccoo_atom	**	0.016	0.015	0.018	0.019	0.015	0.013	0.015	0.015
p_tiny_res	**	0.183	0.103	0.177	0.117	0.199	0.111	0.193	0.111
p_negative_res	*	0.102	0.079	0.117	0.111	0.104	0.072	0.106	0.082
p_Nlys_atom	*	0.004	0.007	0.005	0.010	0.005	0.007	0.005	0.008
p_C_atom	*	0.166	0.065	0.155	0.058	0.164	0.057	0.163	0.059
p_hyd_atom	*	0.113	0.043	0.120	0.045	0.113	0.044	0.115	0.044
p_oxygen_atom	*	0.159	0.044	0.156	0.046	0.153	0.038	0.154	0.041
p_Otyr_atom	*	0.014	0.014	0.013	0.014	0.012	0.013	0.012	0.013

INERTIA_1	*	0.511	0.077	0.513	0.068	0.521	0.070	0.518	0.071
p_hydrophobic_atom	*	0.122	0.042	0.126	0.045	0.121	0.044	0.122	0.044
p_carbone_atom	*	0.676	0.084	0.681	0.073	0.686	0.084	0.684	0.082
PCI	*	0.025	0.012	0.027	0.013	0.026	0.012	0.026	0.012
INERTIA_3	*	0.181	0.048	0.186	0.047	0.186	0.041	0.185	0.043
CONV.SH_COEFF	*	0.978	0.025	0.976	0.028	0.978	0.026	0.978	0.027
p_Carg_atom	*	0.017	0.015	0.016	0.015	0.016	0.017	0.016	0.016
p_Ooh_atom	*	0.017	0.015	0.016	0.015	0.016	0.017	0.016	0.016
PSI[3]	-	0.541	0.083	0.541	0.079	0.537	0.066	0.538	0.071
p_small_res[3]	-	0.437	0.132	0.423	0.137	0.428	0.143	0.428	0.141

[1] Average

[2] Standard deviation

[3] These descriptors are not significant (p-value>$5x10^{-2}$)

Table S2. All the ligand descriptors sorted according to the average comparison test (ANOVA non parametric) between the different degree of promiscuity. The *p*-value range is: $<1x10^{-10}$ (***), $<1x10^{-5}$ (**), <0.01 (*) and >0.01(-). The threshold of significance has been determined with the Bonferroni correction (alpha risk divided by the number of tests). Other columns are the average and the standard deviation of the 53, 257, 2621 and 3488 ligands from the Ligand-Clusters that respectively bind to selective, MP, HP and all the DBS from the DBS4 dataset.

Ligand Descriptor	***p*-value**	**S DBS**		**MP DBS**		**HP DBS**		**Total**	
		Av[1]	Sd[2]	Av	Sd	Av	Sd	Av	Sd
Rings	***	1.11	0.58	1.91	1.10	2.27	0.98	2.22	1.00
RigidB	***	10.57	5.83	15.27	7.38	18.12	6.70	17.74	6.87
HeavyAtoms	***	16.43	5.70	21.17	8.09	23.49	7.46	23.16	7.57
CarbonAtoms	***	11.47	4.00	15.12	6.52	16.82	5.89	16.57	5.98
logSw	***	-1.91	1.63	-2.90	1.70	-3.40	1.46	-3.33	1.50
MW	***	241.37	86.35	309.06	115.39	337.56	107.03	333.32	108.44
Solubility (mg/l)	***	$8.2x10^4$	$1.0x10^5$	$4.6x10^4$	$9.4x10^4$	$2.5x10^4$	$5.9x10^4$	$2.8x10^4$	$6.4x10^4$
NumCharges	***	1.00	0.94	0.77	0.73	0.54	0.71	0.57	0.72
logD	**	0.05	2.49	0.49	2.56	1.40	2.22	1.29	2.27
HBA	**	4.00	2.07	5.25	2.06	5.60	2.21	5.54	2.20
HeteroAtoms	**	4.96	3.09	6.05	2.36	6.67	2.62	6.59	2.62
TotalCharge	**	-0.17	0.80	-0.27	0.80	0.00	0.73	-0.03	0.74
logP	**	1.09	2.19	1.94	1.92	2.42	1.69	2.35	1.74
HBD/HBA	**	5.81	3.08	7.55	2.78	7.93	2.99	7.85	2.99
MaxSizeRing	**	7.26	3.79	7.12	2.92	7.98	2.69	7.89	2.75
Flexibility	*	0.30	0.23	0.23	0.16	0.20	0.13	0.20	0.13
tPSA	*	70.71	42.16	88.54	32.71	89.64	34.60	89.20	34.67
HBD	*	1.81	1.29	2.31	1.25	2.33	1.32	2.32	1.31
RotatableB	*	3.74	2.53	4.04	2.40	4.37	2.53	4.33	2.53
Lipinski Violation[3]	-	0.02	0.14	0.11	0.35	0.12	0.38	0.12	0.37
ratioH/C[3]	-	0.46	0.25	0.45	0.19	0.43	0.18	0.43	0.19

[1] Average

[2] Standard deviation

[3] These descriptors are not significant (*p*-value>0.01)

Table S3. All the ligands descriptors sorted according to the average comparison test (non-parametric ANOVA) between the 53 ligands from the 29 Ligand-Clusters that bind to selective DBS and the 250 ligands from the 39 Mixed Ligand-Clusters that bind to at least one selective DBS. The *p*-value range is: $<1x10^{-5}$ (**), $<5x10^{-2}$(*) and $>5x10^{-2}$(-). The threshold of significance has been determined with the Bonferroni correction (alpha risk divided by the number of tests). Other columns are the average and the standard deviation of each class of ligands.

Ligand Descriptor	***p*-value**	**Selective DBS**		**Mixed**	
		Av[1]	std.dev[2]	Av	Std.dev
HBD	**	1.81	1.29	3.11	1.49
HBD/HBA	**	5.81	3.08	8.70	3.60
HBA	**	4.00	2.07	5.59	2.51
Rings	*	1.11	0.58	1.47	0.76
tPSA	*	70.71	42.16	94.28	36.27
NumCharges	*	1.00	0.94	0.51	0.70
ratioH/C	*	0.46	0.25	0.59	0.30
Flexibility	*	0.30	0.23	0.20	0.17
RigidB	*	10.57	5.83	12.72	6.07
HeteroAtoms	*	4.96	3.09	5.98	2.53
RotatableB	*	3.74	2.53	2.93	2.56
HeavyAtoms	*	16.43	5.70	18.23	6.26
Lipinski.Violation[3]	-	0.02	0.14	0.07	0.27
logD[3]	-	0.05	2.49	-0.66	2.72
MW[3]	-	241.37	86.35	264.92	87.92
logP[3]	-	1.09	2.19	0.52	2.38
Solubility (mg/l)[3]	-	$8.23x10^{4}$	$1.03x10^{4}$	$1.08x10^{5}$	$1.17x10^{5}$
CarbonAtoms[3]	-	11.47	4.00	12.24	5.77
logSw[3]	-	-1.91	1.63	-1.75	1.79
MaxSizeRing[3]	-	7.26	3.79	7.63	3.89
TotalCharge[3]	-	-0.17	0.80	-0.22	0.73

[1] Average

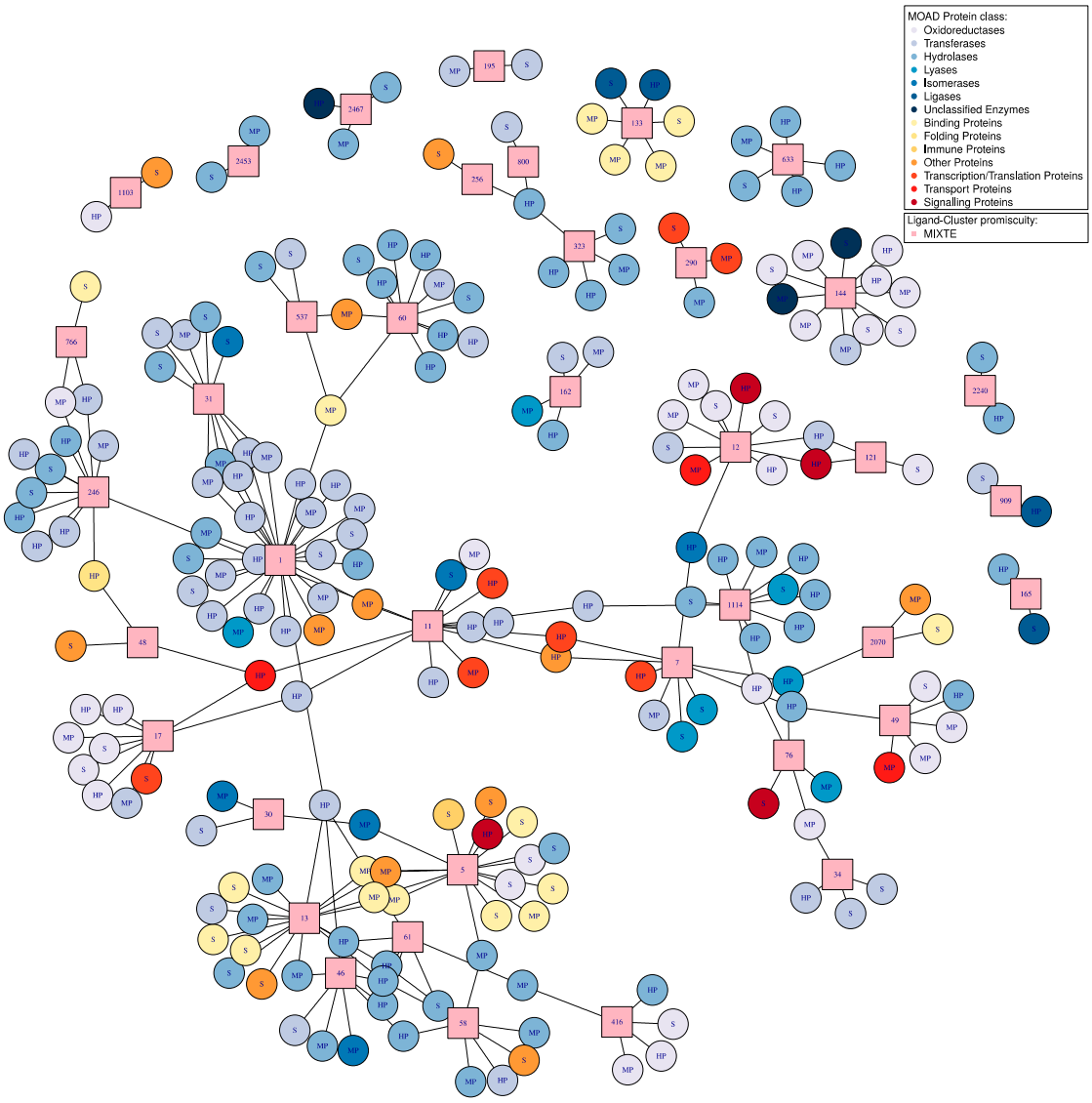
[2] Standard deviation

[3] These descriptors are not significant (*p*-value>0.01)

Table S4. Repartition of the 481 DBS among the 17 MOAD protein classes represented at least once in our DBS4 dataset. The columns respectively represent the number of DBS per MOAD protein class, their proportion, the average number of pocket per DBS, the number of selective, MP and HP DBS, the number of Ligand-Clusters and the average promiscuity per DBS for each class.

MOAD Protein Class	Occurrence	Prop. of class	Average number of pocket/DBS	S	MP	HP	Number of Ligand-Cluster	Average prom.
Transferases	134	27.86	15.38	18	36	80	741	5.5
Hydrolases	126	26.2	15.02	26	37	63	602	4.8
Oxidoreductases	86	17.88	13.49	22	40	24	243	2.8
Other Prot.	24	4.99	11.21	7	13	4	92	3.8
Lyases	20	4.16	16.50	5	8	7	117	5.9
Binding Prot.	20	4.16	11.30	9	10	1	24	1.2
Isomerases	15	3.12	10.07	3	8	4	46	3.1
Transcription/ Translation Prot.	15	3.12	23.53	2	3	10	97	6.5
Signaling Prot.	13	2.7	23.38	1	4	8	69	5.3
Ligases	8	1.66	14.13	3	0	5	39	4.9
Transport Prot.	7	1.46	16.43	0	3	4	41	5.9
Unclassified Enzymes	4	0.83	9.00	1	2	1	9	2.3
Folding Prot.	3	0.62	57.67	0	1	2	94	31.3
Immune Prot.	2	0.42	5.50	1	0	1	5	2.5
Mobility Prot.	2	0.42	20.00	1	0	1	12	6.0
Cell Cycle Prot.	1	0.21	6.00	0	1	0	2	2.0
Structural Prot.	1	0.21	27.00	1	0	0	1	1.0

Figure S5. Network of the 39 promiscuous Ligand-Clusters in interaction with both selective and promiscuous DBS, composed of 213 DBS resulting in 262 interactions. The circles represent the 213 DBS; they are colored according to the MOAD protein class which they belong and named according to their promiscuity. The squares represent the 39 mixed Ligand-Clusters, with their number written inside. The grey lines represent the interactions between DBS and Ligand-Clusters. This network visualization is made by the igraph package.



Chapitre 4. PREDICTION DES INTERACTIONS

Comme mentionné dans le chapitre 1, il existe de nombreux outils computationnels aux objectifs et aux méthodes différentes.

Les approches protéochémométriques développées dans les chapitres 2 et 3 nous ont permis de caractériser finement les interactions avec des partenaires multiples en prenant en compte conjointement les espaces des poches et des ligands. Ces premiers travaux ont été déterminants pour le développement du protocole de prédiction : ils ont mis en évidence les propriétés (descripteurs) importantes à prendre en compte pour la description conjointe poche-ligand et l'étude des interactions multipartenaires. Les méthodes statistiques développées et utilisées ont été validées. Ces travaux nous ont aussi permis de choisir judicieusement les banques de données adaptées à la prédiction d'interactions multipartenaires et à plus long terme, à la polypharmacologie. Les banques de données utile à ce protocole ont été élaborées à partir des observations et des conclusions tirées de mes précédents travaux de thèse.

Dans ce dernier chapitre, je détaille donc notre protocole de prédiction de ces interactions. Il implique, en première partie, la création d'une banque de données qui regroupe un nombre important d'interactions entre protéines et ligands, décrites à l'aide de la méthode protéochémométrique précédente. En seconde partie, l'utilisation des profils statistiques des poches et des ligands ainsi que des métriques de similarité nous permet de proposer une chimiothèque réduite de composés les plus probables à la liaison avec une protéine d'intérêt, dans le but de valider leurs interactions avec des méthodes de docking. Il combinera dans sa forme finale et complète plusieurs méthodes de traitement des données, de filtrage statistique, de méthodes de prédiction afin d'automatiser au mieux tous les traitements et la prédiction des interactions, complémentarisé par un protocole de docking. La plupart de ces méthodes ont été développées au sein du laboratoire (PockDrug, caractérisation conjointe, filtrage statistique, etc.) et sont donc maîtrisées.

Concernant la prédiction des molécules à partir de protéine, il existe des outils qui prédisent les poches de liaison et d'autres outils qui proposent des protocoles de docking et de criblage virtuel, mais peu combinent plusieurs méthodes à des fins de prédiction.

L'originalité de notre outil réside dans la caractérisation fine et explicite (en termes physicochimiques et géométriques) des deux partenaires de l'interaction. De plus, il combine de nombreuses méthodes, permettant une prédiction ne nécessitant qu'une structure tridimensionnelle et fournissant (dans sa forme complète) la validation in silico des interactions prédites. La banque de données créée peut aussi être réutilisée à d'autres fins et d'autres outils peuvent être appliqués dessus.

4.1 Construction d'une banque de données

La première étape du protocole de prédiction est de rassembler tous les complexes protéine-ligands pertinents. La méthode de caractérisation conjointe des complexes servira

ensuite à répertorier les profils de poches qui interagissent avec les profils de ligands. Ces « profils d'interaction » serviront d'apprentissage pour la prédiction.

4.1.1 Matériel et méthodes

Les étapes de la construction de la banque de données sont détaillées dans la figure 5. Bien que le schéma décrive une extraction des poches et des ligands sur la PDB (Berman et al. 2002), les complexes de la première version de la banque de données ont été extraits de la MOAD (L. Hu et al. 2005; Ahmed et al. 2015), base de données d'interactions de haute qualité qui a servi à la mise en place du protocole d'étude de la promiscuité (section 2.5.1). Nous avons testé le protocole de caractérisation des complexes sur MOAD puis ensuite étendu l'extraction des complexes à toute la PDB.

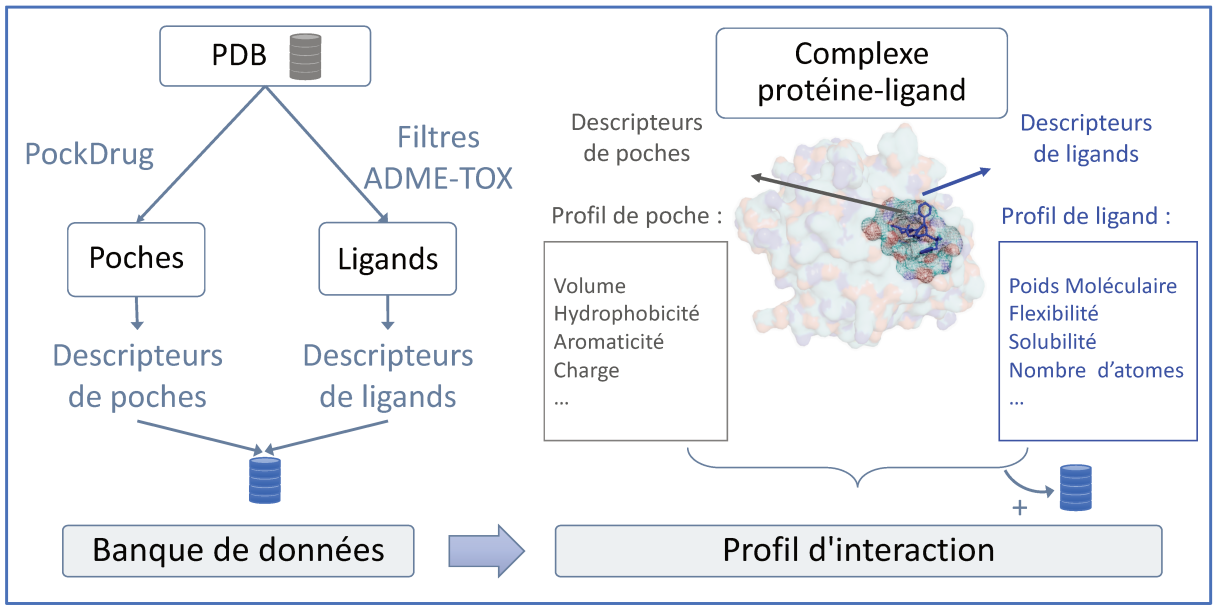


Figure 5 : Schéma du protocole de construction de notre banque de données de profils d'interaction entre les poches et les ligands. La première partie consiste à extraire les complexes pertinents de la PDB, la deuxième partie consiste à caractériser chaque poche et chaque ligand pour en extraire des profils d'interaction, qui constitueront notre banque de données.

4.1.1.1 Extraction des poches et description

De la même manière que pour la caractérisation des Urokinases et l'étude de la promiscuité, les poches ont été extraites de la MOAD puis de la PDB grâce à l'outil PockDrug (Borrel et al. 2015). Les poches des deux types – par proximité (5,5 Å) et par géométrie avec FPocket (Le Guilloux, Schmidtke, and Tuffery 2009) – ont été extraites avec les 52 descripteurs disponibles (cf. section 1.4.1), ces méthodes ayant fait leurs preuves dans les études précédentes.

Un filtre concernant la taille des poches a été appliqué au jeu de données. En effet comme montré dans l'étude de Hussein (Hussein et al. 2015), les poches de moins de 14 résidus ne sont pas significatives, et les poches de plus de 60 résidus sont difficilement estimables par FPocket, ce qui les rend trop différentes de l'estimation par proximité au ligand.

Toutes les poches retenues ont été comparées entre elles au moyen de calcul de la distance euclidienne pondérée à partir des descripteurs. En plus des profils de poches, cette matrice de similarité préalablement calculée compose la banque de données.

Parallèlement, j'ai eu l'opportunité de participer à l'encadrement de stage de fin de Master 2 (Master ISDD – 5 mois – Mlle Indusha Kugathas) qui traitait une autre manière de caractériser les poches avant de prédire leur capacité à interagir avec des ligands promiscuous. En comparaison de la caractérisation avec les 52 descripteurs fournis par PockDrug, l'outil Fuzcav (N. Weill and Rognan 2010) a été utilisé afin d'établir les profils pharmacophoriques des poches. Pour chaque poche, Fuzcav créé un vecteur appelé *Fingerprint* (FP), composé de 4833 cases représentant chacune une combinaison de distances possibles de 3 acides aminés et une combinaison de propriétés pharmacophoriques associées (appelé triplet de pharmacophore). Dans chaque case il est indiqué le nombre de fois pour laquelle ce triplet pharmacophorique apparaît dans la poche (ou 0 à défaut). Mlle Indusha Kugathas a développé des critères de similarité des interactions basées sur des pharmacophores consensus, qui permettent de comparer les poches entre elles efficacement en termes de pharmacophores. Cette approche de similarité présente l'avantage important d'être relativement insensible à la méthode de définition de la poche et peut être appliquée pour des poches liées ou non liées à un ligand. Elle est en cours de mise en place dans notre protocole de prédiction.

4.1.1.2 Sélection des ligands et description

A ce stade du développement de l'outil, aucun filtre n'a été appliqué sur les ligands, ils ont tous été pris en compte dans l'extraction des complexes. Les ligands ont été décrits avec l'outil RDkit (http://www/rdkit.org, un outil open-source) en prévision de l'extension prévue dans le cadre de la prédiction de poche à partir des ligands.

La comparaison des ligands est cependant possible puisqu'une matrice de similarité est calculée entre chacun des ligands de la banque de données. Comme expliqué dans la section *Materials and Method* de l'article présenté en chapitre 3, le coefficient de Tanimoto (Rogers and Tanimoto 1960) est basé sur des *Fingerprints* (ici MACCS) (MDL Information Systems Inc. 2000). Ces dernières sont représentées par une suite de cases appelées « clefs » (correspondant à des propriétés particulières) affectées de la valeur 1 ou 0 selon, respectivement, la présence ou l'absence de certains atomes appartenant à cette clef. Cette suite binaire forme un *Fingerprint* qui est utilisé pour calculer la similarité entre deux molécules.

4.1.2 Résultats : les profils d'interactions

Un profil d'interaction est donc définit comme une association entre le profil de la poche et le profil du ligand. Après construction de notre banque de données, nous avons obtenu 123 561 structures PDB, dont 29 600 d'entre elles ne présentent pas de ligand et donc pas de poche. Notre banque de données contient donc 93 961 structures, 268 799 poches estimées par proximité, 1 979 511 poches estimées par Fpocket et 27 291 ligands différents (en termes de code hétéroatomes référencés dans la PDB). La base de données ZINC propose, quant à elle, 189 100 ligands.

En ayant pris pour base le protocole développé par Mlle Indusha Kugathas (avec ré-implémentation de fonctions afin de prendre en compte de nombreuses exceptions dues aux fichiers PDB non rigoureux) et afin d'anticiper les perspectives de ce projet, j'ai calculé les pharmacophores de toutes les poches précédemment estimées par proximité de la PDB. La banque de données est donc enrichie de 209 677 pharmacophores. En perspective, elle pourra être enrichie des pharmacophores sur les poches estimées par Fpocket, qui correspondent aux poches par proximité (superposées) dans un premier temps. Dans un second temps, les autres poches estimées par Fpocket pourront être intégrées, après avoir vérifié leur pertinence : les poches Fpocket sont des cavités probables et ne sont pas obligatoirement associées à un ligand, elles ne pourront donc pas servir à la proposition de ligand(s) pour une chimiothèque.

4.2 Protocole de prédiction des interactions

Une fois la banque de données constituée, il est possible de prédire les interactions. Lors de la prédiction d'interaction entre protéine(s) et ligand(s) deux cas se présentent : la prédiction du ligand à partir d'une protéine ou l'inverse. Nous avons choisi de commencer le développement de cet outil par la prédiction de ligand à partir d'une structure protéique. Une extension de ce projet sera de développer la partie qui consiste à prédire la poche en partant d'une molécule d'intérêt. Les données recueillies dans notre banque de données permettent cette extension puisque les deux acteurs de l'interaction sont caractérisés. Le principe, illustré en figure 6, sera détaillé dans les sections suivantes.

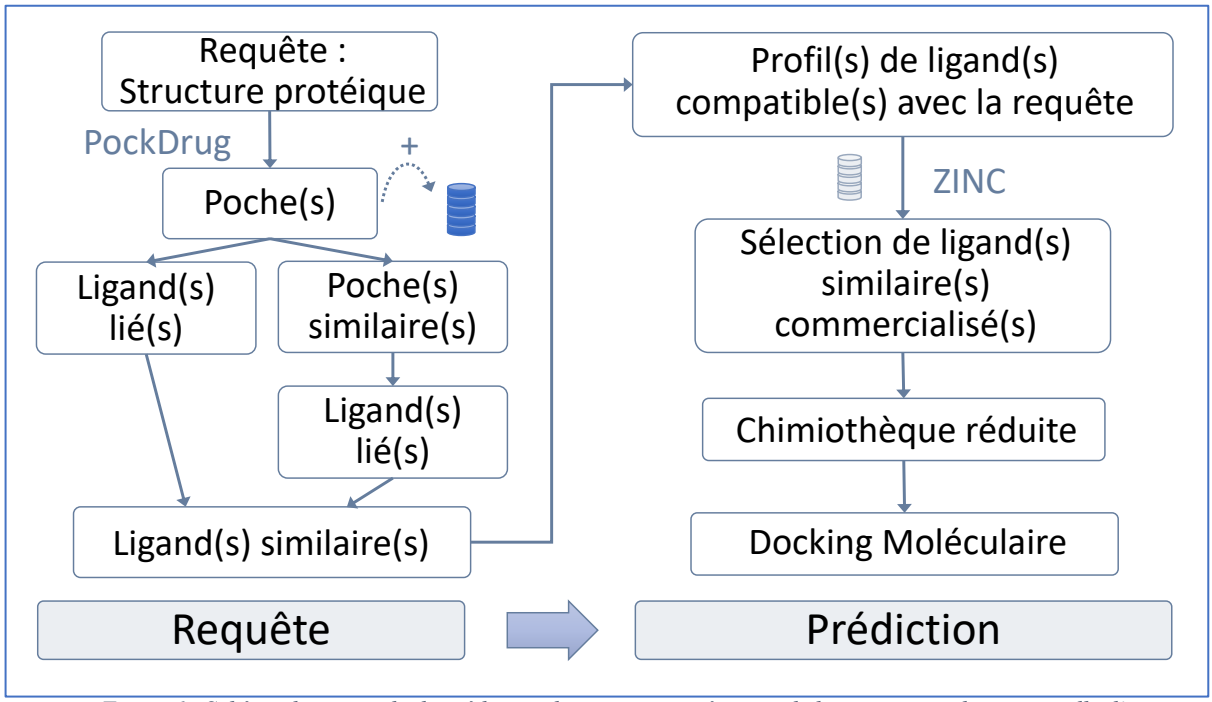


Figure 6 : Schéma du protocole de prédiction des interactions à partir de la structure tridimensionnelle d'une protéine d'intérêt. La première étape consiste à extraire les poches et les ligands, mais aussi les poches proches ainsi que leur ligand. Ces ligands sont filtrés dans une base de données de ligands commercialisés de manière à fournir une chimiothèque réduite de composés commercialisés. Le docking permet de confirmer de manière in silico les énergies de liaisons entre la protéine d'intérêt et les ligands prédits.

4.2.1 Matériel et méthodes

Comme mentionné plus haut, ce protocole a d'abord été établi à l'aide de la base de données d'interactions de haute qualité, MOAD. Bien qu'il ait été étendu à la PDB, les résultats préliminaires présentés concernent la version établie sur MOAD.

4.2.1.1 Extraction des poches d'une protéine d'intérêt

La prédiction des interactions se fait à partir d'une structure tridimensionnelle (ou de son code PDB). Lorsque seul le code PDB est fourni, la structure est téléchargée de la PDB. Les poches de la protéine sont estimées avec l'outil PockDrug, soit par proximité au(x) ligand(s) s'il en existe dans la structure, soit basé sur la géométrie des cavités, par FPocket (cf. section 1.4.1).

Les 52 descripteurs physicochimiques et géométriques sont aussi fournis, le profil de la poche peut donc être créé à cette étape.

4.2.1.2 Construction d'une chimiothèque compatible avec la protéine d'intérêt

Une fois la/les poche(s) extraite(s), il est possible d'en estimer la drugabilité et la promiscuité grâce aux descripteurs les plus informatifs, sélectionnés dans les études précédentes. À notre connaissance, aucun outil ne prédit la drugabilité et la promiscuité d'une poche. Ces informations sont utiles à la suite du protocole, en effet une protéine qui n'aurait aucune proche *druggable* sera plus difficile à cibler et une protéine dont la poche est très promiscuous sera certainement très difficile à lier de manière spécifique par une molécule candidat médicament.

Parallèlement, j'ai eu l'opportunité de participer à l'encadrement d'un deuxième stage de Master 1 (Master Bioinformatique – 3 mois – Mr Bertrand Bouvarel) qui a développé un protocole permettant de prédire le caractère promiscuous d'une poche avec des méthodes de *Deep Learning* et la fonction de linéarisation ReLU (P. Wang et al. 2016). À partir des poches estimées de la MOAD et de leur description par PockDrug, un réseau de neurones à couches cachées a appris un grand nombre d'interactions entre profils de poches et profils de ligands. Les poches promiscuous ont été différenciées des non-promiscuous et le modèle a été développé afin de prédire ce caractère pour des futures poches. La comparaison de ce modèle avec les méthodes classiques de *Machine Learning* (Lavecchia 2014; Rodríguez-Pérez and Bajorath 2018) montre qu'il est plus performant et présente l'avantage de ne pas nécessiter de sélection de variables au préalable. La prédiction de la promiscuité des poches pourra donc être améliorée en intégrant les méthodes de *Deep Learning*.

L'étape suivante du protocole consiste à (i) rechercher les poches similaires à celle(s) de la protéine d'intérêt et leur(s) ligand(s) connu(s) et (ii) si la/les poche(s) extraite(s) ont des ligands, rechercher des ligands similaires.

La notion de similarité est primordiale dans cette partie du protocole et différents critères définissent la similarité entre deux poches ou entre deux ligands.

Pour cette étude, nous avons d'abord étudié des poches protéiques de la MOAD, estimées par proximité (PockDrug) et par géométrie (FPocket). C'est la distance euclidienne pondérée sur les descripteurs qui nous donne la valeur de similarité. Lors de l'étude de la similarité entre

les poches, nous avons différencié les distances entre les poches dites « appariées » et « non-appariées ». Les premières estiment le même site de liaison d'une même protéine, elles se superposent géométriquement, mais l'une est estimée par proximité et l'autre par FPocket. Les poches « non-appariées » estiment quant à elles différents sites de liaison, peu importe la méthode d'estimation. La prise en compte des deux méthodes d'estimation est primordiale puisqu'il n'a pas été trouvé de solution absolue à l'estimation des poches et des sites de liaison. Les poches « appariées » permettent donc de prendre en compte la difficulté de l'estimation des poches. Elles permettent donc aussi de prendre en compte la flexibilité des protéines, montrée comme primordiale pour l'étude des interactions et la polypharmacologie. Une illustration des différences possibles entre les estimations de poches est extraite de l'étude de Borrel et al., (Borrel et al. 2015) en figure 7. C'est le site de liaison de l'inhibiteur de l'interleukine-1beta (code PDB : 1BMQ) estimé par 4 méthodes différentes : Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009), DoGsite (Volkamer et al. 2010) et la proximité (Borrel et al. 2015) à 4 Å et 5.5 Å, qui est illustré.

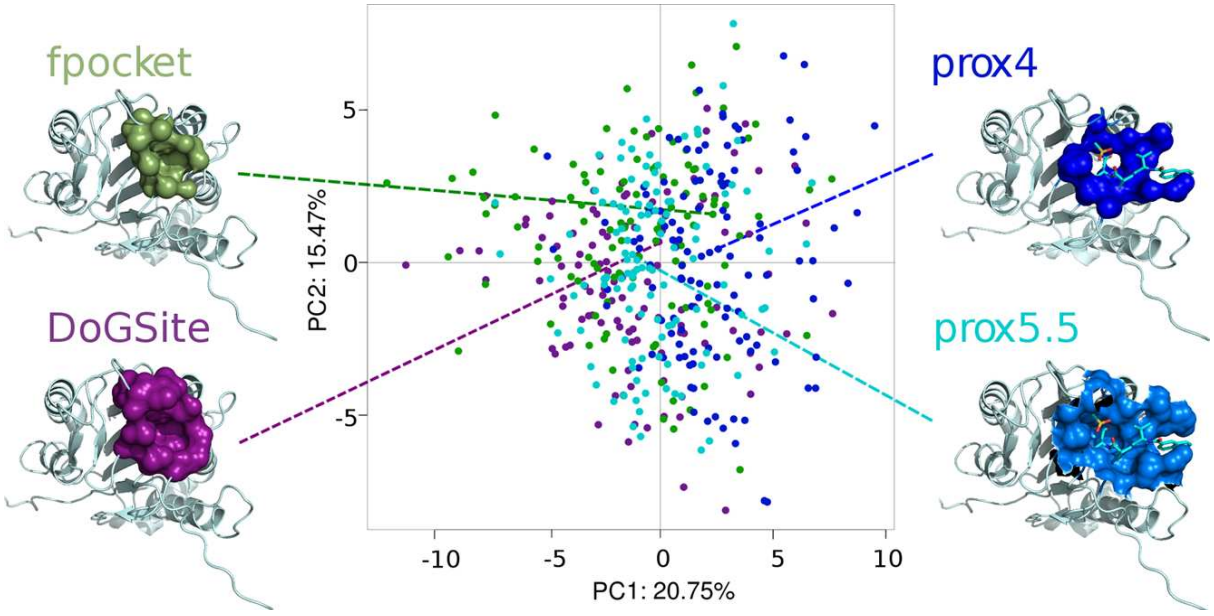


Figure 7 : Exemple de 4 poches « appariées ». Représentation de quatre ensembles de poche NRDLD (Krasowski et al. 2011) *estimés: prox4-NRDLD, prox5.5-NRDLD, fpocket-NRDLD et DoGSite-NRDLD sur le premier plan de l'ACP calculée à l'aide de l'ensemble de 52 descripteurs. Le premier plan de l'ACP explique plus de 35% de la variabilité. Les différentes poches estimées sont colorées en fonction de la méthode d'estimation utilisée: bleu pour prox4-NRDLD, bleu clair pour prox5.5-NRDLD, vert pour fpocket-NRDLD et violet pour DoGSite-NRDLD. Le site de liaison à l'interleukine-1 beta (code pdb : 1BMQ) estimé par les méthodes d'estimation à quatre poches est illustré.Figure tirée du papier de Borrel et al. (*(Borrel et al. 2015).

Comme le montre la figure 7, les différentes méthodes d'estimation donnent des poches qui peuvent différer en taille et en propriétés physico-chimiques, tout en partageant globalement le même espace physico-chimique. Cet exemple nous permet de mettre en avant l'importance de déterminer un seuil de distance qui permet de quantifier les variations possibles entre les méthodes d'estimation des poches. La flexibilité des structures des poches est aussi à prendre en compte lors des différentes estimations.

La similarité (en termes de distance) entre chacune des poches (appariées et non-appariées) calculée est illustrée dans la figure 8 (4 400 poches utilisées). Comme le montre cette figure, il y a une différence remarquable entre les distances des poches appariées et celles non-appariées. Il semblerait que la plupart des poches appariées aient une distance plus faible que la majorité des poches non-appariées, confirmant que l'estimation par Fpocket est fiable. Cette séparation entre les deux types de poches nous a donc servi de base pour déterminer le seuil de distance (ici 3.57) à partir duquel deux poches sont considérées comme similaires.

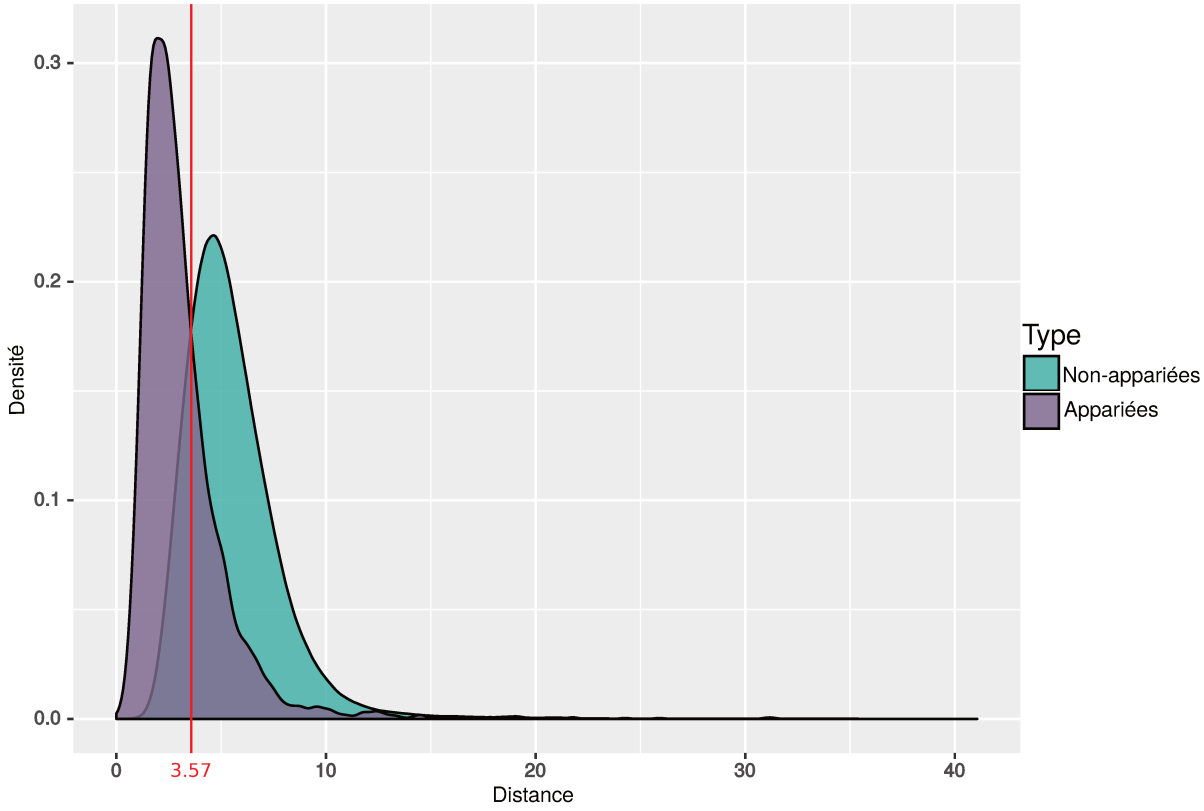


Figure 8 : Histogrammes superposés de la distance entre les poches selon leur type. Les poches appariées sont les poches d'une même protéine qui se superposent, dont une est estimée par proximité et l'autre par Fpocket. Les poches non-appariées sont les poches de protéines différentes (peu importe la méthode d'estimation). La ligne rouge représente la distance seuil choisie pour définir des poches comme similaires.

Une fois les poches similaires déterminées, on extrait leurs ligands qui composeront la chimiothèque. Dans notre cas, l'extraction consiste simplement à retenir le nom et la structure tridimensionnelle du ligand, puisque nous disposons de la structure 3D. Comme mentionné plus haut, si la poche d'intérêt est liée par des ligands connus, ils sont aussi intégrés à la chimiothèque. Cette chimiothèque préliminaire de ligands peut ensuite être enrichie avec des ligands similaires. Le calcul de similarité se fait grâce au coefficient de Tanimoto (cf. section *Materials and Method* de l'article présenté en chapitre 3) et (Rogers and Tanimoto 1960). Le seuil de similarité à partir duquel deux ligands sont considérés comme similaire est de 0,8.

Ces étapes amènent à la création de la chimiothèque de ligands. Les protocoles de docking pourraient être appliqués aux ligands de cette chimiothèque afin de valider la liaison potentielle

bien qu'aucun filtre n'ait été appliqué sur les ligands. Afin d'éliminer les ligands qui ne sont pas pertinents dans l'étude d'interactions médicamenteuses, une étape de réduction de la chimiothèque est nécessaire.

4.2.1.3 Réduction de la chimiothèque

Cette étape consiste à comparer les profils de ligands de la chimiothèque avec ceux contenus dans la base de données ZINC (Sterling and Irwin 2015). Elle regroupe plus de 230 millions de structures tridimensionnelles de composés commercialisés et prêtes à l'emploi, destinée aux protocoles de criblage virtuel et docking. Nous avons choisi de comparer les ligands de la chimiothèque avec une partie de la ZINC : les composés « en stock ». En effet, les composés non disponibles rapidement (sous 2 semaines) sont source d'incertitudes pour la suite du protocole et le filtrage permet de réduire les coûts de comparaison. Nous avons donc extrait 189 100 ligands, tous caractérisés avec RDkit (descripteurs 2D uniquement) afin d'avoir les mêmes descripteurs que les ligands de la PDB.

Seuls les ligands de la ZINC qui ont un profil similaire (coefficient de Tanimoto > 0,8) aux ligands de la chimiothèque constituent notre chimiothèque finale, aussi appelée « chimiothèque réduite » dans la figure 6.

4.2.1.4 Validation computationnelle des interactions par Docking

Cette partie du protocole n'est pas développée à ce stade et fait partie des perspectives du projet. Une fois la chimiothèque réduite établie, il est intéressant de valider de manière in silico les interactions proposées. Pour ce faire, il est prévu d'implémenter un protocole de docking afin d'éliminer les ligands dont la prédiction d'une pose n'est pas correcte et réduire le nombre de candidats potentiels tout en validant leur liaison.

Le principe du docking est d'identifier la conformation correcte d'une molécule dans la poche de liaison d'une protéine cible (dans notre cas) et d'estimer la force de cette liaison. De nos jours, une variété de programmes sont disponibles (Z. Wang et al. 2016) dont 3 des 10 les plus utilisés sont Autodock (Morris et al. 2009), Autodock Vina (Trott and Olson 2009), et Surflex-Dock (Jain 2003). Ils emploient des stratégies et des fonctions de score différentes et sont souvent comparés entre eux.

4.2.2 Résultats : Applications de prédiction des ligands

Bien que tout le protocole ne soit pas automatisé à ce jour, nous avons pu tester la partie automatisée pour la prédiction de ligands partenaires de deux cibles protéiques.

Le fonctionnement du programme de prédiction mis en place est schématisé dans la figure 9. Il prend en compte deux cas : celui dans lequel la structure est connue (n°1) et répertoriée sous un code PDB et celui dans lequel la structure n'est pas connue (n°2). Dans le cas n°1, le programme recherche directement dans la base de données de distances toutes les distances aux poches qui appartiennent au fichier PDB, s'il en existe. Un message averti l'utilisateur du programme dans l'éventualité où ce PDB ne contient pas de ligand et n'a donc pas de poche, ni de distance référencée dans la banque de données. Le programme lance alors une estimation par Fpocket et compare la (ou les) poche(s) la (ou les) plus druggable(s) à celles de la banque de données. Le nombre de plus grandes valeurs de drugabilité à prendre en compte (nommé *k)*

est déterminable par l'utilisateur. Par défaut, *k* est fixé à 1, c'est-à-dire que le programme sélectionnera toutes les poches Fpocket ayant la valeur maximum de drugabilité. Lorsqu'il est fixé à 2, le programme sélectionnera toutes les poches Fpocket ayant les deux plus grandes valeurs de drugabilité.

Dans le cas n°2, le programme ne connait pas le PDB en requête et estime donc les poches par proximité (s'il y a des ligands présents) et/ou par Fpocket. Les poches par proximité sont privilégiées pour la comparaison aux autres poches de la banque de données puisqu'elles permettent d'utiliser le(s) ligand(s) pour enrichir la chimiothèque avec des ligands similaires. Si la structure en requête ne comporte aucun ligand, ce sont donc les poches estimées par Fpocket qui seront comparées à la banque de données de poches.

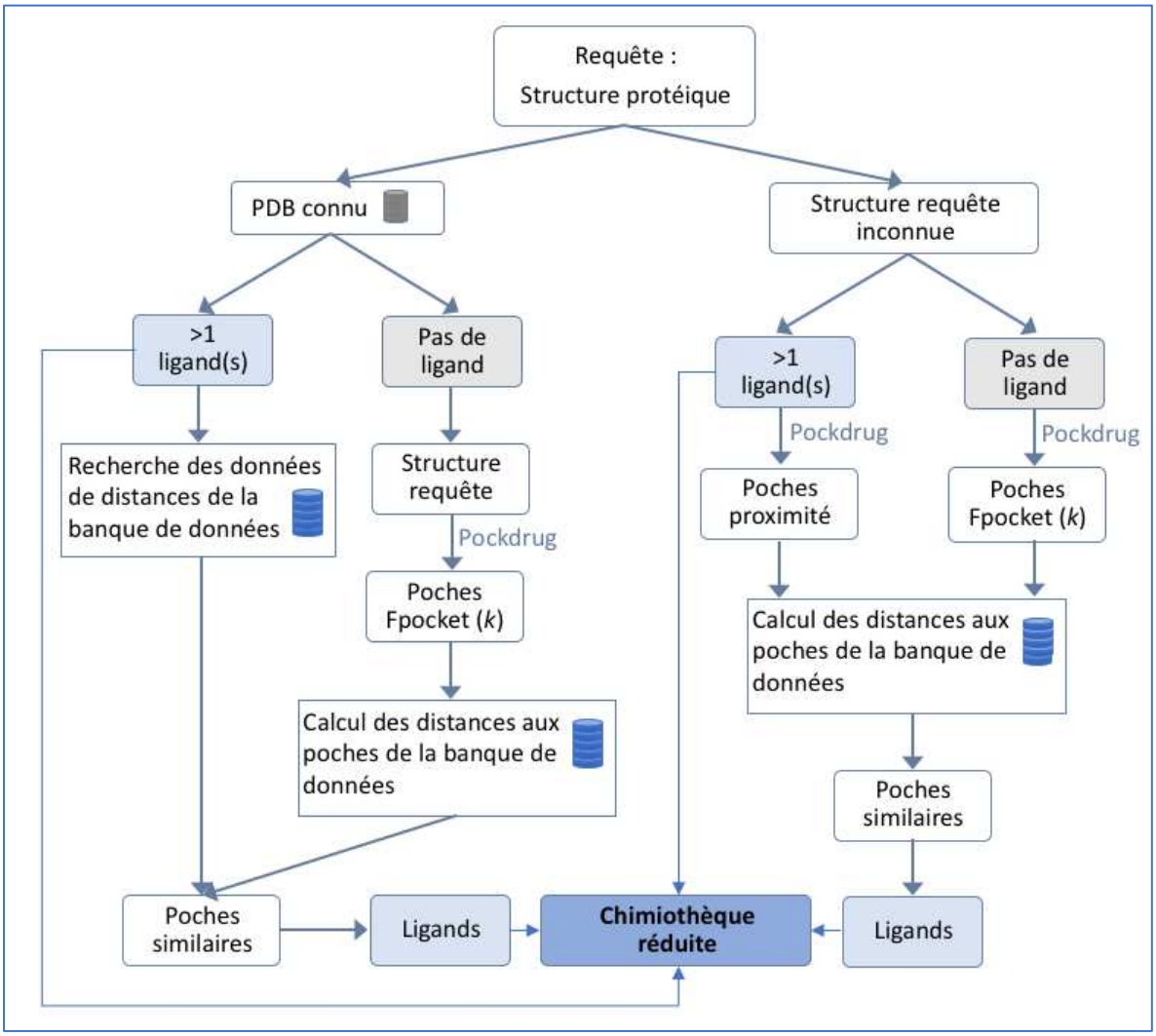


Figure 9 : Fonctionnement schématisé du programme de prédiction des interactions protéine-ligand (en date du 4 novembre 2019). À partir d'une structure requête, le programme fourni une chimiothèque réduite de ligands qui seront utilisés pour enrichir la chimiothèque, la comparer aux composés de la Zinc et appliquer le protocole de docking.

Le seuil déterminé en figure 8 et section 4.2.1.2 (3.57) permet de filtrer les poches proches. Dans les deux cas, les ligands des poches proches trouvées constituent la base de la chimiothèque réduite. Les ligands similaire (Tanimoto >0,8) de ceux qui composent la chimiothèque réduite constituent une chimiothèque plus large qui sera comparé à la Zinc et utilisé pour le protocole de docking.

4.2.2.1 Test sur les Urokinases

La première cible testée est celle des urokinases. En effet, elle nous sert de validation puisque il s'agit d'une cible bien étudiée avec de nombreux ligands liés connus. Il est donc attendu de retrouver les interactions étudiées dans l'étude présentée en chapitre 2.

Les structures tridimensionnelles utilisées sont les suivantes : 1GJ7, 1SQO, 1O3P et 4MNW. La première est la structure de référence désignée lors de l'étude, et les 3 autres sont des représentants des clusters de poches trouvés lors de la caractérisation conjointe des interactions, la dernière étant liée à un peptide. Pour ce test, nous avons utilisé les poches prédites par Fpocket et non par proximité afin de simuler une protéine non présente dans la banque de données initiale.

Pour chacune des 4 structures (1GJ7, 1SQO, 1O3P et 4MNW), on retrouve un certain nombre de poches similaires appartenant à la même protéine, respectivement 18, 39, 25 et 25 (avec des distances comprises entre 0,9 et 1,5). Bien que toutes les poches des urokinases ne soient pas retrouvées, ces résultats sont très encourageants et nous servent de « preuve de concept » pour la suite des applications. Les ligands consentis pour la chimiothèque seront donc les ligands des urokinases.

4.2.2.2 Test sur la grippe NS1

Le protocole partiel a aussi été testé sur une protéine du virus de la grippe, appelée NS1 dont la structure et les domaines sont représentés en figure 10. Il s'agit d'une protéine qui permet notamment de bloquer les défenses antivirales de la cellule infectée ce qui favorise la réplication du virus (Marc 2012). Cette protéine est donc particulièrement d'intérêt dans le développement de médicaments contre la grippe.

Nous avons choisi de tester le domaine RBD avec notre protocole en raison de la bibliographie plus fournie que pour le domaine ED. Aucune poche par proximité n'a pu être estimée puisque c'est une structure non liée. Cette application est donc exploratoire.

L'estimation des poches a permis d'extraire 8 poches. C'est la poche estimée comme la plus probable qui a retenu notre attention pour la suite du test. En effet, elle présente un intérêt majeur par rapport aux autres poches puisqu'elle contient 2 des 3 résidus répertoriés comme « impliqués dans l'interaction de l'ARN avec NS1 » (Marc 2012) qui sont les résidus arginine 38 (R28) et sérine 42 (S42). Cette poche est présentée en figure 11.

Cette poche semble correspondre au site de liaison à l'ARN du domaine. L'analyse de ses propriétés physicochimiques a permis de la comparer aux autres poches de notre banque de données structurales (version sur la PDB).

Les 4 poches les plus proches ainsi que leur ligand et leur rôle sont regroupés dans la table 3 suivante. Elles correspondent à des poches très similaires de distances comprises entre 0,94 et 1,04. Les poches qui correspondent sont représentées en figure 12.

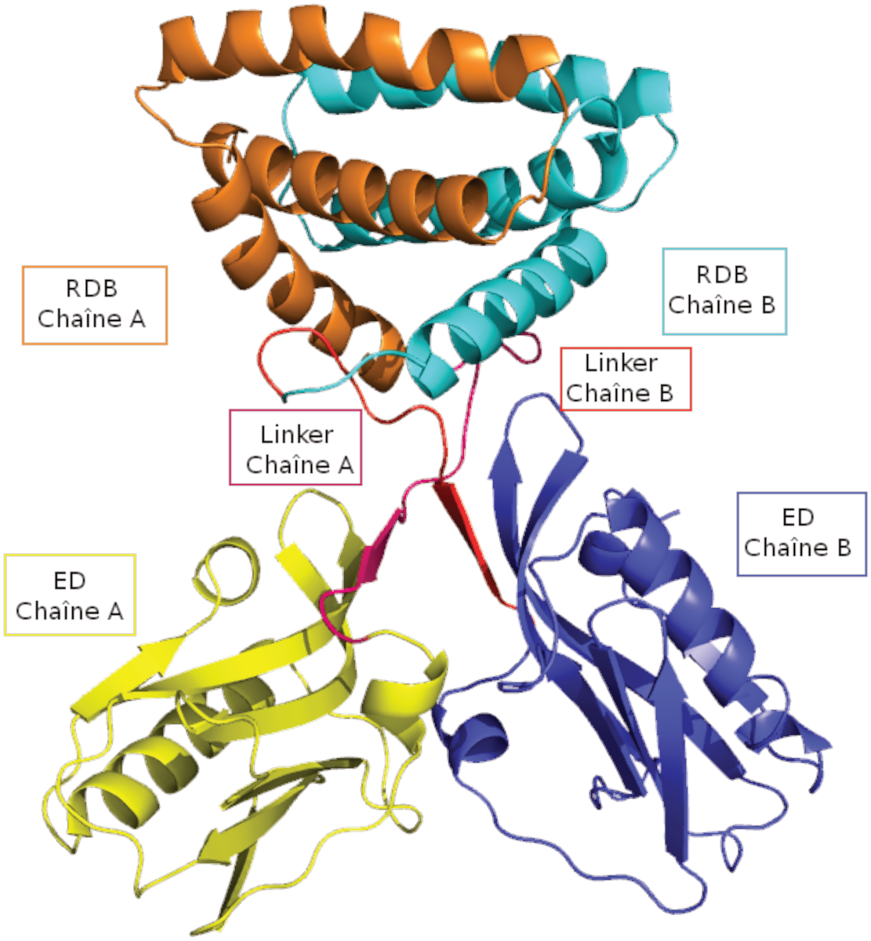


Figure 10 : Structure tridimensionnelle de l'homodimère NS1 issue de la souche H6N6 du virus Influenza A (PDB 4OPH). Chaque domaine de NS1 de chaque chaîne est représenté par une couleur (RBD chaîne A en orange et ED chaîne A en jaune, RBD chaîne B en cyan et ED chaîne B en bleu). Le « linker » d'un monomère est représenté en rose pour la chaîne A et en rouge pour la chaîne B.

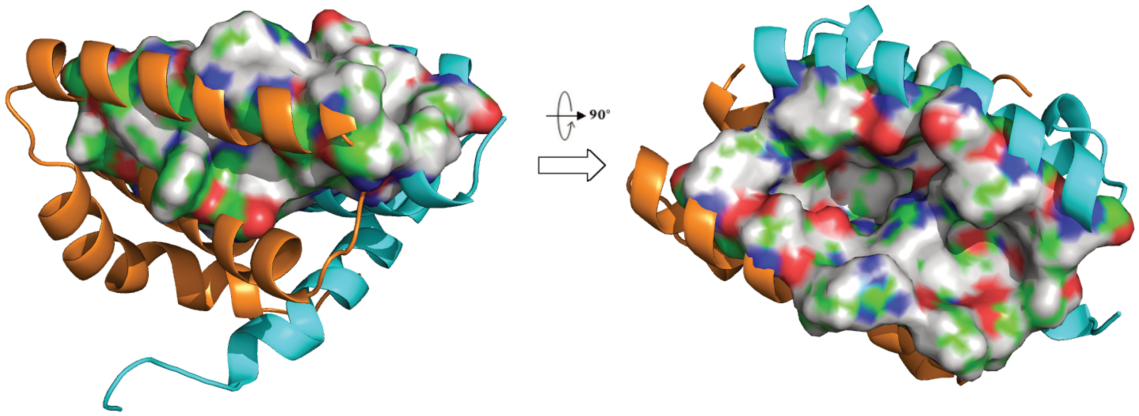


Figure 11 : Représentation du domaine RBD de la protéine NS1 (en cartoon) et de sa poche estimée (en surface) comme la plus probable par PockDrug à l'aide de PyMOL. La chaîne A est colorée en orange et la chaîne B en cyan. La poche est colorée en fonction de la nature des atomes qui la composent : les carbones en vert, les hydrogènes en blanc, l'azote en bleu et l'oxygène en rouge. Une rotation de 90° vers le dessus de la poche nous permet de visualiser la profondeur de la poche et la nature des atomes de la cavité.

La Table 3 nous indique que les poches proches sont issues de protéines de nature très diverse et impliquées dans des rôles différents. Aucune protéine en rapport avec le virus de la grippe ne se trouve parmi les poches les plus proches. La visualisation des poches en 3D nous montre qu'elles sont de formes très différentes. La poche de la protéine 3G8E est cylindrique, le ligand est peu accessible, tandis que celle de la protéine 2V50 est très plate et ouverte, le ligand est très accessible. Les poches des deux dernières protéines sont de géométries intermédiaires.

Table 3 : Liste des poches les plus proches obtenues selon le protocole de prédiction. Le nom de la protéine, celui du ligand, la distance à la poche de NS1 et le rôle de la protéine sont indiqués. Les distances ayant une étoile (*) signifient que d'autres poches de cette même protéine ont été retrouvées comme similaires à la poche de NS1 (avec une valeur de distance plus grande), la poche la plus proche est détaillée.

Protéine (Code PDB)	Ligand (Code HET)	Distance à NS1	Rôle connu
Vistafine (3G8E)	FK866 (IS1)	0,94*	Biosynthèse du NAD+
MexB (2V50)	Dodecyl-α-D-Maltoside (LMT)	0,95	Transport transmembranaire (résistance multi-médicament)
MurD (2X5O)	(R)-32 (VSV)	1,00*	Biogénèse de la paroi cellulaire
Tubuline (4X1Y)	PRD_002154 (3WV)	1,04	Constituant des microtubules (cytosquelette)

Les ligands (IS1, LMT, VSV et 3WV), dont la structure 2D est représentée en figure 12, sont les premiers candidats de notre chimiothèque réduite. Ils ont globalement une forme semblable entre eux (calcul de similarité en cours) et semblent correspondre à certains antagonistes obtenus par des approches de docking décrit par la littérature (Engel 2013) (La récupération de ces ligands en cours) : ils sont polycycliques et de grande taille.

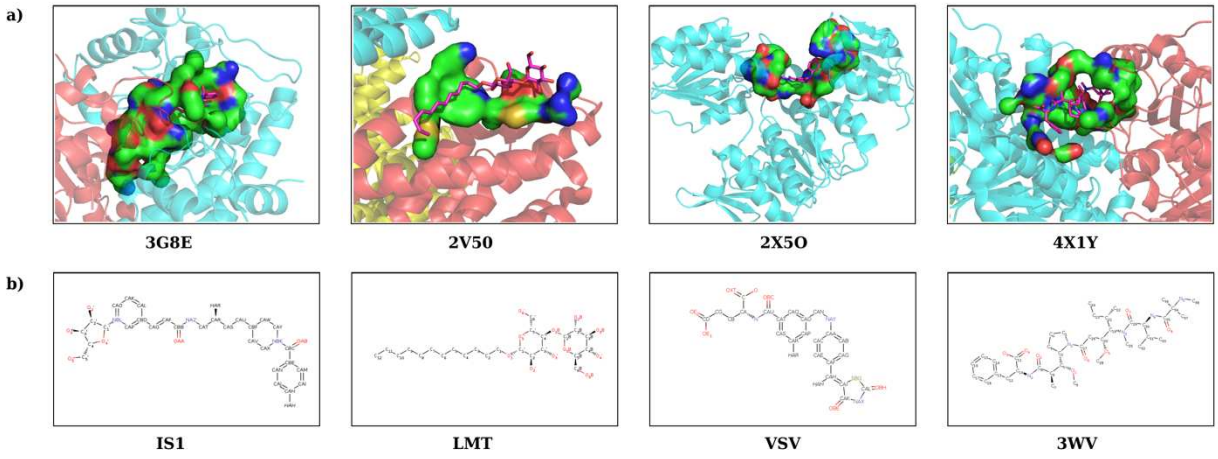


Figure 12 : a) Représentation des poches estimées comme les plus proches de la poche de NS1. Les protéines sont représentées en cartoon et colorées selon leur chaînes, les poches en surface et colorées selon les atomes qui les composent (carbones en vert) et les ligands en sticks, colorés selon les atomes qui les composent (carbones en magenta). b) Représentation (fournie par la PDB) en 2D des ligands correspondants, leur nom (code HET de la PDB) est indiqué en dessous.

4.3 Perspectives

Les méthodes développées permettent de caractériser les profils physicochimiques et géométriques des sites de liaison, ceux des ligands et leur correspondance statistique. Elle permettent aussi de déterminer leur promiscuité et de caractériser les principales types d'interactions associés.

Comme mentionné plus haut, cet outil est en cours de développement, les perspectives sont donc nombreuses. En premier il est nécessaire d'automatiser la fin du protocole, c'est-à-dire le calcul de la similarité des ligands, la comparaison avec la ZINC. Enfin des méthodes de docking seront ensuite appliquées.

Ensuite, les stages de Mlle Kughatas et M. Bouvarel ont permis d'explorer de nouvelles méthodes qui peuvent être intégrées au protocole. Les pharmacophores peuvent améliorer la description des poches en prenant en compte une certaine flexibilité des poches, cette notion étant importante lors des simulations de dynamique moléculaire et lorsque la protéine est sujette à des changements de conformations. Les aspects de *Deep Learning* sont aussi très prometteurs en étant rapides. La prédiction de la drugabilité, de la promiscuité et des interactions en utilisant les descripteurs et les pharmacophore est donc possible.

Enfin, à plus long terme la mise à disposition de l'outil pour la communauté scientifique est la perspective dans l'optique d'optimiser les processus de découverte des médicaments, de réduire les effets secondaires des molécules médicamenteuses et de proposer des molécules pour le recyclage de médicaments.

CONCLUSION ET PERSPECTIVES

De nos jours, une des principales difficultés dans la conception de médicaments est d'éviter les effets secondaires néfastes, dus à l'interaction du médicament avec une cible inattendue, et de réduire les coûts de développement en réutilisant des médicaments existants et déjà commercialisés qui ont un effet bénéfique autre que celui pour lequel ils sont indiqués.

La base de ces effets secondaires (bénéfiques autant que néfastes) est donc l'interaction entre le médicament et au moins une autre protéine que celle ciblée. La découverte de la capacité à avoir plusieurs partenaires (promiscuité) a invalidé le concept clé–serrure développé au début du XXe siècle et remis fortement en question la sélection des molécules candidates dans le développement de médicaments. D'autres modèles ont été décrits, comme le modèle de Koshland, et permettent de mieux comprendre les mécanismes d'interactions, primordiaux pour le développement de médicaments. La notion de flexibilité dans les modèles est devenue incontournable et doit être prise en compte lors de l'étude des interactions.

Les méthodes computationnelles récentes s'attellent donc à étudier les interactions, en particulier celles impliquant plusieurs partenaires, en profitant largement de l'augmentation du nombre de données de bonne qualité, notamment en termes de structures protéiques et en proposant de nombreux outils aux méthodes et objectifs divers et variés.

Dans ce contexte, l'objectif de mon projet de thèse consistait à modéliser et caractériser de manière optimale les interactions entre les protéines et les ligands afin de développer une méthode de prédiction des interactions. Le but à plus long terme étant d'améliorer le processus de sélection des molécules pour la découverte de médicaments.

Pour atteindre cet objectif, il a été nécessaire de développer une méthode capable de caractériser une interaction entre une protéine et un ligand ayant des propriétés semblables aux molécules médicamenteuses. Il existe de nombreuses méthodes permettant la caractérisation précise de chaque acteur de l'interaction, dont la description physicochimique et géométrique à l'aide de variables appelées descripteurs. De nombreux descripteurs ayant été développés au sein de l'équipe, c'est par cette méthode qu'ont été décrits les sites de liaison et les ligands de mes travaux.

Pour entraîner la caractérisation des sites de liaison et des ligands, le site catalytique de l'inhibiteur de l'urokinase a été pris comme exemple. Son implication dans les voies de signalisation de l'inflammation et le fait qu'elle est multipartenaire (aussi dit promiscuous) en fait une cible thérapeutique d'intérêt. Ces travaux de caractérisation ont d'abord mis en évidence une diversité des ligands se liant au même site de liaison, les poches estimées par proximité à ces ligands sont donc aussi diverses. Permettant de différencier les profils, 9 descripteurs de poches et 9 autres descripteurs de ligands sont ressortis comme informatifs. Ils concernent l'hydrophobicité, la charge et la taille des protéines et le poids moléculaire, la solubilité et la flexibilité des ligands. La caractérisation conjointe a permis de faire ressortir 5

profils types de complexes. Des caractéristiques variées pour les poches et pour les ligands ont été mises en avant, même au sein d'une unique famille de protéines. La pertinence de notre méthode d'estimation et de caractérisation des poches pour la description d'un unique site de liaison a été démontrée une fois de plus. Combinées à la classification de ces profils, les différentes méthodes forment notre protocole protéochémométrique.

Nos résultats confirment que ce protocole peut établir des correspondances statistiques entre les profils de poches et de ligands. Cette correspondance de profils poche-ligand résultante peut être utilisée pour évaluer l'affinité de certains profils de ligands pour certains profils de poches et inversement. Cette analyse permet de détecter les modifications des conformations 3D, dues aux mutations ou à des différences conformationnelles provenant de la liaison avec un ligand, autre protéine, un ARN ou aucune molécule. D'autres données pourraient venir enrichir cette approche comme l'utilisation de pharmacophores ou les valeur d'activité. Ces travaux sont l'étude préliminaire à la prédiction des interactions. En effet, comprendre les mécanismes d'interaction entre les ligands et les protéines est primordial pour la suite du développement des méthodes de prédiction. Ces travaux ont donné lieu à une publication.

Lors d'une seconde étude, les méthodes de caractérisation conjointe des profils utilisées précédemment ont été appliquées sur un nombre élevé de complexes d'une banque de données d'interactions de haute qualité dans le but de caractériser la promiscuité des sites de liaison et d'établir leur rôle dans la polypharmacologie. Notre attention s'est portée sur les interactions multipartenaires et les profils ont été classés selon leur capacité à interagir avec un ou plusieurs partenaires. Les résultats de cette étude sont nombreux.

Il a été mis en avant la fréquence élevée (80 %) des sites de liaison promiscuous, dont plus de la moitié sont liés par plus de 4 ligands différents. Les poches qui décrivent ces sites de liaison sont de grande taille et hydrophobes. Les 20 % des sites de liaison restants sont appelés sélectifs, et tendent à être moins favorables aux interactions en ayant des propriétés opposées. Concernant les ligands, un résultat tributaire du choix de la MOAD comme base du jeu de données a été mis en évidence : seuls 18 % des ligands sont promiscuous. Cette faible fréquence peut-être due en partie au fait que les molécules ne sont pas testées sur d'autres protéines que celles visées, les informations d'interaction sont donc manquantes, mais aussi due au filtrage de qualité appliqué aux données. Globalement, une majorité des sites de liaison interagissent donc avec plusieurs ligands qui, eux, ne lient en majorité qu'un type de sites de liaison. Aussi, ces travaux ont montré que la promiscuité des sites de liaison n'était pas dépendante de la famille de protéine traitée.

Cette seconde étude a mis en exergue l'importance de la promiscuité des sites de liaison dans la polypharmacologie et l'efficacité de la caractérisation conjointe des poches et des ligands (une publication a validé cette étude). La prise en compte de la promiscuité des sites de liaison devrait contribuer plusieurs améliorations. Tout d'abord, elle pourrait expliquer la promiscuité du médicament ou de détection de *« off-target »* de la cible, ce qui conduit aux effets secondaires souvent néfastes et qui ralentissent énormément les progrès en matière de développement de médicaments. Comprendre l'origine des interactions non voulues serait une

aide précieuse lors de l'optimisation des molécules en vue de leur tests *in vivo*. Aussi, elle pourrait présenter des hypothèses de repositionnement, souvent considéré comme offrant d'importantes opportunités pour la recherche et le développement de médicaments en réduisant les coûts et le temps de développement.

Le protocole protéochémométrique et l'étude de la promiscuité ont été déterminants dans l'étude des interactions protéine–ligand et ont permis d'entamer le développement d'un protocole complexe de prédiction de ces interactions. Ils ont notamment mis en évidence les propriétés déterminantes lors des interactions (dont la flexibilité), celles qui caractérisent la promiscuité (hydrophobicité) et ont permis d'orienter certains choix et critères pour le développement de l'outil de prédiction.

À partir d'une structure 3D de protéine « requête » (liée à un ligand ou non) le protocole de prédiction dont le développement est en cours estime ses poches de liaison, les caractérise, et les compare à toutes les poches liées d'une banque de données préalablement construite. Les ligands des poches similaires (et les ligands de la protéine s'il en existe) sont extraits et comparés aux ligands d'une banque de données de composés commercialisés. Ainsi, une chimiothèque de composés est proposée. Le développement de ces parties est partiellement réalisé. La suite du protocole devrait permettre de proposer une chimiothèque sélectionnée pour appliquer ensuite des approches de docking des composés sur la protéine requête, afin d'estimer in silico l'affinité et les énergies de liaison des interactions.

Le protocole de prédiction, en développement, a pu être testé sur deux protéines d'intérêt thérapeutique, l'urokinase et la protéine NS1 de la grippe. Les résultats ont montré sur l'urokinase que le programme est capable de retrouver des poches similaires de la même protéine et fait office de preuve de concept. La structure de NS1, issue de dynamique moléculaire et sans ligand est encore peu documentée. Les poches similaires obtenues sont issues de protéines très différentes de celle de la grippe et leurs ligands semblent correspondre à ceux obtenus par docking décrits dans la bibliographie sur la protéine NS1. C'est donc une piste à explorer plus méticuleusement lorsque le programme sera terminé. Pour ces derniers travaux, des perspectives ont déjà été pressenties, comme l'utilisation de pharmacophores pour la description des poches et des ligands et des méthodes plus poussées de *Deep Learning*.

Il est évident que ce programme de prédiction des interactions est très ambitieux et qu'il reste plusieurs étapes à améliorer voire finir : l'intégration des pharmacophores qui ont démontré leur pertinence, l'automatisation de la recherche de ligands similaires et enfin le protocole de docking à mettre en place et à tester.

En conclusion de ces différents travaux de thèse, l'importance de la caractérisation conjointe et de la prise en compte de la promiscuité des sites de liaison ont été soulignées. Malgré les difficultés liées aux fichiers PDB non-rigoureux, aux méthodes d'estimations des poches et aux choix des descripteurs ou des seuils les plus pertinents, le protocole protéochémométrique et l'étude de la promiscuité me semblent pertinents et une base satisfaisante pour la suite des travaux de prédiction. Le protocole de prédiction développé durant ma thèse, ainsi que les nombreuses autres permettant la prédiction d'interactions entre

les protéines et les ligands, sont prometteuses et jouent un rôle important dans l'avancée sur la compréhension des mécanismes des effets secondaires et de la recherche de nouveaux médicaments.

BIBLIOGRAPHIE

Ahmed, Aqeel, Richard D. Smith, Jordan J. Clark, James B. Dunbar, Jr Carlson, and Heather A. Carlson. 2015. "Recent Improvements to Binding MOAD: A Resource for Protein–Ligand Binding Affinities and Structure." *Nucleic Acids Research* 43 (D1): D465–69. https://doi.org/10.1093/nar/gku1088.

Andrade, E L, A F Bento, J Cavalli, S K Oliveira, R C Schwanke, J M Siqueira, C S Freitas, R Marcon, and J B Calixto. 2016. "Non-Clinical Studies in the Process of New Drug Development – Part II : Good Laboratory Practice , Metabolism , Pharmacokinetics , Safety and Dose Translation to Clinical Studies" 49: 1–19. https://doi.org/10.1590/1414-431X20165646.

Anighoro, Andrew, Jürgen Bajorath, and Giulio Rastelli. 2014. "Polypharmacology: Challenges and Opportunities in Drug Discovery." *Journal of Medicinal Chemistry* 57 (19): 7874–87. https://doi.org/10.1021/jm5006463.

ANSM. 2018. "Rapport d'activité."

Arunan, Elangannan, Gautam R. Desiraju, Roger A. Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C. Clary, et al. 2011. "Definition of the Hydrogen Bond (IUPAC Recommendations 2011)." *Pure and Applied Chemistry* 83 (8): 1637–41. https://doi.org/10.1351/pac-rec-10-01-02.

Ashburn, Ted T., and Karl B. Thor. 2004. "Drug Repositioning: Identifying and Developing New Uses for Existing Drugs." *Nature Reviews Drug Discovery* 3 (8): 673–83. https://doi.org/10.1038/nrd1468.

Barak, Nir. 2008. "Betahistine: What's New on the Agenda?" *Expert Opinion on Investigational Drugs* 17 (5): 795–804. https://doi.org/10.1517/13543784.17.5.795.

Barelier, Sarah, Teague Sterling, Matthew J. O'Meara, and Brian K. Shoichet. 2015. "The Recognition of Identical Ligands by Unrelated Proteins." *ACS Chemical Biology* 10 (12): 2772–84. https://doi.org/10.1021/acschembio.5b00683.

Berman, Helen M., Tammy Battistuz, T. N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, et al. 2002. "The Protein Data Bank." *Acta Crystallographica Section D Biological Crystallography* 58 (6): 899–907. https://doi.org/10.1107/S0907444902003451.

Bisgin, Halil, Zhichao Liu, Hong Fang, Reagan Kelly, Xiaowei Xu, and Weida Tong. 2014. "A Phenome-Guided Drug Repositioning through a Latent Variable Model." *BMC Bioinformatics* 15 (1): 1–12. https://doi.org/10.1186/1471-2105-15-267.

Borrel, Alexandre, Leslie Regad, Henri Xhaard, Michel Petitjean, and Anne-Claude Claude Camproux. 2015. "PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties." *Journal of Chemical Information and Modeling* 55 (4): 882–95. https://doi.org/10.1021/ci5006004.

Borrotti, Matteo, Davide De March, Debora Slanzi, and Irene Poli. 2014. "Designing Lead Optimisation of MMP-12 Inhibitors" 2014. https://doi.org/10.1155/2014/258627.

Bosc, Nicolas, Berthold Wroblowski, Samia Aci-Sèche, Christophe Meyer, and Pascal Bonnet. 2015. "A Proteometric Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues." *ACS Chemical Biology* 10 (12): 2827–40. https://doi.org/10.1021/acschembio.5b00555.

Buckley, Benjamin J, Michael J Kelso, Umar Ali, and Marie Ranson. 2018. "The Urokinase Plasminogen Activation System in Rheumatoid Arthritis: Pathophysiological Roles and Prospective Therapeutic Targets." *Current Drug Targets* 20: 970–81. https://doi.org/10.2174/1389450120666181204164140.

Buyse, Marc. 2016. "Phase III Design : Principles." *Chin Clin Oncol* 5 (1): 1–13. https://doi.org/10.3978/j.issn.2304-3865.2014.08.05.

Campillos, Monica, Michael Kuhn, Anne-claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. "Drug Target Identification Using Side-Effect Similarity" 321 (July): 263–67.

Chaudhari, Rajan, Zhi Tan, Beibei Huang, and Shuxing Zhang. 2017. "Computational Polypharmacology: A New Paradigm for Drug Discovery." *Expert Opinion on Drug Discovery* 12 (3): 279–91. https://doi.org/10.1080/17460441.2017.1280024.

Choy, Young Bin, and Mark R. Prausnitz. 2011. "The Rule of Five for Non-Oral Routes of Drug Delivery: Ophthalmic, Inhalation and Transdermal." *Pharmaceutical Research* 28 (5): 943–48. https://doi.org/10.1007/s11095-010-0292-6.

Cochran, Blake J., David R. Croucher, Sergei Lobov, Darren N. Saunders, and Marie Ranson. 2011. "Dependence on Endocytic Receptor Binding via a Minimal Binding Motif Underlies the Differential Prognostic Profiles of SerpinE1 and SerpinB2 in Cancer." *Journal of Biological Chemistry* 286 (27): 24467–75. https://doi.org/10.1074/jbc.M111.225706.

Congreve, Miles, Robin Carr, Chris Murray, and Harren Jhoti. 2003. "A 'Rule of Three' for Fragment-Based Lead Discovery?" *Drug Discovery Today* 8 (19): 876–77. https://doi.org/10.1016/S1359-6446(03)02831-9.

Croucher, David R., Darren N. Saunders, Sergei Lobov, and Marie Ranson. 2008. "Revisiting the Biological Roles of PAI2 (SERPINB2) in Cancer." *Nature Reviews Cancer* 8 (7): 535–45. https://doi.org/10.1038/nrc2400.

Desaphy, Jérémy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. 2015. "Sc-PDB: A 3D-Database of

Ligandable Binding Sites-10 Years On." *Nucleic Acids Research* 43 (D1). https://doi.org/10.1093/nar/gku928.

Desaphy, Jérémy, Eric Raimbaud, Pierre Ducrot, and Didier Rognan. 2013. "Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs." *Journal of Chemical Information and Modeling* 53 (3): 623–37. https://doi.org/10.1021/ci300566n.

Dessailly, Benoit H., Marc F. Lensink, Christine A. Orengo, and Shoshana J. Wodak. 2008. "LigASite - A Database of Biologically Relevant Binding Sites in Proteins with Known Apo-Structures." *Nucleic Acids Research* 36 (SUPPL. 1): 667–73. https://doi.org/10.1093/nar/gkm839.

Diharce, Julien, Mickaël Cueto, Massimiliano Beltramo, Vincent Aucagne, and Pascal Bonnet. 2019. "In Silico Peptide Ligation: Iterative Residue Docking and Linking as a New Approach to Predict Protein-Peptide Interactions." *Molecules (Basel, Switzerland)* 24 (7). https://doi.org/10.3390/molecules24071351.

Dougherty, Dennis A. 2013. "The Cation-π Interaction" 46 (4): 885–893. https://doi.org/10.1021/ar300265y.

Du, Xing, Yi Li, Yuan Ling Xia, Shi Meng Ai, Jing Liang, Peng Sang, Xing Lai Ji, and Shu Qun Liu. 2016. "Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods." *International Journal of Molecular Sciences* 17 (2): 1–34. https://doi.org/10.3390/ijms17020144.

Ehrt, Christiane, Tobias Brinkjost, and Oliver Koch. 2018. "A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs)." *PLoS Computational Biology* 14 (11): 1–50. https://doi.org/10.1371/journal.pcbi.1006483.

Ekins, Sean. 2004. "Predicting Undesirable Drug Interactions with Promiscuous Proteins in Silico" 9 (6): 276–85.

Feder, Gene. 1993. "Paradigm Lost: A Celebration of Paracelsus on His Quincentenary." *The Lancet* 341: 1396–98.

Fogel, David B. 2018. "Factors Associated with Clinical Trials That Fail and Opportunities for Improving the Likelihood of Success : A Review." *Contemporary Clinical Trials Communications* 11 (August): 156–64. https://doi.org/10.1016/j.conctc.2018.08.001.

Franco, Paola, Ciro Iaccarino, Ferdinando Chiaradonna, Anna Brandazza, Carlo Iavarone, M. Rosaria Mastronicola, M. Luisa Nolli, and M. Patrizia Stoppelli. 1997. "Phosphorylation of Human Pro-Urokinase on Ser138/303 Impairs Its Receptor-Dependent Ability to Promote Myelomonocytic Adherence and Motility." *Journal of Cell Biology* 137 (3): 779–91. https://doi.org/10.1083/jcb.137.3.779.

Gao, Mu, and Jeffrey Skolnick. 2013. "APoc: Large-Scale Identification of Similar Protein Pockets." *Bioinformatics* 29 (5). https://doi.org/10.1093/bioinformatics/btt024.

Gold, Nicolas D., and Richard M. Jackson. 2005. "SitesBase: A Database for Structure-Based Protein-Ligand Binding Site Comparisons." *Nucleic Acids Research* 34: D231–34. https://doi.org/10.1093/nar/gkj062.

Goldstein, Irwin, Arthur L. Burnett, Raymond C. Rosen, Peter W. Park, and Vera J. Stecher. 2019. "The Serendipitous Story of Sildenafil: An Unexpected Oral Therapy for Erectile Dysfunction." *Sexual Medicine Reviews* 7 (1): 115–28. https://doi.org/10.1016/j.sxmr.2018.06.005.

Gong, Jiayu, Chaoqian Cai, Xiaofeng Liu, Xin Ku, Hualiang Jiang, Daqi Gao, and Honglin Li. 2013. "ChemMapper: A Versatile Web Server for Exploring Pharmacology and Chemical Structure Association Based on Molecular 3D Similarity Method." *Bioinformatics* 29 (14): 1827–29. https://doi.org/10.1093/bioinformatics/btt270.

Guilloux, Vincent Le, Peter Schmidtke, and Pierre Tuffery. 2009. "Fpocket: An Open Source Platform for Ligand Pocket Detection." *BMC Bioinformatics* 10: 168. https://doi.org/10.1186/1471-2105-10-168.

Hajduk, Philip J., Jeffrey R. Huth, and Stephen W. Fesik. 2005. "Druggability Indices for Protein Targets Derived from NMr-Based Screening Data." *Journal of Medicinal Chemistry* 48 (7): 2518–25. https://doi.org/10.1021/jm049131r.

Halgren, THomas A: 2009. "Identifying and Characterizing Binding Sites and Assessing Druggability." *Journal of Chemical Information and Modeling* 49 (2): 377–89.

Harrison, Richard K. 2016. "Phase II and Phase III Failures : 2013 – 2015." *Nature Reviews Drug Discovery* 15 (12): 817–18. https://doi.org/10.1038/nrd.2016.184.

Haupt, V. Joachim, Simone Daminelli, and Michael Schroeder. 2013. "Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key." *PLoS ONE* 8 (6). https://doi.org/10.1371/journal.pone.0065894.

Hodos, Rachel A, Brian A Kidd, Shameer Khader, Ben P Readhead, and Joel T Dudley. 2016. "Computational Approaches to Drug Repurposing and Pharmacology." *Wiley Interdiscip Rev Syst Biol Med.* 8 (3): 186–210. https://doi.org/10.1002/wsbm.1337.

Hopkins, Andrew L., György M. Keserü, Paul D. Leeson, David C. Rees, and Charles H. Reynolds. 2014. "The Role of Ligand Efficiency Metrics in Drug Discovery." *Nature Reviews Drug Discovery* 13 (2). https://doi.org/10.1038/nrd4163.

Hu, Liegi, Mark L. Benson, Richard D. Smith, Michael G. Lerner, and Heather A. Carlson. 2005. "Binding MOAD (Mother Of All Databases)." *Proteins: Structure, Function, and Bioinformatics* 60 (3): 333–40. https://doi.org/10.1002/prot.20512.

Hu, Ye, and Jürgen Bajorath. 2013. "Compound Promiscuity: What Can We Learn from Current Data?" *Drug*

Discovery Today 18 (13–14): 644–50. https://doi.org/10.1016/j.drudis.2013.03.002.
———. 2015. "Quantifying the Tendency of Therapeutic Target Proteins to Bind Promiscuous or Selective Compounds." *PLoS ONE* 10 (5). https://doi.org/10.1371/journal.pone.0126838.
Huang, Ying, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. 2010. "CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences." *Bioinformatics (Oxford, England)* 26 (5): 680–82. https://doi.org/10.1093/bioinformatics/btq003.
Hussein, Hiba Abi, Alexandre Borrel, Colette Geneix, Michel Petitjean, Leslie Regad, and Anne-Claude Claude Camproux. 2015. "PockDrug-Server: A New Web Server for Predicting Pocket Druggability on Holo and Apo Proteins." *Nucleic Acids Research* 43 (W1): W436-42. https://doi.org/10.1093/nar/gkv462.
Hussein, Hiba Abi, Colette Geneix, Michel Petitjean, Alexandre Borrel, Delphine Flatters, and Anne-Claude Camproux. 2017. "Global Vision of Druggability Issues: Applications and Perspectives." *Drug Discovery Today* 22 (2): 404–15. http://www.ncbi.nlm.nih.gov/pubmed/27939283.
Ianiro, G., F. Mangiola, T.A. Di Rienzo, S. Bibbò, F. Franceschi, A.V. Greco, and A. Gasbarrini. 2014. "Levothyroxine Absorption in Health and Disease , and New Therapeutic Perspectives." *European Review for Medical and Pharmacological Sciences* 18: 451–56.
Jain, Ajay N. 2003. "Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine." *Journal of Medicinal Chemistry* 46 (4): 499–511. https://doi.org/10.1021/jm020406h.
Kimes, Patrick K., Yufeng Liu, David Neil Hayes, and James Stephen Marron. 2017. "Statistical Significance for Hierarchical Clustering." *Biometrics* 73 (3): 811–21. https://doi.org/10.1111/biom.12647.
Knobloch, J., D. Jungck, and A. Koch. 2017. "The Molecular Mechanisms of Thalidomide Teratogenicity and Implications for Modern Medicine." *Current Molecular Medicine* 17 (2). https://doi.org/10.2174/1566524017666170331162315.
Konc, Janez, and Dušanka Janežič. 2010. "ProBiS Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment." *Bioinformatics* 26 (9): 1160–68. https://doi.org/10.1093/bioinformatics/btq100.
Koshland, Daniel E. 1958. "Application of a Theory of Enzyme Specificity to Protein." *Proceedings of the National Academy of Sciences* 44: 98–104.
Krasowski, Agata, Daniel Muthas, Aurijit Sarkar, Stefan Schmitt, and Ruth Brenk. 2011. "DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set." *Journal of Chemical Information and Modeling* 51 (11): 2829–42. https://doi.org/10.1021/ci200266d.
Kuhn, Michael, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. "The SIDER Database of Drugs and Side Effects." *Nucleic Acids Research* 44 (D1): D1075–79. https://doi.org/10.1093/nar/gkv1075.
Kumar, Akhil, Ashish Tiwari, and Ashok Sharma. 2018. "Changing Paradigm from One Target One Ligand Towards Multi-Target Directed Ligand Design for Key Drug Targets of Alzheimer Disease: An Important Role of In Silico Methods in Multi-Target Directed Ligands Design." *Current Neuropharmacology* 16 (6): 726–39. https://doi.org/10.2174/1570159x16666180315141643.
Kurczab, Rafał, Paweł Śliwa, Krzysztof Rataj, Rafał Kafel, and Andrzej J. Bojarski. 2018. "Salt Bridge in Ligand-Protein Complexes - Systematic Theoretical and Statistical Investigations." *Journal of Chemical Information and Modeling* 58 (11): 2224–38. https://doi.org/10.1021/acs.jcim.8b00266.
Kwon, Sunghark, and Hyun Ho Park. 2019. "Structural Consideration of the Working Mechanism of Fold Type I Transaminases From Eubacteria : Overt and Covert Movement." *Computational and Structural Biotechnology Journal* 17: 1031–39. https://doi.org/10.1016/j.csbj.2019.07.007.
Kyte, Jack, and Russell F Doolittle. 1982. "A Simple Method for Displaying the Hydropathic Character of a Protein." *J. Mol. Biol* 157: 105–32.
Lamb, Audrey L., T. Joseph Kappock, and Nicholas R. Silvaggi. 2015. "You Are Lost without a Map: Navigating the Sea of Protein Structures." *Biochim Biophys Acta* 1854 (4): 258–68. https://doi.org/10.1016/j.bbapap.2014.12.021.
Laskowski, Roman A., Jagoda Jabłońska, Lukáš Pravda, Radka Svobodová Vařeková, and Janet M. Thornton. 2018. "PDBsum: Structural Summaries of PDB Entries." *Protein Science* 27 (1): 129–34. https://doi.org/10.1002/pro.3289.
Lavecchia, Antonio. 2014. "Machine-Learning Approaches in Drug Discovery : Methods and Applications." *Drug Discovery Today* 00 (00): 1–14. https://doi.org/10.1016/j.drudis.2014.10.012.
Leach, Andrew R, Brian K Shoichet, and Catherine E Peishoff. 2006. "Docking and Scoring Perspecti V E" 49 (20). https://doi.org/10.1021/jm060999m.
Lee, Aeri, Kyoungyeul Lee, and Dongsup Kim. 2016. "Expert Opinion on Drug Discovery Using Reverse Docking for Target Identification and Its Applications for Drug Discovery Discovery." *Expert Opinion on Drug Discovery* 11 (7): 707–15. https://doi.org/10.1080/17460441.2016.1190706.
Li, Jiao, and Zhiyong Lu. 2013. "Pathway-Based Drug Repositioning Using Causal Inference." *BMC Bioinformatics* 14 (SUPPL16): S3. https://doi.org/10.1186/1471-2105-14-S16-S3.
Lipinski, Christopher A, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. 2001. "Experimental and

Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings." *Advanced Drug Delivery Reviews* 46 (1): 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

Lo, Yu Chen, Stefano E. Rensi, Wen Torng, and Russ B. Altman. 2018. "Machine Learning in Chemoinformatics and Drug Discovery." *Drug Discovery Today* 23 (8): 1538–46. https://doi.org/10.1016/j.drudis.2018.05.010.

Luo, Yunan, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. 2017. "A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information." *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-00680-8.

Malhotra, Ranjan P., Edward Meier, Gail Torkildsen, Paul J. Gomes, and Mark C. Jasek. 2019. "Safety of Cetirizine Ophthalmic Solution 0.24% for the Treatment of Allergic Conjunctivitis in Adult and Pediatric Subjects." *Clinical Ophthalmology* 13: 403–13. https://doi.org/10.2147/OPTH.S186092.

Marc, Daniel. 2012. "NS1 Des Virus Influenza: Une Protéine Très «influente»." *Virologie*. Vol. 16. https://doi.org/10.1684/vir.2012.0444.

MDL Information Systems Inc. 2000. "MACCS Drug Data Report, Release 2000.2." San Leandro, CA, USA, USA: MDL Information Systems, Inc.

Meng, Fan Rong, Zhu Hong You, Xing Chen, Yong Zhou, and Ji Yong An. 2017. "Prediction of Drug–Target Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures." *Molecules* 22 (7). https://doi.org/10.3390/molecules22071119.

Mestres, Jordi, Elisabet Gregori-Puigjané, Sergi Valverde, and Ricard V Solé. 2009. "The Topology of Drug-Target Interaction Networks: Implicit Dependence on Drug Properties and Target Families." *Molecular BioSystems* 5 (9): 1051–57. https://doi.org/10.1039/b905821b.

Meyers, Joshua, Nicola E A Chessum, Salyha Ali, N Yi Mok, Birgit Wilding, A Elisa Pasqua, Martin Rowlands, et al. 2018. "Privileged Structures and Polypharmacology within and between Protein Families." *ACS Medicinal Chemistry Letters* 9 (12): 1199–1204. https://doi.org/10.1021/acsmedchemlett.8b00364.

Milletti, Francesca, and Anna Vulpetti. 2010. "Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe." *Journal of Chemical Information and Modeling* 50 (8): 1418–31. https://doi.org/10.1021/ci1001263.

Miteva, Maria A, Stephanie Violas, Matthieu Montes, David Gomez, Pierre Tuffery, and Bruno O Villoutreix. 2006. "FAF-Drugs: Free ADME/Tox Filtering of Compound Collections." *Nucleic Acids Research* 34 (Web Server issue): W738-44. https://doi.org/10.1093/nar/gkl065.

Morris, Garrett M, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. 2009. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility." *Journal of Computational Chemistry* 30 (16): 2785–91. https://doi.org/10.1002/jcc.21256.AutoDock4.

Nickel, Janette, Bjoern Oliver Gohlke, Jevgeni Erehman, Priyanka Banerjee, Wen Wei Rong, Andrean Goede, Mathias Dunkel, and Robert Preissner. 2014. "SuperPred: Update on Drug Classification and Target Prediction." *Nucleic Acids Research* 42 (W1): 26–31. https://doi.org/10.1093/nar/gku477.

Pabon, Nicolas A, and Carlos J Camacho. 2017. "Probing Protein Flexibility Reveals a Mechanism for Selective Promiscuity." *ELife* 6 (April). https://doi.org/10.7554/eLife.22889.

Pagel, Walter. 1982. *PARACELSUS An Introduction to Philosophical Medicine in the Era of the Renaissance*. Edited by Karger. 2nd, revis ed. Basel; New York.

Pérot, Stéphanie, Leslie Regad, Christelle Reynès, Olivier Spérandio, Maria A. Miteva, Bruno O. Villoutreix, Anne-Claude Claude Camproux, et al. 2013. "Insights into an Original Pocket-Ligand Pair Classification: A Promising Tool for Ligand Profile Prediction." *PLoS ONE* 8 (6): e63730. https://doi.org/10.1371/journal.pone.0063730.

Petitjean, Michel. 2014. "RADI Version 4.0." http://petitjeanmichel.free.fr.

Pu, Limeng, Rajiv Gandhi Govindaraj, Jeffrey Mitchell Lemoine, Hsiao-Chun Wu, and Michal Brylinski. 2019. "DeepDrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network." *PLOS Computational Biology* 15 (2): e1006718. https://doi.org/10.1371/journal.pcbi.1006718.

Raber, Inbar, Cian P. McCarthy, Muthiah Vaduganathan, Deepak L. Bhatt, David A. Wood, John G.F. Cleland, Roger S. Blumenthal, and John W. McEvoy. 2019. "The Rise and Fall of Aspirin in the Primary Prevention of Cardiovascular Disease." *The Lancet* 393 (10186): 2155–67. https://doi.org/10.1016/S0140-6736(19)30541-0.

Revue Prescrire. 2005. "Éviter Les Rétinoïdes En Application Cutanée Pendant La Grossesse (Suite)." Paris.

Rodríguez-Pérez, Raquel, and Jürgen Bajorath. 2018. "Prediction of Compound Pro Fi Ling Matrices , Part II : Relative Performance of Multitask Deep Learning and Random Forest Classi Fi Cation on the Basis of Varying Amounts of Training Data." *ACS Omega* 3: 12033−12040. https://doi.org/10.1021/acsomega.8b01682.

Rogers, David J, and Taffee T Tanimoto. 1960. "A Computer Program for Classifying Plants." *Science* 132 (3434):

1115–18.
Rognan, D. 2012. “Fragment-Based Drug Design Discovery and X-Ray Crystallography.” In *Topics in Current Chemistry*, edited by Thomas G. Davies and Marko Hyvönen, 317:201–22. Springer-Verlag Berlin Heidelberg 2012. https://doi.org/10.1007/978-3-642-27540-1.
Rognan, Didier. 2017. “The Impact of in Silico Screening in the Discovery of Novel and Safer Drug Candidates.” *Pharmacology and Therapeutics* 175: 47–66. https://doi.org/10.1016/j.pharmthera.2017.02.034.
Rossi, Alfredo, Carmen Cantisani, Luca Melis, Alessandra Iorio, Elisabetta Scali, and Stefano Calvieri. 2012. “Minoxidil Use in Dermatology, Side Effects and Recent Patents.” *Recent Patents on Inflammation & Allergy Drug Discovery* 6 (2): 130–36. https://doi.org/10.2174/187221312800166859.
Schaeffer, Laurent. 2008. “The Role of Functional Groups in Drug–Receptor Interactions.” *The Practice of Medicinal Chemistry*, January, 359–78. https://doi.org/10.1016/B978-0-12-417205-0.00014-6.
Shterev, Ivo D, David B Dunson, Cliburn Chan, and Gregory D Sempowski. 2018. “Bayesian Multi-Plate High-Throughput Screening of Compounds.” *Scientific Reports*, 1–10. https://doi.org/10.1038/s41598-018-27531-w.
Silva, Franck Da, Jeremy Desaphy, and Didier Rognan. 2018. “IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions.” *ChemMedChem* 13 (6): 507–10. https://doi.org/10.1002/cmdc.201700505.
Skalic, Miha, Gerard Martínez-Rosell, José Jiménez, and Gianni De Fabritiis. 2018. “PlayMolecule BindScope: Large Scale CNN-Based Virtual Screening on the Web.” *Bioinformatics* 35 (7): 1–2. https://doi.org/10.1093/bioinformatics/bty758.
Spraggon, Glen, Christopher Phillips, Ursula K Nowak, Christopher P Ponting, Derek Saunders, Christopher M Dobson, David I Stuart, and E.Yvonne Jones. 1995. “The Crystal Structure of the Catalytic Domain of Human Urokinase-Type Plasminogen Activator.” *Structure* 3 (7): 681–91. https://doi.org/10.1016/S0969-2126(01)00203-9.
Sterling, Teague, and John J. Irwin. 2015. “ZINC 15 - Ligand Discovery for Everyone.” *Journal of Chemical Information and Modeling* 55 (11): 2324–37. https://doi.org/10.1021/acs.jcim.5b00559.
Storer, Barry E. 1989. “Design and Analysis of Phase I Clinical Trials.” *Biometrics* 45 (3): 925–37. http://www.jstor.org/stable/2531693.
The UniProt Consortium. 2019. “UniProt: A Worldwide Hub of Protein Knowledge.” *Nucleic Acids Research* 47 (D1): D506–15. https://doi.org/10.1093/nar/gky1049.
Tian, Wei, Chang Chen, Xue Lei, Jieling Zhao, and Jie Liang. 2018. “CASTp 3.0: Computed Atlas of Surface Topography of Proteins.” *Nucleic Acids Research* 46 (W1): W363–67. https://doi.org/10.1093/nar/gky473.
Tonddast-Navaei, Sam, Bharath Srinivasan, and Jeffrey Skolnick. 2017. “On the Importance of COmposite Protein Multiple LIGand (COLIG) Interactions in Protein Pockets.” *J Comput Chem.* 38 (15): 1252–59. https://doi.org/10.1002/cncr.27633.Percutaneous.
Triki, Dhoha, Telli Billot, Benoit Visseaux, Diane Descamps, Delphine Flatters, Anne Claude Camproux, and Leslie Regad. 2018. “Exploration of the Effect of Sequence Variations Located inside the Binding Pocket of HIV-1 and HIV-2 Proteases.” *Scientific Reports* 8 (1): 30–40. https://doi.org/10.1038/s41598-018-24124-5.
Triki, Dhoha, Sandrine Fartek, Benoit Visseaux, Diane Descamps, Anne Claude Camproux, and Leslie Regad. 2018. “Characterizing the Structural Variability of HIV-2 Protease upon the Binding of Diverse Ligands Using a Structural Alphabet Approach.” *Journal of Biomolecular Structure and Dynamics* 0 (0): 1–13. https://doi.org/10.1080/07391102.2018.1562985.
Trott, Oleg, and Arthur J. Olson. 2009. “Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.” *Journal of Computational Chemistry* 31: 455–61. https://doi.org/10.1002/jcc.21334.
Veber, Daniel F., Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. 2002. “Molecular Properties That Influence the Oral Bioavailability of Drug Candidates.” *Journal of Medicinal Chemistry* 45: 2615–23. https://doi.org/10.1021/JM020017N.
Volkamer, Andrea, Axel Griewel, Thomas Grombacher, and Matthias Rarey. 2010. “Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets.” *Journal of Chemical Information and Modeling* 50 (11): 2041–52. https://doi.org/10.1021/ci100241y.
Wang, Pu, Ruiquan Ge, Xuan Xiao, Yunpeng Cai, and Guoqing Wang. 2016. “Rectified-Linear-Unit-Based Deep Learning for Biomedical Multi-Label Data.” *Interdisciplinary Sciences: Computational Life Sciences* 9 (3): 419–22. https://doi.org/10.1007/s12539-016-0196-1.
Wang, Zhe, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. 2016. “Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power.” *Physical Chemistry Chemical Physics* 18 (18): 12964–75. https://doi.org/10.1039/c6cp01555g.
Ward, Joe H Jr. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American*

Statistical Association 58 (301): 236–44.
Wass, Mark N., Lawrence A. Kelley, and Michael J E Sternberg. 2010. "3DLigandSite: Predicting Ligand-Binding Sites Using Similar Structures." *Nucleic Acids Research* 38 (SUPPL. 2): 469–73. https://doi.org/10.1093/nar/gkq406.
Weill, Alain, Michel Paı, Philippe Tuppin, Jean-paul Fagot, Anke Neumann, and Dominique Simon. 2010. "Benfluorex and Valvular Heart Disease : A Cohort Study of a Million People with Diabetes Mellitus." *Pharmacoepidemiology and Drug Safety* 19: 1256–62. https://doi.org/10.1002/pds.2044.
Weill, Nathanaël, and Didier Rognan. 2010. "Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites." *Journal of Chemical Information and Modeling* 50 (1): 123–35. https://doi.org/10.1021/ci900349y.
Wen, Ming, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, and Yonghuan Yun. 2017. "Deep-Learning-Based Drug − Target Interaction Prediction." https://doi.org/10.1021/acs.jproteome.6b00618.
Wilkes, Scott. 2008. "The Use of Bupropion SR in Cigarette Smoking Cessation." *International Journal of COPD* 3 (1): 45–53.
Wu, CY, and JJ. Wittick. 1977. "Separation of Five Major Alkaloids in Gum Opium and Quantitation of Morphine, Codeine, and Thebaine by Isocratic Reverse Phase High Performance Liquid Chromatography." *Anal Chem* 49 (3): 359–63.
Xia, Xuhua. 2017. "Bioinformatics and Drug Discovery," 1709–26. https://doi.org/10.2174/1568026617666161116143440.
Xu, Hua, Melinda C. Aldrich, Qingxia Chen, Hongfang Liu, Neeraja B. Peterson, Qi Dai, Mia Levy, et al. 2015. "Validating Drug Repurposing Signals Using Electronic Health Records: A Case Study of Metformin Associated with Reduced Cancer Mortality." *Journal of the American Medical Informatics Association* 22 (1): 179–91. https://doi.org/10.1136/amiajnl-2014-002649.
Yamanishi, Yoshihiro, Masaaki Kotera, Yuki Moriya, Ryusuke Sawada, Minoru Kanehisa, and Susumu Goto. 2014. "DINIES: Drug-Target Interaction Network Inference Engine Based on Supervised Analysis." *Nucleic Acids Research* 42 (W1): 39–45. https://doi.org/10.1093/nar/gku337.
Yang, Lun, and Pankaj Agarwal. 2011. "Systematic Drug Repositioning Based on Clinical Side-Effects." *PLoS ONE* 6 (12). https://doi.org/10.1371/journal.pone.0028025.
Yao, Zhi Jiang, Jie Dong, Yu Jing Che, Min Feng Zhu, Ming Wen, Ning Ning Wang, Shan Wang, Ai Ping Lu, and Dong Sheng Cao. 2016. "TargetNet: A Web Service for Predicting Potential Drug–Target Interaction Profiling via Multi-Target SAR Models." *Journal of Computer-Aided Molecular Design* 30 (5): 413–24. https://doi.org/10.1007/s10822-016-9915-2.
Ye, Hao, Qi Liu, and Jia Wei. 2014. "Construction of Drug Network Based on Side Effects and Its Application for Drug Repositioning." *PLoS ONE* 9 (2). https://doi.org/10.1371/journal.pone.0087864.
Yella, Jaswanth K., Suryanarayana Yaddanapudi, Yunguan Wang, and Anil G. Jegga. 2018. "Changing Trends in Computational Drug Repositioning." *Pharmaceuticals* 11 (2): 1–21. https://doi.org/10.3390/ph11020057.
Zhang, Lu, Jianjun Tan, Dan Han, and Hao Zhu. 2017. "From Machine Learning to Deep Learning : Progress in Machine Intelligence for Rational Drug Discovery." *Drug Discovery Today* 22 (11): 1680–85. https://doi.org/10.1016/j.drudis.2017.08.010.
Zhang, Yang, and Jeffrey Skolnick. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score." *Nucleic Acids Research* 33 (7): 2302–9. https://doi.org/10.1093/nar/gki524.

ANNEXES

Article n°1 :

Cerisier N, Regad L, Triki D, Camproux AC, Petitjean M. Cavity Versus Ligand Shape Descriptors: Application to Urokinase Binding Pockets, *J Comput Biol.* 2017 Jun 1. doi: 10.1089/cmb.2017.0061

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 0, 2017
Mary Ann Liebert, Inc.
Pp. 1–4
DOI: 10.1089/cmb.2017.0061

Communication

Cavity Versus Ligand Shape Descriptors: Application to Urokinase Binding Pockets

NATACHA CERISIER,[1] LESLIE REGAD,[1] DHOHA TRIKI,[1] ANNE-CLAUDE CAMPROUX,[1] and MICHEL PETITJEAN[1,2]

ABSTRACT

We analyzed 78 binding pockets of the human urokinase plasminogen activator (uPA) catalytic domain extracted from a data set of crystallized uPA–ligand complexes. These binding pockets were computed with an original geometric method that does *NOT* involve any arbitrary parameter, such as cutoff distances, angles, and so on. We measured the deviation from convexity of each pocket shape with the pocket convexity index (PCI). We defined a new pocket descriptor called distributional sphericity coefficient (DISC), which indicates to which extent the protein atoms of a given pocket lie on the surface of a sphere. The DISC values were computed with the freeware *PCI*. The pocket descriptors and their high correspondences with ligand descriptors are crucial for polypharmacology prediction. We found that the protein heavy atoms lining the urokinases binding pockets are either located on the surface of their convex hull or lie close to this surface. We also found that the radii of the urokinases binding pockets and the radii of their ligands are highly correlated ($r = 0.9$).

Keywords: algorithms, protein binding pockets, statistics, urokinases.

1. INTRODUCTION

SEVERAL THOUSANDS OF MOLECULAR DESCRIPTORS are known (Todeschini and Consonni, 2008). Although they are suitable for protein ligands, most of them are meaningless for protein pockets. This is particularly true for geometric descriptors, because the ligand shape is commonly associated with some envelope separating the ligand to its exterior, whereas the shape of a protein pocket is rather associated with the boundary of a cavity internal to the protein. Thus, we used our own pocket descriptors (Section 3). Then, the definition of a cavity inside a protein is highly polemical: see the many algorithms cited by Pérot et al. (2010)

[1]MTi, INSERM UMR-S 973, Université Paris Diderot, Paris, France.
[2]Epôle de Génoinformatique, Institut Jacques Monod, CNRS, UMR7592, Université Paris Diderot, Paris, France.

and by Benkaidali et al. (2014). The main problem encountered in pocket calculation algorithms, and more generally in modeling algorithms, is the existence of arbitrary parameters having a crucial effect on the results. It is why we built our own pocket calculation algorithm, which is parameter free (Section 2.2). We applied this calculation to a set of human urokinase plasminogen activator (uPA) catalytic domains. It is an attractive therapeutic target in cancer because it plays an essential role in the process of tumor cell migration and metastasis (Andreasen et al., 2000). Furthermore, the uPA receptor system is known as a strategic therapeutic target (Degryse, 2013).

2. METHODS

2.1. Data preparation

From the Protein Data Bank (PDB; Berman et al., 2000), we extracted a set of 97 crystallized uPA catalytic domain–ligand complexes. To remove nonspecific and nonbiological ligands, we removed the crystallization additives and salts. We also removed all hydrogen atoms to get a homogeneous set. In the case of polymers with multiple ligands (e.g., 2VNT PDB code), we duplicated the file in the ones where there are ligands, to launch an automated treatment for pocket calculations. The final working set contained 71 human urokinase catalytic domains, from where 78 pockets are extracted, each pocket containing one ligand.

2.2. Pocket calculation algorithm

We define a protein pocket as protein atoms extracted using the two following steps: (1) for each ligand atom we retain its closest neighboring atom in the protein and (2) in the case of multiple copies of a protein atom, we retain only one to get a nonredundant set. The drawback of this algorithm is that it cannot apply if there is no ligand. However, the strong advantage of our algorithm is its simplicity, particularly the fact that it does not require any parameter.

3. RESULTS AND DISCUSSION

We calculated the pocket descriptors of Table 1 with the *PCI* freeware. The 78 PCI values ranging from 0 to 0.04 indicate highly convex pockets. The pocket sphericity index values range from 0.14 to 0.50: the largest inscribed sphere (radius R_i) is smaller than the radius of the smallest circumscribed sphere (radius R_h): the pockets are a bit flat. It is stressed that a pocket could be nearly flat while its bounding atoms indeed lie on the surface on a sphere. The distributional sphericity measured by the distributional sphericity coefficient (DISC) parameter shows to which extent the pocket protein atoms lie on the surface of a sphere (Appendix). The DISC values range from 0.01 to 0.27. Thus, the pocket shapes are moderately fitted by spheres. To study the correspondence between pocket and ligand's shapes, we looked at the correlation between the pocket radii R (Appendix) and the radii R_{hL} of the smallest sphere enclosing the ligand. We found a high correlation coefficient $r(R, R_{hL}) = 0.89$. The correlation coefficient $r(R_h, R_{hL}) = 0.91$ is also high. These results show that our method is suitable to estimate pockets guided by the ligand. This

Table 1. The Main Pocket Descriptors Computed with *PCI*

N	Number of pocket atoms
R^h	Radius of the convex hull of the N atoms (Petitjean, 1992)
R^i	Radius of the largest sphere inscribed in the convex hull
PSI	Pocket sphericity index[a]: R_i/R_h; takes values in [0;1]
PCI	Pocket convexity index[a,b]; takes values in [0;1]
R	Pocket radius (see Theorem 4.1 in Appendix)
DISC	Distributional sphericity coefficient (Appendix)

[a]These indices were first mentioned by Borrel et al. (2015).
[b]Ratio of the squared quadratic mean distance of the N atoms to their hull, to R_i^2.

estimation is a crucial step to predict interaction partners and off targets, and to address polypharmacology (Abi Hussein et al., 2017). Our combined use of descriptors with a free parameter pocket estimation permitted us to evaluate the adaptation of the pocket to the size of the ligand.

4. APPENDIX: CALCULATION OF THE BEST FITTING SPHERE

Consider n points $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^d$, $d \geq 1$. The best fitting sphere is defined so that its center c minimizes the variance V of the population of the squared distances of the n points to c. The radius R of the best fitting sphere is the square root of the mean of these n squared distances. When the minimized variance is null, all the n points lie on the boundary of a sphere of radius R centered on c.

The calculation of R and c are done according to Theorem 4.1. For convenience, this theorem is presented for random vectors. The case of n points in $\mathbb{R}^d$ is retrieved through a finite discrete random vector. In what follows, the quote denotes the transposition operator and the symbol E denotes the expectation operator.

Let X be a random vector taking values in $\mathbb{R}^d$. The random variable expressing the squared length of $X-c$ is $Z=(X-c)'(X-c)$. The variance of Z is $V=E(Z-EZ)^2$, assumed to exist. The variance matrix of X is $\mathbf{K}=EYY'$, with $Y=X-EX$. $\mathbf{K}$ is assumed to be of full rank. V_0 is the variance of $Y'Y$. We set $\gamma=EYY'Y$ and $\tau=c-EX$.

Theorem 4.1. *The center of the best fitting sphere is* $c=EX+\mathbf{K}^{-1}\gamma/2$ *and its squared radius* $R^2=EY'Y+\tau'\tau$. *The minimized variance is* $V=V_0-\gamma'\mathbf{K}^{-1}\gamma$.

Proof. $Z=(Y-\tau)'(Y-\tau)$ and $Z-EZ=Y'Y-EY'Y+\tau'\tau$. The variance of $Z-EZ$ is $V=V_0-4\tau'\gamma+4\tau'\mathbf{K}\tau$. The gradient of this variance with respect to τ is $\nabla V=-4\gamma+8\mathbf{K}\tau$. Thus we deduce that the optimal value of τ is $\mathbf{K}^{-1}\gamma/2$. The minimal variance and the optimal radius are deduced from the latter. ■

The support of X lies on the boundary of a sphere if and only if $V=0$. When $V>0$ we need to evaluate how the distribution of X deviates from this ideal case. Having $c=EX+\mathbf{K}^{-1}\gamma/2$, we could look at the normalized variance $\Delta=V/R^4$. Unfortunately, Δ can be arbitrarily large. Thus we define the quantity DISC, which takes values in [0;1]:

Definition 4.1. $\mathrm{DISC}=\Delta/(1+\Delta)$ *is the distributional sphericity coefficient.* $\mathrm{DISC}=V/(V+R^4)$.

DISC is null if and only if the support of X lies on the boundary of a sphere. The larger DISC is, the more the distribution of X deviates from this ideal situation. DISC is insensitive to isometries and scaling.

When $d=1$, the value $\mathrm{DISC}=0$ is reached if and only if the random variable X follows a Bernouilli distribution. Thus, when $d=1$, DISC may also be viewed as a bimodality coefficient.

Still when $d=1$, we retrieve in corollary 4 an interesting inequality, which goes back to Pearson (1929). Let $\mathcal{S}$ be the skewness of X, that is, its reduced centered third order moment, and $\mathcal{K}$ its kurtosis, that is, its reduced centered fourth order moment, assumed to be existing.

Corollary 4.1. $\mathcal{K} \geq \mathcal{S}^2+1$.

Proof. Set $d=1$ in Theorem 4.1. $V=\sigma^4(\mathcal{K}-1-\mathcal{S}^2)$, σ being the standard deviation of X. Write that $V \geq 0$. ■

This inequality was also mentioned by Petitjean (2013), as a consequence of a result about geometric docking (Petitjean, 2004). Theorem 4.1 can also be viewed as a consequence of this result of Petitjean (2004).

AVAILABILITY OF *PCI*

Free binaries and documentation of *PCI* are available through a software repository located at http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abi Hussein, H., Geneix, C., Petitjean, M., et al. 2017. Global vision of druggability issues, applications and perspectives. *Drug Discov. Today*. 22, 404–415.

Andreasen, P.A., Egelund, R., and Petersen, H.H. 2000. The plasminogen activation system in tumor growth, invasion, and metastasis. *Cell. Mol. Life Sci.* 57, 25–40.

Benkaidali, L., André, F., Maouche, et al. 2014. Computing cavities, channels, pores and pockets in proteins from non spherical ligands models. *Bioinformatics*. 30, 792–800.

Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.

Borrel, A., Regad, L., Xhaard, H., et al. 2015. PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties. *J. Chem. Inf. Model.* 55, 882–895.

Degryse, B. 2013. Editorial: The urokinase receptor system as strategic therapeutic target: Challenges for the 21st century. *Curr. Pharm. Des.* 17, 1872–1873.

Pearson, K. 1929. Editorial note to Inequalities for moments of frequency functions and for various statistical constants, by J.Shohat. *Biometrika*. 21, 361–375.

Pérot, S., Sperandio, O., Miteva, M.A., et al. 2010. Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discov. Today*. 15, 656–667.

Petitjean, M. 1992. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* 32, 331–337.

Petitjean, M. 2004. From shape similarity to shape complementarity: Toward a docking theory. *J. Math. Chem.* 35, 147–158.

Petitjean, M. 2013. The chiral index: Applications to multivariate distributions and to 3D molecular graphs, 11–16. *In* Zadnik Stirn, L., Žerovnik, J., Povh, J., Drobne, S., and Lisec, A., eds. *Proceedings of SOR'13, the 12th International Symposium on Operational Research in Slovenia*. Slovenian Society INFORMATIKA (SDI), Section for Operations Research (SOR), Ljubljana, Slovenia.

Todeschini, R., and Consonni, V. 2008. *Handbook of Molecular Descriptors*. Wiley, New York.

Address correspondence to:
Dr. Michel Petitjean
MTi, INSERM UMR-S 973
Université Paris Diderot
35 rue Hélène Brion
75205 Paris Cedex 13
France

E-mail: michel.petitjean@univ-paris-diderot.fr;
petitjean.chiral@gmail.com