UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire d'Innovation Thérapeutique

THÈSE présentée par :

Franck Da Silva

soutenue le : **23 Septembre 2016**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**
Discipline/ Spécialité : Chimie/Chémoinformatique

Cartographie des interfaces protéine-protéine et recherche de cavités droguables

THÈSE dirigée par :
Dr. ROGNAN Didier Directeur de recherche, CNRS

RAPPORTEURS :
Dr. MORELLI Xavier Directeur de recherche, CNRS
Dr.SPERANDIO Olivier Chargé de recherche, INSERM

AUTRES MEMBRES DU JURY :
Dr. BAADEN Marc Directeur de Recherche, CNRS

Remerciements

Tous d'abord je souhaiterais remercier Didier Rognan pour m'avoir proposé cette thèse et m'avoir accepté dans l'équipe Chémogénomique Structurale au sein de la Faculté de Pharmacie d'Illkirch. Je tiens ici à remercier l'équipe complète du laboratoire dont Jérémy, Guillaume et Esther car ce travail ne serait pas ce qu'il est sans leur contribution. Ce fut un réel plaisir de travailler avec eux tant pour leur qualités respectives que pour leur envie de partager leurs expériences, leurs méthodes et leurs savoir-faire.

Mes remerciements s'adressent à toutes les personnes qui un jour ont fait partie de cet extraordinaire voyage. Je ne sais pas par où commencer, peut-être par Pierre Poulain et Denis Métivier qui m'ont initié à la bio-informatique et m'ont transmis leur passion avec brio. Un grand merci aussi à Marc Baaden et Jacques Chomiler qui m'ont fait confiance et apporté mes premières expérience en laboratoire. Ce fut deux vraies expériences qui m'ont appris beaucoup et sans lesquelles je ne serais pas là aujourd'hui.

Je remercie encore Didier pour m'avoir accordé le privilège d'effectuer ma thèse sous sa tutelle. Je le remercie pour sa confiance et pour tout ce qu'il m'a appris ainsi que pour toujours avoir été là à mon écoute, merci.

Le milieu scientifique est très important mais il ne faut pas oublier la famille, je remercie mes grands-parents qui, même si je suis le petit dernier, ont toujours été là pour moi, je vous remercie pour vous être toujours souciés de moi. Je voudrais aussi dire un grand merci à ma famille. Un grand bisou à mes parents qui, même si ils ne comprennent pas l'ensemble du contenu de ces travaux, s'y intéressent. Une pensée aussi à ma sœur et sa petite famille qui sont une grande source d'énergie. Et enfin il n'y a pas de mots pour te remercier Aimie, sans qui rien de tout cela n'aurait été possible. Tu es celle qui n'abandonne jamais, qui a toujours su et du me réconforter et me remotiver, tu es remarquable et unique.

Enfin, je tiens à remercier tous les membres de mon jury, pour avoir accepté de juger ce travail.

J'attache beaucoup d'importance à ces trois années et toutes les personnes qui y ont contribué car cela m'a fais évoluer et me rapprocher des personnes que j'aime.

Pour finir, un merci à vous lecteur qui, quelque soit vos motivations, lisez ces quelques mots.

Merci

Table des matières

Résumé

Le domaine du vivant est un ensemble complexe. Le vivant est entre autres une suite d'interactions entre des protéines et des petites molécules appelées ligands. Les protéines sont des ensembles d'atomes organisés et stabilisés dans un espace tridimensionnel. Chaque protéine possède un rôle précis, réalisant des interactions définies par son agencement propre. Il existe des méthodes pour définir les éléments qui interagissent entre eux, et créer ce que l'on appelle l'Interactome, mais la nature des interactions reste quant à elle difficile à décrire.
La Bioinformatique structurale et la Cheminformatique sont deux disciplines visant à modéliser et étudier in-silico les protéines et les ligands. Cette approche théorique se fait selon deux principaux axes qui sont, la compréhension et la description des mécanismes impliqués ainsi que l'identification de nouvelles molécules pour le traitement de maladies. Ces études sont fréquemment basées sur la Protein Data Bank (PDB), une base de données regroupant près de 120 000 protéines représentées sous la forme de leur structure tridimensionnelle. Le nombre d'interfaces protéine-protéine (PPI) et protéine-ligand explose depuis quelques années grâce à aux développements spectaculaires en biologie structurale.

Moduler les interactions protéine-protéine par des ligands de faible poids moléculaire est une approche nouvelle et prometteuse pour la découverte de médicaments, ouvrant de nouvelles voies thérapeutiques et étendant les champs d'application des cibles macromoléculaires actuellement connues. Une des difficultés principales de cette stratégie réside dans le fait que les interfaces protéine-protéine sont la plupart du temps plates et peu propices à l'accrochage de petites molécules inhibitrices. Cependant, des travaux récents de recherche d'inhibiteurs d'interactions montrent des résultats encourageants. La détection et la caractérisation des PPI est donc une étape clé de la découverte de nouveaux candidats médicaments.

Le premier chapitre de cette thèse s'intéresse aux interfaces protéine-protéine, il définit l'ensemble des méthodes disponibles actuellement pour caractériser une PPI. La caractérisation d'une protéine a évolué au fil des avancées technologiques majeures aussi bien d'un point de vue expérimental que théorique. La discrimination des interfaces m'intéresse plus particulièrement, l'ensemble des structures sont résolues sous des conditions très strictes et restrictives qui peuvent entrainer la création d'artéfacts structuraux ne représentant pas l'état physiologique in-vivo des protéines. Différencier une structure biologiquement

pertinente a toujours été un élément important pour l'étude des PPI. Ce chapitre montre également l'importance des jeux de données d'études en pointant les défauts des jeux dits "historiques" ainsi qu'en expliquant la création d'un nouveau jeu de données équilibré permettant de réaliser un apprentissage efficient.

Le second chapitre s'intéresse aux outils développés et/ou améliorés au sein du laboratoire. La majorité des travaux de cette thèse sont intégrés à une suite logicielle (IChem) d'analyse d'interactions atomiques à visé pharmacologique. La première partie décrit le fonctionnement global du logiciel. Une grille composée d'une multitude de voxels de 1,5Å de dimension est placé autour des molécules, cela permet de calculer une valeur d'enfouissement pour chaque point de l'espace en regardant son environnement. La grille définit les voxels de protéines, de surfaces ainsi que les voxels trop éloignés. Les sites d'interaction et plus précisément les cavités formées à la surface des protéines sont déterminées à l'aide des valeurs d'enfouissement. A chaque voxel est aussi associée une propriété physico-chimique représentant la propriété de l'atome attendu si un ligand se trouvait dans cette espace. Cette propriété est définie par des règles géométriques strictes, de distances et d'angles. L'expérience a montré qu'il existe généralement plus d'une cavité par protéine, la suite du chapitre décrit la manière de déterminer plusieurs cavités à la surface d'une protéine ainsi que la manière de les trier. Le tri s'effectue principalement sur deux critères qui sont la taille de la cavité et la droguabilité, celle-ci étant une valeur représentant la faculté qu'a une cavité à accueillir une molécule droguable de petite taille. La droguabilité est calculé par un modèle d'apprentissage (svm) sur 76 cavités contenant ou non un ligand. Le modèle binaire obtient une précision de 88%. Une autre manière de représenter un site de liaison d'un ligand est le pharmacophore, il s'agit d'un ensemble de pseudos-atomes typés représentant les atomes interagissant avec la protéine. Le pharmacophore est en général créé à partir d'un ou plusieurs ligands. Je montre ici qu'il est possible de le créer à partir d'une cavité en l'absence de tout ligand, en sélectionnant et regroupant correctement les bonnes propriétés. Dans ce chapitre, je montre qu'il est possible d'obtenir des phamacophores de moins de 40 propriétés conservant 70% des liaisons protéine-ligand et permettant d'aligner un ligand à moins de 3Å de sa position cristallographique. Cette technique a déjà été utilisée mais n'est pas encore finalisée notamment dans la priorisation des poses à sélectionner.

Le troisième chapitre décrit le fonctionnement d'IChemPIC, un logiciel permettant de déterminer, décrire puis classer les interfaces protéine-protéine. La procédure d'analyse complète y est décrite en commençant par le jeu d'études créé spécifiquement pour entrainer le logiciel. Les PPI sont détectées par des règles géométriques puis par des règles d'interactions. Cette partie montre aussi comment transformer un objet tridimensionnel en un vecteur, un ensemble de 45 descripteurs étudiables par un modèle d'apprentissage basé sur des forêts aléatoires ("Random forest"). Le modèle de prédiction finale est un consensus de 10 forêts aléatoires réalisées avec les mêmes paramètres. La classification des interfaces protéine-protéine n'est pas nouvelle mais je montre qu'IChemPIC, avec une prédiction à 75%, est aussi précis voir meilleur que les outils actuellement disponibles. Surtout, IchemPIC est la seule méthode à prédire aussi bien les interfaces biologiquement pertinentes que les artéfacts de cristallisation. C'est aussi la méthode présentant le domaine d'applicabilité le plus large, pouvant traiter les interfaces de 200 à 2500Å². Cette dernière particularité est le but de sa création, pouvoir étudier les petites interfaces biologiquement pertinentes qui ont longtemps été oubliées.

Le dernier chapitre présente l'application des travaux précédents sur l'ensemble de la PDB. Le processus d'analyse est décrit, en commençant par la récupération des structures de protéines puis l'application de filtres sur la précision et les méthodes de résolution des structures. L'utilisation d'un nouvel algorithme pour détecter les interfaces ainsi que les cavités et la création de pharmacophores y est décrite. Près de 400 000 interfaces potentielles ont été vérifiées, 60 000 dimères interagissant suffisamment entre eux pour définir une PPI dont 40% ont été prédits étant biologiquement pertinents. Aux niveaux des cavités, nous trouvons en moyenne 7 cavités à proximité des interfaces de dimère dont en moyenne 2 sont prédites droguables. Il nous reste encore à achever une description précise de la totalité des cavités et pharmacophores associés aux interfaces dans le but de déterminer ceux à étudier et valider expérimentalement. Ce chapitre met aussi en avant les limites de la méthode sur les complexes oligomériques complexes.

Pour conclure, ces travaux consistent en une étude complète des interactions protéine-protéine de structure cristallographique connue. Les logiciels spécialement conçus pour cette étude sont rassemblés dans un seul et même outil. Ils ont été appliqués à l'ensemble de la Protein Data Bank. Nous sommes désormais capables d'analyser automatiquement chaque

entrée de la PDB, d'en extraire les chaines en interaction, et de la caractériser précisément à l'échelle atomique. Chaque interface voit son environnement décrit sous forme de cavité droguable et de ligand attenants, ainsi que sous forme de pharmacophore directement généré à partir de la cavité. Ces nouvelles méthodologies permettent une meilleure compréhension et une meilleure sélection des cibles autour des interfaces protéine-protéine. A court terme, nous ambitionnons de sélectionner un panel divers de cavités allostériques à l'interface, de les cribler in silico afin d'identifier de nouveaux modulateurs allostériques de PPI.

Chapitre 1

Les interactions protéine-protéine

1.1. Introduction

La recherche pharmaceutique a pour principal support d'études les interactions protéine-ligand. Le développement de candidats-médicament est un procédé complexe dans lequel on cherche fréquemment à reproduire la nature en mimant l'interaction entre une protéine et son ligand endogène[1]. Depuis quelques années, les interactions protéine-protéine sont devenues une source d'inspiration majeure pour la recherche académique et pharmaceutique, dans le but de pouvoir identifier des molécules de faibles poids moléculaire capables de les moduler sélectivement.

Les interactions protéine-protéine composent ce que l'on appelle l'interactome[2], un réseau immense constitué de toutes les interactions protéine-protéine connues (**Figure 1.1**).

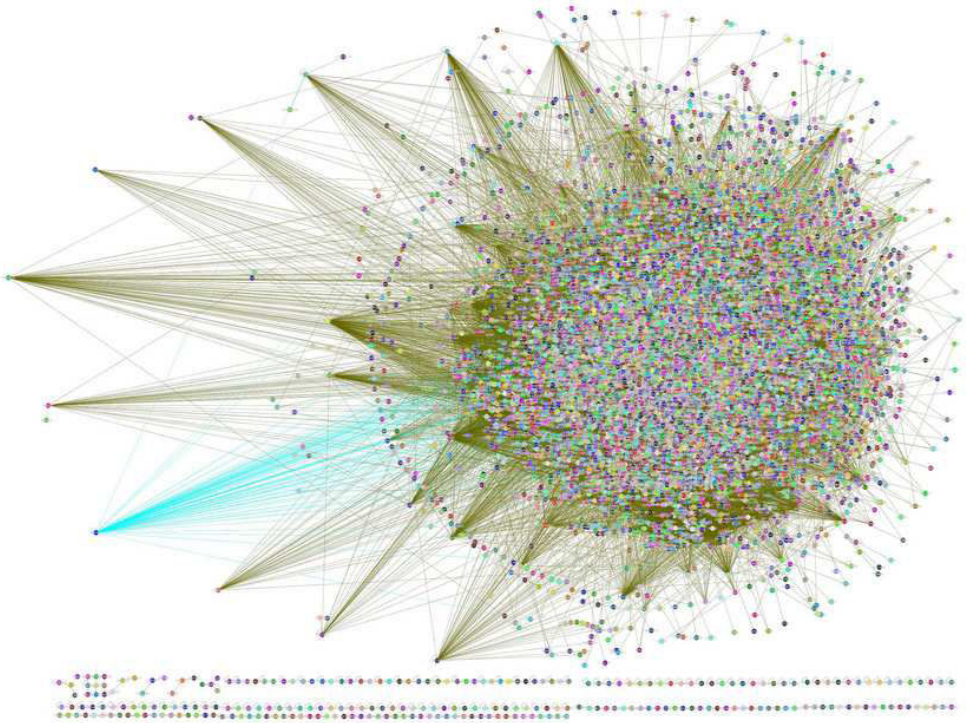


***Figure 1.1** : Représentation de l'interactome humain sous forme d'un graphe, chaque noeud représente une protéine et chaque arrête une interaction. (https://www.flickr.com/photos/andytrop/5232042116)*

L'interactome de la levure est estimé à environ 20 000 interactions, celui de l'espèce humaine étant estimé à environ 650 000 interactions[3] ayant chacune une fonction. Ces prédictions sont fondées principalement sur la taille et le nombre de gènes dans le génome de l'espèce étudiée. Les interactions les plus connues ont des missions de structure (ex: oligomérisation de l'actine) ou des fonctions de transmission de message (ex: récepteurs couplés aux protéines G). Le principe de pouvoir influencer ce réseau par une molécule de faible poids moléculaire (ici nommée petite molécule) ouvre un grand nombre de possibilités pour la recherche pharmaceutique. Les interactions protéine-protéine jouent un rôle majeur dans le vivant[2], elles sont au cœur du fonctionnement de la cellule. Pouvoir moduler ces interactions ouvre l'accès à l'étude et à de possibles traitements[4] de maladies pour lesquelles les méthodes traditionnelles appliquées à des cibles uniques se sont révélées en échec.

La recherche d'une petite molécule pouvant moduler l'effet d'une protéine unique se fait usuellement par l'étude d'une cavité contenant déjà un ligand. Le problème est différent pour une interface protéine-protéine qui, dans la plupart des cas, ne possède ni ligand ni cavité préalablement identifiés. Une autre particularité des interfaces protéine-protéines vient du fait qu'elles sont souvent de grande taille et avec peu de relief, ce qui complique la recherche de molécules inhibitrices. Tester expérimentalement l'ensemble de ces informations demanderait énormément de temps et des moyens infinis. L'accès à l'information biologique est cher et même impossible quand cela concerne l'ensemble des interfaces protéine-protéine humaines. Un moyen d'observer ces interfaces plus facilement est la résolution de structures cristallines. Depuis près de 50 ans, une quantité d'information phénoménale de structure de protéines est mise à la disposition de tous. Ces informations sont rassemblées dans une grande base de données internationale nommée Protein Data Bank (PDB)[5]. Cette base recense actuellement près de 120 000 entrées. Une des méthodes les plus rationnelles pour exploiter cette information est de concevoir des petites molécules altérant la stabilité de ces interfaces. Pour cela, il est nécessaire de déterminer les zones d'interactions protéine-protéine, ainsi que de trouver des sites de liaison potentiels pour de petites molécules. Ces sites de liaison sont des cavités localisées à la surface des protéines constituantes, dans des zones favorables à leur interaction. Ces sites sont flexibles et possèdent une répartition des propriétés physico-chimique unique dues à l'alignement des acides aminés les composant. La complémentarité du site de liaison et de son ligand forme un complexe permettant la réalisation d'interactions non covalentes la stabilisant fortement.

La définition d'un site de liaison et de la complémentarité avec son ligand peut être extrapolée à l'étude des interactions protéine-protéine ou le ligand est une macromolécule. Les zones d'interaction protéine-protéine sont plus grandes qu'un site de liaison de ligand classique mais la complémentarité de formes et de propriétés physico-chimiques reste un principe universel. Toutes les connaissances et méthodes mises au point pour la compréhension des interactions protéine-ligand peuvent donc être adaptées à l'étude des interactions protéine-protéine. L'analyse de la forme et des interactions des interfaces protéine-protéine ouvrent notamment la perspective d'une possible sélection de petites molécules potentiellement actives pour moduler ces interfaces particulières.

Le développement d'outils informatiques réellement adaptés à l'étude des interfaces protéine-protéine et l'identification de ces modulateurs est donc un enjeu majeur de la recherche pharmaceutique moderne.

Ce chapitre décrit de manière globale les particularités d'un site d'interaction entre deux protéines. Dans un premier temps, nous décrirons l'architecture de l'interface ainsi que sa composition. Ces informations nous conduirons au potentiel de developabilité d'un site d'interaction, grâce au type d'interactions présentes. Par la suite nous nous intéresserons aux données rassemblées au cours des années sur le sujet, celles ayant servies de données d'apprentissage. Enfin, nous retracerons l'histoire de la caractérisation des interfaces protéine-protéine au travers de l'analyse des différentes méthodes et outils développés depuis les années 2000.

1.2. Protéines

L'histoire de la science des protéines est un parfait exemple de montagnes russes. Le terme protéine est apparu en 1838[6] mais il a fallu attendre des années pour qu'une représentation des protéines comme une molécule bien définie soit reconnue par la communauté. Même après que cette définition fut acceptée, les observations des protéines et de leurs rôles n'ont pas toujours été admises et furent toujours sujet à controverse. Des avancées majeures ont notamment été ignorées pendant des années comme la cristallisation de l'uréase[7]. Depuis longtemps l'avancée autour des protéines est en dents de scie, avec des bonds majeurs suivis de périodes de recul ou de stagnation. Certains concepts furent énoncés bien avant que les moyens techniques permettent de les valider. La notion de protéines a beaucoup évoluée au cours des dernières années jusqu'à avoir plusieurs niveaux de définitions. Il existe des définitions dites structurales ou fonctionnelles d'une protéine. D'un point de vue fonctionnel, une protéine est une entité formée d'une ou plusieurs chaines peptidiques ayant une fonction définie dans un organisme. D'un point de vue structural, une protéine peut être définie comme une chaine peptidique.

1.2.1. Anatomie

La résolution des structures cristallines[8] de protéines a permis les plus grandes avancées sur ces macromolécules. Connaitre la composition et la forme tridimensionnelle des protéines permet en effet de bien mieux les caractériser. De ce fait, une protéine possède plusieurs niveaux de repliement (**Figure 1.2**). Elle est composée d'une structure primaire qui est une suite d'acides aminés. La structure secondaire est le repliement de cette chaîne en blocs, il en existe 3 types les hélices alpha, les brins bêtas et les boucles. La structure tertiaire est l'assemblement des structures secondaires. Enfin les structures quaternaires sont les assemblements de chaînes entre elles, ces assemblages sont la cible de cette thèse. La définition de protéines est floue à ce niveau-là, les protéines pouvant être une chaîne peptidique ou un assemblage de plusieurs chaînes. Pour le reste de cette étude, la définition de protéine correspond à la structure tertiaire des protéines, c'est-à-dire une suite unique et continue d'acides aminés structurés en trois dimensions.

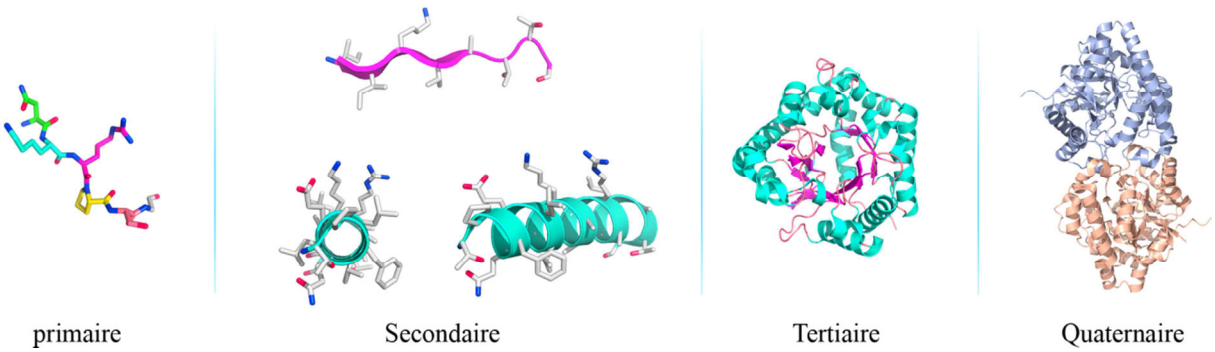


***Figure 1.2** : Représentation des différents niveaux de structures applicables a une protéine. La structure primaire est l'enchainement propre d'acides aminés, chaque couleur correspond à un acide aminé. La structure secondaire est l'alignement tridimensionnel spécifique des acides aminé, les brins béta sont en rose et l'hélice alpha en cyan. La structure tertiaire est l'agencement tridimensionnel des structures secondaires. La structure dite quaternaire est l'assemblement de plusieurs protéines repliées. La structure quaternaire est le niveau qui nous intéresse le plus pour cette thèse.*

Les protéines sont composées de 20 acides aminés standards, la forme générale de la protéine dépendent de l'enchainement en acides aminés[9]. Les acides aminés vont jouer un rôle sur les structures secondaires mais aussi tertiaires de la protéine, les acides aminés dit hydrophobes (isoleucine, valine, leucine, cystéine, méthionine, alanine et phénylalanine) auront tendance à composer le cœur de la protéine pour laisser des acides aminés plus polaires au contact avec le solvant.

1.2.2. Classes

Les protéines peuvent être classées en plusieurs catégories selon des critères distincts. Elles sont en général classées selon leur forme, leur fonction ou même encore selon leur localisation au sein de la cellule. Ces classements se regroupent la plupart du temps mais il n'existe pas encore de classement universel. Le classement structural principal, SCOP[10] repose uniquement sur la description des protéines, il existe aussi des classements basés sur l'homologie de séquence tel que CATH[11]. SCOP trie les protéines dans 7 catégories principales, les protéines tout alpha, les protéines tout béta, deux catégories avec des hélices alphas et des brins bêtas, les protéines à domaines distincts, les protéines membranaires et les petites protéines. Pour la recherche pharmaceutique, ce classement est boudé au profit d'un classement fonctionnel des protéines les regroupant par rôle biologique au sein de la cellule. La principale banque de données classant les protéines par fonction se nomme GO. Le projet Gene Ontology (GO) fournit un vocabulaire contrôlé pour décrire un gène et l'ensemble des

produits du gène dans tout l'organisme. Ce vocabulaire contrôlé est développé indépendamment des bases de données existantes. Il y a 3 catégories disjointes : composant cellulaire, fonction moléculaire et processus biologique. Une tâche importante est de cartographier les termes GO[12] avec les produits de gènes (protéine) dans les bases de données via l'annotation automatique (électronique) ou manuelle. GO est un projet ambitieux mais le résultat de l'annotation sur les protéines est souvent inutilisable. La base de données principale de protéines (Uniprot) a composé son propre classement réalisé manuellement (au sein de la SwissProt) et basé sur le vocabulaire de GO. Les classes permettent d'avoir une information sur la structure quaternaire des protéines par homologie avec les protéines de la même famille connue.

1.3. Interface protéine-protéine (structure quaternaire)

1.3.1. Zone d'interaction

La taille des surfaces d'interaction entre les protéines est très variable[13], allant de 200 à 4000 Å². Les zones d'interactions **(Figure 1.3)** entre les protéines sont généralement décrites comme grandes et avec peu de reliefs. L'étude de ces zones est un élément clé de la compréhension des mécanismes d'interaction entre les protéines.

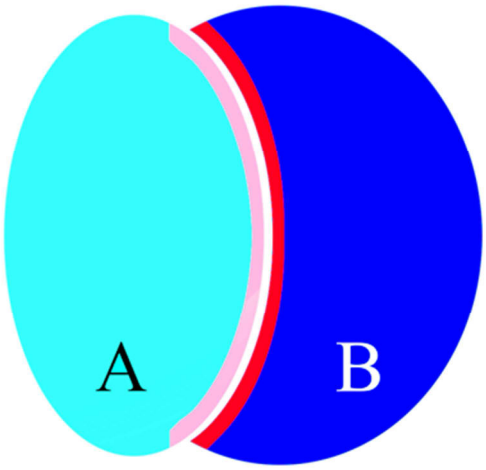


***Figure 1.3**: Représentation d'une zone d'interaction entre deux protéines A et B. Les résidus des protéines A et B réalisant une interaction sont colorés en rose et rouge, respectivement. La somme des interactions crée ce que nous appelons la zone d'interaction.*

1.3.2. Flexibilité

Divers types d'interfaces sont possibles en fonction de la flexibilité structurale des partenaires. Une protéine n'est pas un ensemble rigide mais un assemblage d'éléments mobiles ordonnés. Malgré la structure imposée à la protéine, les chaînes latérales des acides aminés sont mobiles tout comme certaines boucles. Les boucles sont des chaines d'acides aminés structurés de manière moins rigide que le reste de la protéine. Elles peuvent être tellement flexibles qu'il arrive que l'on ne puisse pas déterminer leur position par diffraction des rayons X. Ces éléments mobiles font que la liaison des deux protéines en interaction n'est jamais complétement rigide. L'échelle de modification de la structure varie d'aucun

changement conformationel à un changement conformationnel radical d'une des deux protéines. Le changement conformationnel de la structure des protéines est un des plus gros problèmes pour la compréhension des mécanismes d'interaction[14]. Différents modèles de reconnaissance se sont succédés au cours du temps avant de cohabiter:

i) le modèle rigide (clef serrure) implique la fixation de deux structures rigides (Emil Fisher 1894);

ii) ii) le modèle par ajustement induit ("induced fit") décrit chaque partenaire comme légèrement flexible[15]. Ce modèle a été mis en place car la complémentarité entre les protéines partenaires est beaucoup plus faible sur les structures résolues séparément que sur la structure du complexe correspondant. Le changement de conformation est influencé par des contraintes électrostatiques, l'énergie de desolvatation[16] ou des interactions de van der Waals. Les transformations citées ici ont lieu après le phénomène de reconnaissance des protéines;

iii) iii) le modèle "tout mobile"[17,18], décrit les structures des protéines résolues comme incomplètes, les protéines alternant entre un ensemble de conformations inconnues parmi lesquelles sont présentes celles permettant la reconnaissance protéine-protéine

La coexistence de ces trois modèles montre que nous ne connaissons encore qu'imparfaitement les mécanismes de reconnaissance et de fixation des protéines, cette difficulté se retrouvant aussi au niveau des techniques experimentales qui essayent de prédire les interactions protéine-protéine.

1.3.3.Dynamique

Le modèle prédisant que les structures résolues ne sont qu'une image d'une interaction mobile a incité à l'étude de la dynamique des interfaces protéine-protéine. La méthode utilisée par les protéines pour se reconnaitre est encore méconnue mais nous avons des informations sur la stabilité des interfaces. Certains complexes sont souvent résolus par diffraction des rayons X mais aussi grâce à la résonance magnétique nucléaire (RMN.) La RMN apporte une flexibilité et une liberté impossible en diffraction. Ces données ont confirmé la présence de plusieurs modes d'interaction possible entre deux partenaires. Des études en dynamique moléculaire sur des interfaces connues ont aussi montré la présence de plusieurs zones d'interaction[19]. Les dynamiques de 2µs montrent l'alternance entre différente formes stables toutes les 50-100ns. L'étude montre une réorganisation de la zone d'interaction entre chaque forme, par le biais de liaisons hydrogènes et de ponts salins. La connaissance des transitions entre les différentes interactions est un élément clé pour l'avenir de la modulation des interfaces.

1.3.4.Statut d'oligomérisation

Le statut d'oligomérisation (**Figure 1.4**) correspond à la structure quaternaire quand on parle de détermination d'interface protéine-protéine.

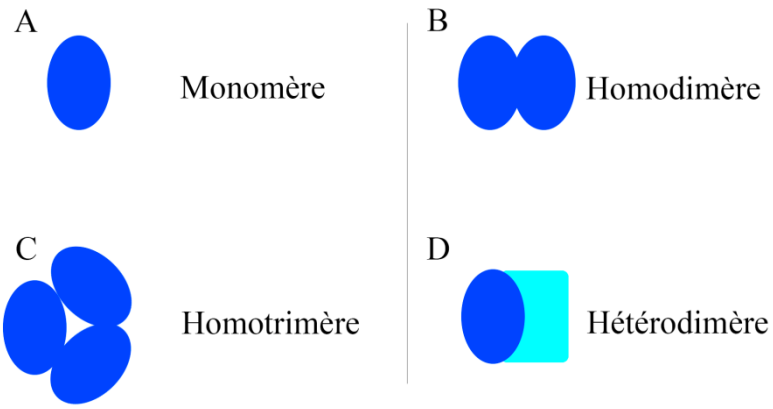


***Figure 1.4**: Modèle de représentation de différents statuts d'oligomérisation. A. le statut monomère correspond à une protéine seule ; B. Le statut homodimère correspond à deux protéines identiques liées entre elles ; C. La même protéine s'associe trois fois ; D. Tous les statuts sont dits homo- pour une association de protéines identiques ou hétéro- s'il s'agit de protéines différentes. Les unités biologiques les plus grosses actuellement (ribosomes d'eukaryotes) comprennent plus de 120 protéines.*

La détermination expérimentale du statut d'oligomérisation d'une protéine en solution est particulièrement difficile. Il existe beaucoup de méthodes de caractérisation de l'état d'oligomérisation des protéines reposant sur des techniques différentes et qui ont des cibles variées. On peut citer parmi les plus récentes :

- i) Le croisement du couplage covalent avec la spectrométrie de masse[20] qui observe les contacts au niveau du résidu. Cette méthode donne beaucoup d'informations mais requiert beaucoup de matériel protéique ainsi que de ressources informatiques ;
- ii) la vitesse de sédimentation par ultracentrifugation[21] qui donne des informations à propos de l'arrangement irrégulier du complexe. Cette technique a l'avantage de donner des informations très précises et n'a pas besoin de marquage des protéines cibles, cependant elle coûte très cher et ne fonctionne pas en présence de détergents ;
- iii) La chromatographie par exclusion de taille[22] donne des informations sur les stœchiométrie globale du complexe mais fonctionne mal sur des protéines membranaires.

Ce ne sont que des exemples parmi une longue liste d'outils (**Tableau 1.1**) disponibles afin d'analyser les complexes protéine-protéine expérimentalement. La confirmation du statut d'oligomérisation d'un complexe protéique est souvent validée après la comparaison de résultats issus de plusieurs méthodes. Les méthodes expérimentales prouvent souvent la taille du complexe (nombre de chaînes impliquées) mais il est très difficile de faire le lien avec la structure protéique. Les partenaires de l'interface (stœchiométrie globale) sont connus mais pas l'interface en elle-même.

Tableau 1.1: *Tableau non exhaustif des méthodes expérimentales d'assignation de statut d'oligomérisation et de structure quaternaire de protéines.*

Méthode	Résultats	Avantages	Inconvénients
Chromatographie par exclusion de taille[22,23]	Stœchiométrie globale	- protéine native - Calcul de l'équilibre de l'interaction - Condition variables	- Masse imprécise - Résultats indirects
Diffusion de la lumière multi-angles (MALS)[24]	Stoechiométrie globale	- protéine native - masse précise - fonctionne avec détergents	- Besoin de calibration - Surveillance constante nécessaire
Ultracentrifugation[21]	Stoechiométrie globale	- masse précise - protéine native	- Coute cher - Analyse des résultats difficile en cas de mélange
Diffusion des rayons X aux petits angles (SAXS)[25]	Arrangement brut du complexe	- protéine native - donne la forme du complexe	- Qualité des résultats variable - Utilisation d'un synchrotron
Diffusion de la lumière dynamique (DLS)[26]	Arrangement brut du complexe	- protéine native - peu de protéine nécessaire - pas de perte de protéine	- Faible résolution - Les agrégats pose problème
Vitesse de sédimentation avec centrifugation analytique[21]	Arrangement brut du complexe	- masse précise - protéine native	- Coute cher - Masse moléculaire mal défini
Microscopie électronique[27]	Arrangement brut du complexe	- marche sur des données à faible résolution - protéine native	- Ne fonctionne pas sur les petites protéines - Problème sur les sous-unités biologiques
Le couplage covalent-spectrométrie de masse (MS)[20]	Contacts entre les résidus	- Donne la structure tertiaire - Marche sur les assemblages complexes - Utilisable à haut débit	- Matériel coûteux - Besoin informatique élevé
Impression chimique et MS[28]	Contacts entre les résidus	- Décrit les résidus d'interfaces - Facile d'utilisation - Facile d'utilisation	- Matériel coûteux - Peu répandu
Mutagénèse dirigée[29]	Contacts entre les résidus	- Décrit les résidus d'interfaces - Définit les points chauds d'interaction	- Besoin de connaissance structurale - Besoin d'information complémentaire

1.4. Informatisation des méthodes d'étude des interfaces protéine-protéine

1.4.1. Diffraction des rayons X

La diffraction de rayon X est une méthode visant à observer la diffraction résultante des électrons d'une protéine. La diffraction de ceux-ci étant trop faible pour être concrètement observée, cette méthode se base sur la redondance des électrons présents dans un solide voire même plus précisément dans un cristal pur. L'avantage de l'état cristallin est la répétition quasi parfaite des éléments au sein de celui-ci, il y a une bonne amplification du signal. L'utilisation de cristaux est l'élément clé de la méthode mais aussi son principal défaut, une protéine à l'état biologique étant dans un environnement aqueux. De plus, il est difficile de prédire l'effet de la cristallisation sur la conformation de la protéine cible. Cette technique nécessite d'avoir des protéines peu flexibles et parfaitement alignées. La création de cristaux de protéine est un procédé complexe et capricieux mais essentiel. La qualité du cristal est directement reflétée dans la précision de la structure (résolution). Le résultat de la diffraction électronique est une carte tridimensionnelle de la densité atomique représentant la probabilité de présence d'un électron pour chaque ensemble de coordonnées possible. La carte résultante c'est à dire la densité électronique est alors interprétée informatiquement pour déterminer l'emplacement de chaque atome. La partie résolue est ce que l'on appelle la maille cristallographique, c'est le plus petit élément qui par des opérations de translation peut reconstruire l'ensemble du cristal (Figure 1.5). A l'intérieur de cette maille se retrouve l'unité asymétrique (la structure principalement distribuée dans la base de données PDB) qui contient la plus petite portion de structure permettant par symétrie de retrouver la maille cristalline.
Dans la plupart des cas, l'unité asymétrique correspond à l'état d'oligomérisation de la protéine native. Un élément clé avec cette méthode est que le cristal de molécules contient un grand nombre d'interfaces appelées contacts cristallins ou plus communément interfaces cristallines. Elles sont la plus part du temps indissociables des interfaces dites biologiques. La complexification des structures de ces 20 dernières années a mis en avant le problème de distinction entre les interfaces biologiques et les contacts cristallins, les erreurs sont devenues un problème récurrent.

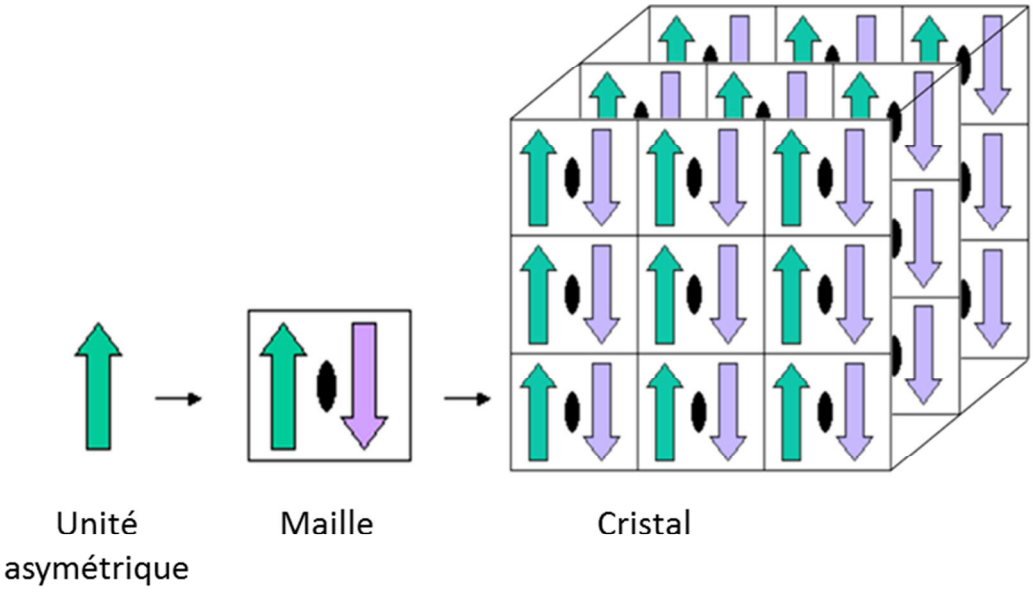


Figure 1.5: *Représentation de l'unité asymétrique (source site web de la RSCB PDB). L'unité asymétrique permet de recréer la maille par des opérations de symétrie. La maille permet de recréer l'ensemble du cristal grâce à des translations.*

1.4.2. Prédiction de la véracité des interfaces présente dans les structures de protéine

Statuts d'oligomerisation

Le statut d'oligomérisation est aussi étudié à partir des structures 3D résolues par diffraction des rayons X. Depuis des années, les cristallographes sont incités à décrire dans les fichiers de structures l'unité biologique qu'ils pensent être la plus probable dans l'unité asymétrique. La PDB stocke cette information dans la REMARK 350 du fichier PDB (ou la catégorie pdb_struct_assembly dans les fichiers mmcif) afin de le décrire sur le site web[30]. L'information soumise par les cristallographes est souvent remise en cause car composée de données expérimentales manquantes ou erronées lors du dépôt de la structure. Des études ont estimé que le taux d'erreur présent dans les unités biologiques, en se basant sur des notions de similarités de séquences et d'homologie, était de l'ordre de 15%[31]. Les informations décrites par les cristallographes décrivent les structures qu'ils ont réussi à résoudre. Du fait que les protéines sont des molécules flexibles, la structure résolue n'est qu'une conformation parmi plusieurs possibles. Les interactions entre les protéines auraient plusieurs états plus ou moins stables et nous observerions un état d'équilibre dans les structures cristallines. Dans cet état

dit d'équilibre, toutes les interfaces présentes ne sont donc pas forcément dans leur forme la plus stable en solution. Cette observation peut expliquer pourquoi l'énergie de dissociation des interfaces ainsi que leur taille ne sont pas de bons critères afin de prédire la véracité biologique d'une interface. Certaines interfaces ont d'ailleurs des énergies libres de dissociation du même ordre voir plus faibles que celle des contacts cristallins pouvant induire un mauvais jugement et même l'impossibilité de cristalliser une structure contenant l'interaction[32]. Deux entités ont complété ces informations au sein de plusieurs bases de données (PiQSi[31] et PDBWiki[33]) contenant les informations quaternaire des protéines basées notamment sur l'homologie de séquence et vérifiées à la main.

De nombreuses méthodes de prédiction de la pertinence biologique d'une interface protéine-protéine sont disponibles (**Tableau 1.2**). Elles sont décrites plus précisément dans le Chapitre 2 de cette thèse.

Tableau 1.2: *Tableau récapitulatif des outils de classification des interfaces protéine-protéine*

Méthode	Propriété prédite	Précision	Jeu de données	Serveur (URL)
Janin[34]	Aire d'interface	–	Janin monomères	×
PQS[35]	Différence de SAS[a]	–	–	Hors ligne
PITA[36]	SAS enfouie Score des atomes appariés	0.81	Ponstingl2003	http://www.ebi.ac.uk/thornton-srv/databases/pita/
PISA[37]	ΔG_{diss} estimée	0.90	Ponstingl2003	http://www.ebi.ac.uk/pdbe/pisa/
NOXclass[38]	Probabilité estimée par SVM	0.92	BNCP-CS	http://noxclass.bioinf.mpi-inf.mpg.de/
DiMoVo[39]	Tessellation de Voronoi	0.90 0.83	BNCP-CS Ponstingl2003	Hors ligne
COMP[40]	Complémentarité de 3 propriétés de surface	0.89	Tschuchiya2008	Hors ligne
CRK[41]	*Ratio Ka=Ks* ratios A partir des séquences codantes	0.84	Schärer2010	×
IPAC[42]	Vecteur Bayésien Naif	≥ 0.90	Ponstingl2000 Ponstingl2003 Benchmark3.0 (Hwang et al., 2008)	✓
EPPIC[43]	Nombre de résidus de noyau et homologie de séquence	0.90	Ponstingl2003 (monomères and dimères)	✓
ECR[44]	Géométrie	0.78 0.80	Test-18 Dey2010 (Dey et al., 2010)	×
Luo et al.[45]	Forêt aléatoire 46 descripteurs	0.91 0.92	Bernauer2008 Ponstingl2003	×
Liu et al.[46]	B-facteur	0.94 0.90 0.87	BNCP-CS Ponstingl2000 Bahadur2004	×
PiQSi[31]	Inférence des structures quaternaires par homologie	0.95	Levy2007	×
ProtCID[47]	Conservation des interfaces au sein du cristal	–	Ponstingl2000 Bahadur2004	×

Nous ne mentionnerons ici que deux méthodes, l'une d'entre elles (PISA) étant utilisée par la Protein Data Bank pour caractériser la structure quaternaire de ses entrées, l'autre (EPPIC) étant très récente et présentant l'avantage d'avoir été entrainée sur un jeu de données récentes.

PISA

PISA est le principal outil de détermination de statut d'oligomérisation au sein des structures des protéines[48]. Le but principal est de définir la forme la plus pertinente de la structure présente dans un fichier de coordonnées. Il cherche à définir l'assemblement optimal, c'est à dire le plus stable dans une unité asymétrique. Le fonctionnement de PISA suit différentes étapes (**Tableau 1.3**).

Tableau 1.3: *Tableau récapitulatif des trois actions principales réalisées par PISA*

Etape	Méthodes
1	Vérification de la symétrie
2	Calcul d'énergie
3	Classement des interfaces (taille > complexité > énergie)

La première vérification est une étape purement géométrique suivant les deux règles ci-dessous :

- si une interface est validée, toutes les interfaces similaires sont validées.

- Il ne peut pas y avoir d'interfaces entre deux protéines identiques parallèlement positionnées, c'est-à-dire qu'une translation n'est pas une symétrie suffisante pour que l'interface soit biologiquement viable

La seconde règle est une dérivée logique de la première qui fait que si deux protéines parfaitement alignées sont en interaction, par reproduction de l'interface nous obtenons une répétition du complexe de manière illimité.

Afin de déduire la stabilité chimique d'un ensemble complexe, il faut calculer l'énergie libre de Gibbs, l'énergie de dissociation ΔG^0_{diss}, en utilisant l'équation.

$$\Delta G^0_{diss} = -\Delta G_{int} - T\Delta S \quad (1)$$

Ou ΔG_{int} est l'enthalpie d'interaction, plus précisément la somme de l'énergie de toutes les interactions électrostatique, et $T\Delta S$ est un modèle simplifié d'entropie. Comme

cela a été discuté dans la stabilité des macromoléculaires complexes, PISA suppose que les assemblées se dissocient en sous-unités stables le long d'un motif avec $une\ valeur\ \Delta G^0_{diss}$ minimale. En outre, des considérations de symétrie, le motif de dissociation ne peut pas contenir une interface engagée si une interface équivalente est désengagée quelque part ailleurs dans le cristal. Si dans un complexe, une des interactions a un ΔG^0_{diss} négatif, l'ensemble du complexe ou la sous-unité impliquée est instable. PISA considère que l'identification de dissociation probable des motifs est un sous-produit important de la méthode. Il y a finalement trois règles pour déterminer les meilleures interfaces : i) Les assemblées de grande taille sont prioritaires sur les petites ; ii) les homomères sont prioritaires vis à vis des hétéromères ; iii) les assemblées avec énergie libre de dissociation ΔG^0_{diss} élevée sont préférées à celles avec une valeur de ΔG^0_{diss} inférieure.

EPPIC

EPPIC pour « Evolutionary Protein-Protein Interface Classifier »[43] est un classifieur d'interaction protéine-protéine discriminant les interfaces pertinentes des contacts cristallins basés sur des critères évolutionnaires. EPPIC est intéressant car il prend en compte beaucoup de descripteurs mis au point au cours des années sur les interfaces protéine-protéine. Après plusieurs versions basées uniquement sur des critères évolutionnaires la dernière version d'EPPIC rajoute deux nouveaux critères afin d'améliorer les prédictions du classifieur. Les trois critères pour la classification sont donc :

- le critère évolutionnaire basé sur de l'homologie de séquence avec des structures connues ;
- un critère géométrique: la taille du cœur de l'interface ;
- une estimation de la pression de sélection: l'entropie de séquence.

Le critère géométrique a été utilisé au cours de différentes études des interfaces protéine-protéine, il s'agit généralement d'un critère basé sur une classification des résidus présents à l'interface. L'idée de diviser les résidus de l'interface, à savoir ceux qui plus enfouis qu'un seuil déterminé, dans différentes classes sont apparues au début de la littérature sur les interfaces protéine-protéine. Une équipe a proposé une première classification basée sur les atomes plutôt que des résidus[49], en les divisant en 3 classes qu'ils appelaient A, B et C. La classe B contenant les atomes les plus enfouis, les deux autres classes A et C correspondant à des atomes moins enfouis. Cette classification a évolué autour de la notion de résidus de base comme les résidus ayant au moins une partie de leur chaine latérale entièrement enfouie[50]. La

troisième étape de la classification des résidus d'interface se fait sur les résidus dont l'enfouissement change lors de la dimérisation, un résidu enfoui à plus de 95% est un résidu du noyau d'interface. EPPIC utilise désormais une classification à catégories niveau, résidus noyau, jante et de soutien[51]. EPPIC a étudié les différents critères avant de conserver la méthode basé sur le pourcentage d'enfouissement qu'ils décrivent comme le plus discriminant, ainsi une interface avec plus de 6 résidus de cœur sont considérés comme biologiquement pertinente.

Le paramètre entropie de séquence est une recherche de protéine par similarité de séquence. Ils regroupent entre eux des protéines avec plus de 60% d'identité de séquence et descendent à 50% dans le cas de groupes de plus de 10 séquences. Une interaction est définie pertinente si les protéines du groupe correspondant possèdent au moins 8 résidus du cœur en commun et enfouis à plus de 70%.

1.4.3.Jeux de données d'interfaces protéine-protéine

Le principal problème des classifieurs d'interface protéine-protéine est que la majorité des outils développés au cours des années sont basés sur de structures et des jeux de données anciens et déséquilibrés. Bien que la rigueur avec laquelle le jeu d'apprentissage a été construit soit un élément clé de succès de toute prédiction, peu de groupes de recherche se sont intéressés sérieusement à ce problème. Dans les jeux de données classiquement utilisés, nous avons très vite identifié deux problèmes majeurs remettant en cause la qualité des outils disponibles :

- i) la classification des interfaces dans une classe (biologique) ou une autre (contact cristallin) n'est pas toujours vérifiée par des données expérimentales obtenues depuis lors. De même, beaucoup de structures contenant des contacts cristallins sont qualitativement désuètes.

- ii) Les données sont totalement déséquilibrées vis-à-vis du simple critère de la taille de l'interface, les interfaces cristallographiques étant d'une taille très largement inférieure à celle des interfaces biologiquement relevantes (**Figure 1.6**).

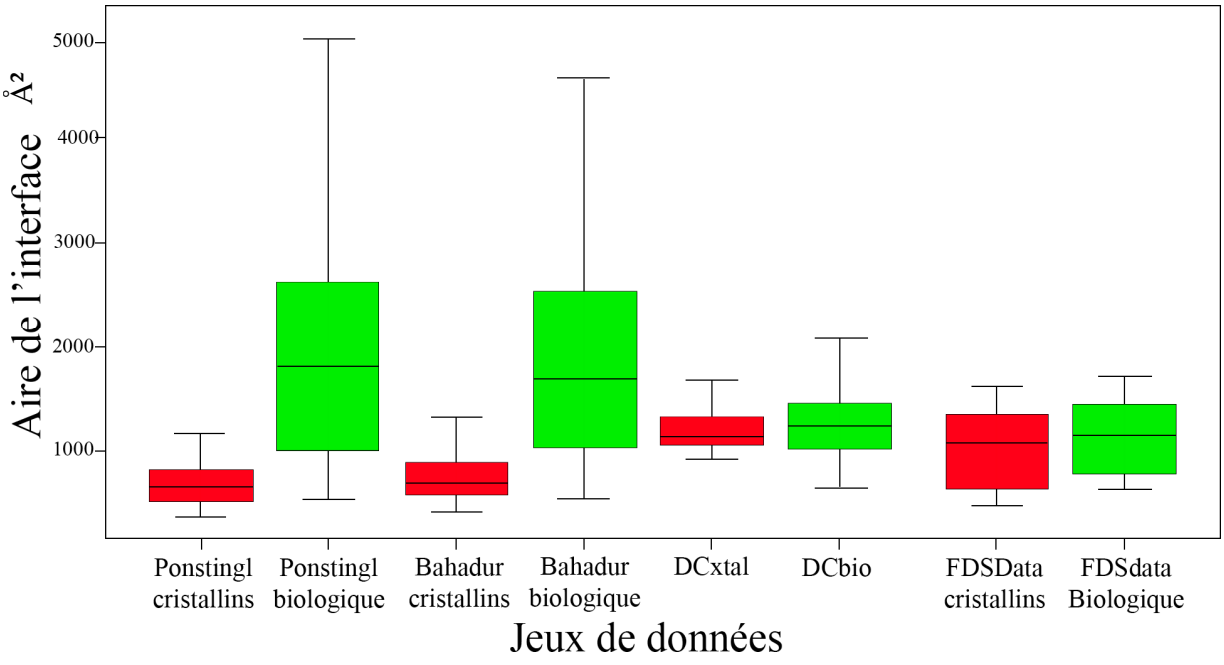


Figure 1.6: *Boites à moustaches présentant les deux jeux les plus utilisés (Ponstingl et Bahadur), le dernier jeu (DC) créé pour EPPIC, et le jeu de données (FDS) créé lors de cette thèse.*

Ponstingl

Le jeu de données Ponstingl[52] est le premier jeu de données utilisé par la communauté afin d'étudier les classifieurs d'interfaces protéine-protéine. Ce jeu est composé de 172 protéines non-redondantes qui ont pour caractéristiques d'être solubles dans l'eau. Le statut oligomérique en solution est connu grâce à des données expérimentales (monomère ou multimère).

Bahadur

Bahadur[53] est un jeu de données publié en 2004 créé pour étudier la géométrie et les propriétés physico-chimiques des interfaces protéine-protéine. Il contient 70 hétérodimères et 122 homodimères biologiquement pertinents. Les contacts cristallins sont représentés par 188 homodimères réalisant des contacts semblables aux interfaces biologiquement pertinentes. Le principal problème de ce jeu de données est la très faible taille des interfaces définies comme contacts cristallins.

Duarte-Capitani (DC)

Le jeu de données DC[43] est le premier à prendre en compte les erreurs du passé, il a été créé afin de mieux déterminer l'efficacité des classifieurs d'interface protéine-protéine. Pour créer le jeu, les auteurs se sont dirigés vers 3 axes d'optimisation principaux :

i) utiliser uniquement les entrées pour lesquelles la structure oligomérique est clairement vérifiée expérimentalement;
ii) utiliser des structures de contacts cristallins avec des filtres beaucoup plus stricts: résolution inférieure à 1.8Å, rfree inférieur à 30% et un seuil d'identité de séquence de 90%. Sur les structures dont les données expérimentales prouvent que l'homodimère n'existe pas, les cristaux sont reconstruits afin de générer les interfaces cristallines;
iii) Ne conserver que les interfaces d'une taille variant entre 1 000 et 1 500Å².

Le jeu de données est un très bon jeu de validation et de comparaison de classifieurs. Le principal problème est le troisième axe qui supprime une grande partie d'interfaces biologiquement pertinentes (ex: p53-mdm2) dont la surface est inférieure à la valeur seuil de 1000 $Å^2$.

FDS

La méthode ayant conduit à la création de ce jeu de données[54] est décrite en détail dans le chapitre 2. Ce jeu de données est composé de 200 interfaces protéine-protéine biologiquement pertinentes (sélectionnées et curées manuellement) et 200 interfaces correspondant à des contacts cristallins impliquant des protéines connues pour exister à l'état monomérique en solution.

1.4.4.Bases de données d'interfaces protéine-protéine

La rapidité avec laquelle les données génomiques et protéomiques sont générées nécessite le développement d'outils et de ressources pour la gestion des données qui permettent l'intégration des informations provenant de sources disparates.

Hprd

La base de données de référence de protéines humaines (http://www.hprd.org) est une ressource Web basée sur des technologies "open source" stockant pour chaque protéine un nombre considérable d'informations dont les interactions protéine-protéine, les modifications post-traductionnelles, les relations enzyme-substrat et les associations avec des maladies (**Figure 1.7**).

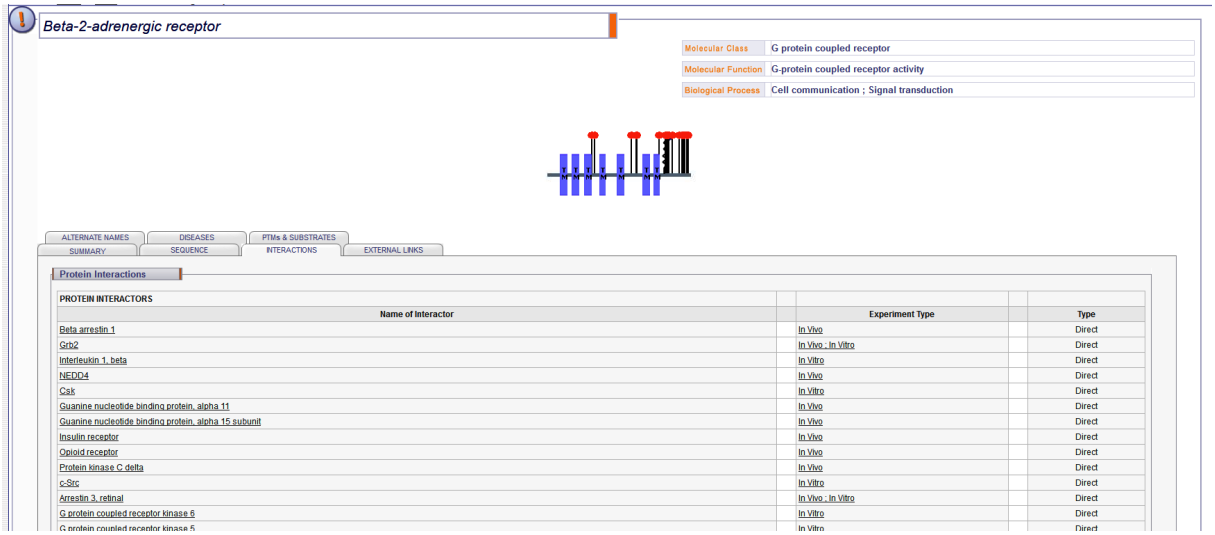


Figure 1.7: *Capture d'écran des informations disponibles pour le récepteur béta2 adrénergique humain dans la base de données Hprd.*

Cette information a été extraite manuellement par une lecture critique de la littérature publiée par les biologistes et à travers l'analyse bio-informatique des séquences protéiques. Cette base de données a pour visée une utilisation médicale. La complexité des données de protéines sont difficiles à présenter sans méthodes de visualisation appropriées. Bien sûr, il serait désirable de profiter d'une vision intégrée alliant données de génomique ainsi de protéomique comme cela a été démontré dans le cas de certaines voies métaboliques dans la levure[55].

BioGRID

La "Biological General Repository for Interaction Datasets" (BioGRID: http://thebiogrid.org) est une base de données d'accès ouverte qui abrite les interactions génétiques, chimiques et de protéines issus de la littérature pour tous les principaux modèles d'organisme (**Figure 1.8**).

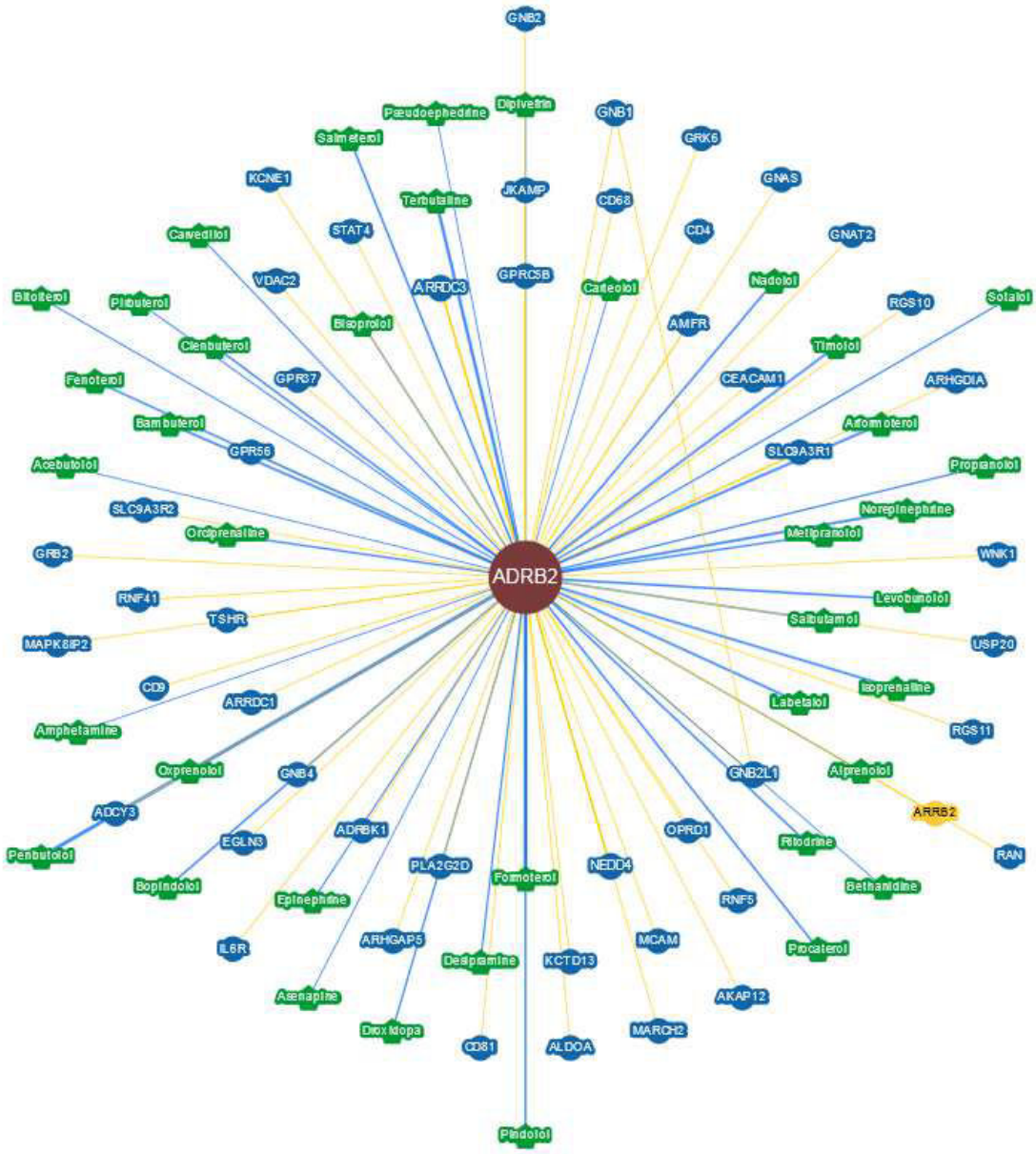


Figure 1.8: *Capture d'écran des informations disponibles pour le gène du récepteur béta2 adrénergique humain dans la base de données BioGRID. L'interactome est centré sur la requête et affiche à la fois les gènes (même espèce, bleu; autre espèce: jaune) et les ligands (vert) en interaction. directes(traits jaunes et bleus) .*

BioGRID contient aujourd'hui 1 069 563 interactions dont 468 948 interactions non redondantes, décrites comme physiques. Il est à noter que BioGRID possède des algorithmes d'apprentissage supervisés de traitement des publications développés au cours du temps afin de faciliter son actualisation.

EPPICDB

EPPICDB[56] est un site web (http://www.eppic-web.org/ewui/) permettant de prédire l'état d'oligomérisation de 99.32 % des structures présentes dans la PDB. EPPICDB montre pour chaque entrée PDB la structure de l'unité asymétrique ainsi que l'ensemble des interfaces réalisables deux à deux grâce aux axes de symétrie fournis par les auteurs (**Figure 1.9**).

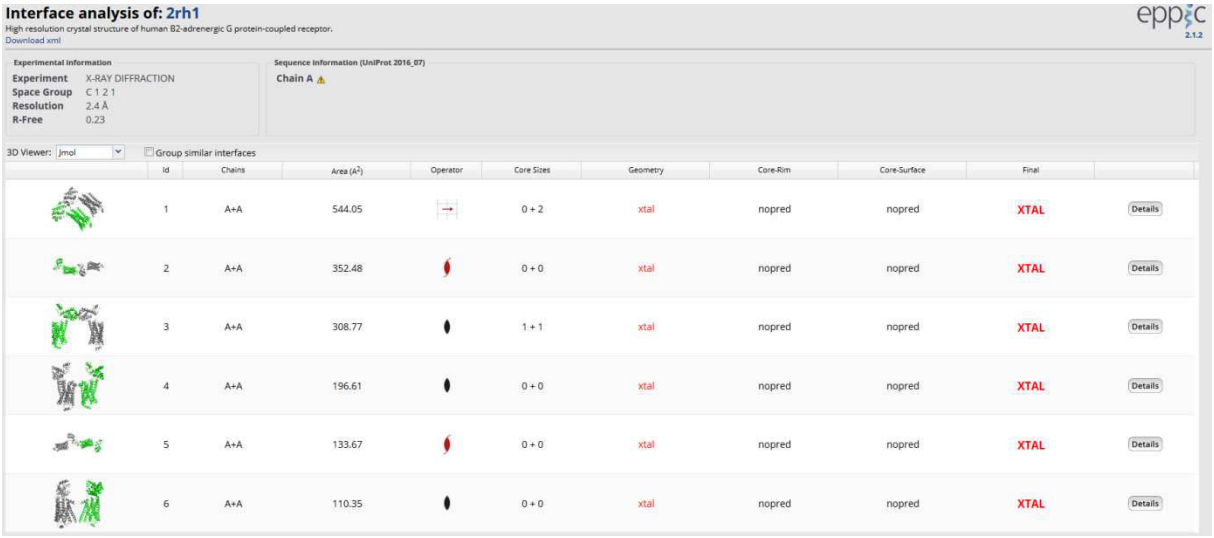


Figure 1.9: *Capture d'écran des informations disponibles pour le récepteur béta2 adrénergique humain (PDB ID 2rh1) dans la base de données EPPICDB.*

Elle se présente comme la meilleure alternative aux informations données par PISA dans la PDB. Les nouvelles entrées sont mise à jour tous les mois à l'aide de leur outil de prédiction.

1.5. Modulation des interfaces protéine-protéine par des petites molécules

1.5.1. Introduction

Historiquement, les ligands endogènes ont été identifiés bien avant que leurs récepteurs ne soient isolés et que leurs structures ne soient résolues[4]. Les premières informations sur la structure chimique du ligand étaient basiques et rares par rapport aux informations disponibles actuellement. Toutefois, ces informations précoces se sont souvent avérées suffisantes pour déduire la pertinence physiologique des ligands même si les mécanismes physiologiques sous-jacents restaient inconnus. Aujourd'hui, la découverte de candidats médicaments est souvent contrainte par la cible biologique choisie, l'existence de modèles animaux transgéniques, la biologie moléculaire, ou encore le génie génétique. Par conséquent, la connaissance de la structure de la protéine cible est un atout crucial, pour prédire et visualiser des ligands liés à leur cible. De nombreuses techniques expérimentales ont été mises au point afin de visualiser la présence d'interaction entre une protéine et une molécule, les structures de nombreux complexes protéine-ligand ont été résolues. Cependant les outils chemoinformatiques existants ne sont toujours pas capables de décrire précisément les mécanismes impliquant la reconnaissance d'un ligand par sa protéine cible.

1.5.2. Petites molécules aux interfaces protéine-protéine

Les interfaces protéine-protéine sont des éléments clés dans tous les processus biologiques et leur dérégulation implique souvent des maladies, cela en fait des cibles thérapeutiques privilégiées. L'importance de ces cibles n'influe malheureusement en rien sur la difficulté à les moduler de manière rationnelle par des petites molécules. Ces interfaces étaient encore considérées comme non droguables il y a une vingtaine d'années, notamment par l'absence de cavités profondes et enfouies et de par leur taille considérable (1000-2000Å^2 [57]).

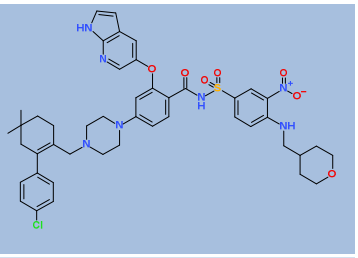
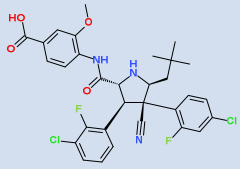
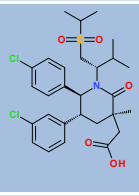
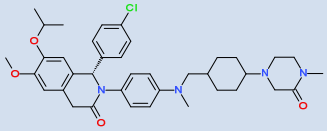
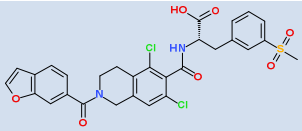
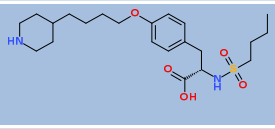
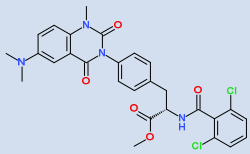
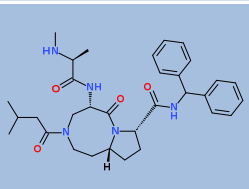
Un autre obstacle provenait de l'absence de petits ligands endogènes pouvant être utilisés comme support de modification structurale afin d'identifier les premières touches

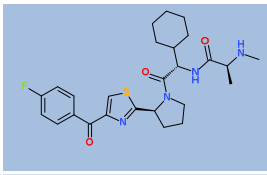
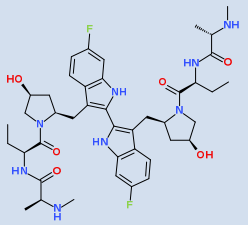
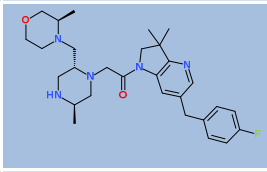
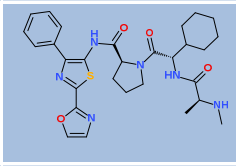
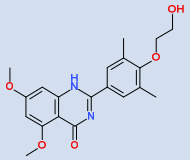
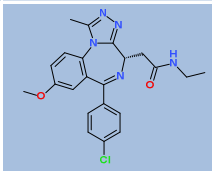
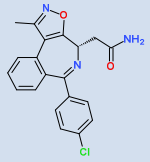
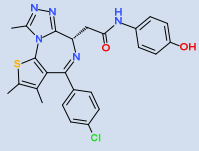
d'intérêt. La recherche d'inhibiteurs d'interfaces change à partir de 1995 quand est découvert le fonctionement des molécules naturelles comme les taxanes, la rapamycine[58] et les cyclosporines. Quelques années plus tard, le tirofiban, un antagoniste d'intégrine rentre en phase clinique. Des études de mutagénèse dirigée ont montré que, dans l'interface, toutes les interactions n'ont pas la même importance. Il existe des aminoacides clés appelés points chauds indispensables à la stabilité du complexes[4,59,60].L'avantage de ces points chauds est qu'ils sont souvent situés au cœur de l'interface protéine-protéine et qu'ils sont étalés sur une surface pouvant être comparé à la taille de petites molécules médicamenteuses. Ce sont des zones hydrophobes montrant une flexibilité assez élevée. Ces caractéristiques nous montrent que certaines interfaces protéine-protéine pourraient avoir des zones d'interaction de taille suffisamment petite pour être efficacement modulées par des molécules "drug-like ".

Depuis lors, des inhibiteurs puissants ont été développés pour des complexes de protéines diverses. En 2016, on recense 19 composés en phase d'études clinique active (**Tableau 1.4**). Beaucoup de ces candidats cliniques d'inhibition d'interface ont une efficacité élevée, typiques de molécules "drug-like" et sont disponibles par voie orale. De manière intéressante, les inhibiteurs d'interface les plus prometteurs sont issus de travaux multidisciplinaires alliant entre autres outils moderne d'analyse, étude de la structure tridimensionnelle, biomarqueurs et calculs théoriques. Les propriétés de ces inhibiteurs[61] sont légèrement différentes de celles des molécules "drug-like" classiques, par exemple l'efficacité du ligand (ΔG/HA) des inhibiteurs d'interface protéine-protéine est autour de 0.20 ce qui est inférieur à la valeur de 0.3 classiquement observées pour des inhibiteurs de protéines globulaires. Les paramètres géométriques des inhibiteurs d'interfaces sont aussi différents[61], la partie hydrophobe des ligands exposant au solvant est plus petite que sur les ligands classiques et cette zone est beaucoup plus localisé sur les extrémités. Les ligands des interfaces protéine-protéine sont aussi plus globulaires que les ligands classiques.

Il est à noter que durant ces dernières années, beaucoup d'avancées majeures dans la création d'inhibiteurs d'interfaces ont eu pour source des peptides et peptidomimétiques[62–64].

Tableau 1.4: *Tableau récapitulatif des inhibiteurs d'interface protéine-protéine passés en phase clinique (issu de Scott et al., Nature Rev. Drug. Discov., 2016)*

Molécule	PPI	Indication	Phase clinique	Structure
Venetoclax	Bcl2-Bak	Cancer	Pré-enregistrement	
Idasanutlin	Mdm2-p53	Cancer	Phase III	
AMG232	Mdm2-p53	Cancer	Phase I/II	
CGM097	Mdm2-p53	Cancer	Phase I	
DS-3032b	Mdm2-p53	Cancer	Phase I	
ALRN-6924	Mdm2-p53	Cancer	Phase I	
MK-8242	Mdm2-p53	Cancer	Phase I	
Lifitegrast	Lfa1-Icam1	Syndrome de l'oeil sec	Pré-enregistrement	
Tirofiban	GpIIb/IIIa-Fibronogène	Thrombose	Approuvé	
Carotegrast methyl	Intégrine α4-	Colite ulcéreuse	Phase III	
AT-406	Xiap-Smac	Cancer	Phase II	

Molécule	PPI	Indication	Phase clinique	Structure
LCL-161	Xiap-Smac	Cancer	Phase II	
Birinapant	Xiap-Smac	Cancer	Phase I/II	
ASTX-660	Xiap-Ciap1	Cancer	Phase I/II	
AEG40826	Xiap-Smac	Cancer	Phase I	
CUDC-427	Xiap-Smac	Cancer	Phase I	
Apabetalone	Bet-Histone	Diabète type 2	Phase III	
GSK525762	Bet-Histone	Cancer	Phase I/II	
CPI-0610	Bet-Histone	Cancer	Phase I	
TEN-010	Bet-Histone	Cancer	Phase I	
OTX015	Bet-Histone	Cancer	Phase I	

1.5.3.Conception de médicaments

La conception de médicaments (Drug Design) fait appel à la recherche de molécules affines pour une cible d'intérêt qui peut être une protéine, un complexe protéique ou une voie métabolique. Divers algorithmes et méthodes existent depuis longtemps et les progrès informatiques incessants permettent d'accélérer toujours plus le processus. La conception de médicaments assistée par ordinateur permet d'accélérer cette phase de criblage en évitant notamment de tester expérimentalement des milliers de composés. Le but est de réaliser un premier tri de composé à l'aide de tests *in-silico,* dont les trois principaux sont l'arrimage moléculaire, la recherche de pharmacophores et l'étude des modes d'interactions.

L'arrimage moléculaire

Cette méthode a pour principe d'essayer de positionner une molécule de faible poids moléculaire au sein d'une protéine[65]. L'arrimage moléculaire commence avec le positionnement d'une protéine dans un champ de force. Chacune des conformations du ligand, soit pré-enregistrée soit calculée à la volée, est positionnée à la surface de la protéine et subit de nombreuses translations et rotations. Chaque position du complexe est conservée et se voit attribuer un score grâce à une fonction de score définie. Cette fonction est en général basée sur la complémentarité entre les deux entités et prend en compte en priorité les liaisons hydrogènes et les contacts hydrophobes[66]. Au final seule la pose ayant le meilleur score est conservée pour chaque molécule. La chimiothèque criblée est ensuite triée par score de docking décroissant afin de prioriser les molécules les plus prometteuses (touches) à la validation expérimentale in vitro.

Cette technique est une des plus intuitives pour prédire des complexes protéine/ligand[67] cependant elle possède de nombreux inconvénients. La notation des interactions est problématique pour prédire l'affinité mais fonctionne bien pour positionner les ligands. L'applicabilité de la méthode aux interfaces protéine-protéine a récemment été vérifiée avec les logiciels AutoDock[68] et Glide[69] dans le but de reproduire les poses de d'inhibiteurs de PPI dont la position cristallographique est connue[70]. Les résultats obtenus (54% de bonne prédiction sur les 80 inhibiteurs d'interfaces testés) sont en fait très proches de ceux obtenus avec des ligands de protéine globulaires. L'arrimage moléculaire a notamment été utilisé ces dernières années pour identifier des inhibiteurs de PPI par criblage virtuel (**Figure 1.10**)

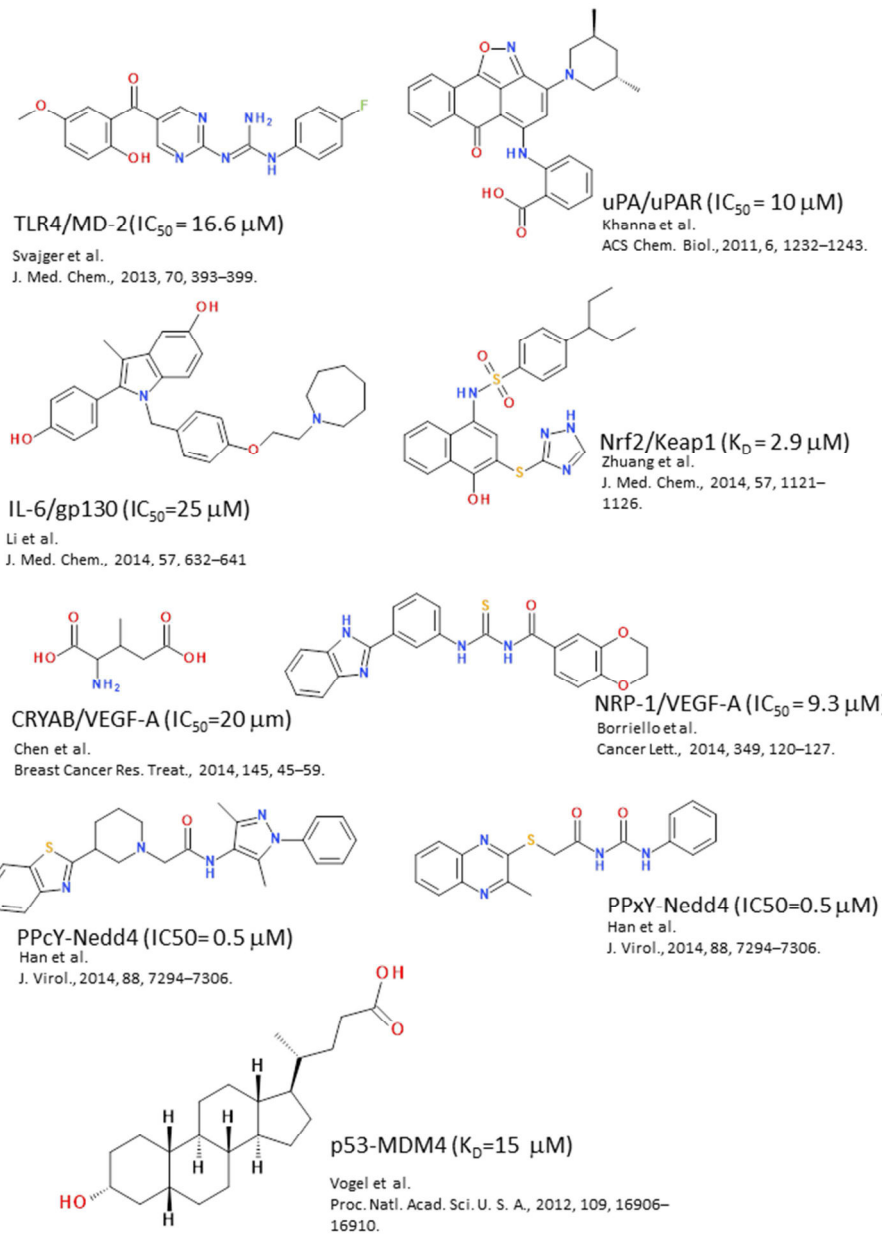


Figure 1.10: *Inhibiteurs d'interfaces protéine-protéine découverts par arrimage moléculaire. Pour chaque inhibiteur, l'interface et sa constante de liaison sont précisées.*

La recherche de pharmacophores

Le pharmacophore est la représentation d'un « ensemble de propriétés stériques et électroniques défini à partir d'une interaction entre deux entités moléculaires et nécessaire pour induire une réponse biologique » (IUPAC 1998). Il se présente sous la forme d'un ensemble de sphères typées et orientées représentant les interactions entre les deux entités, c'est une version simpliste et plus souple du ligand dans la position d'un complexe donnée[71,72]. Il existe aussi des pharmacophores basés uniquement sur les ligands qui eux

permettent de reproduire une molécule en trois dimension. Les propriétés étant représentées par des sphères, il y a une tolérance pour l'alignement d'autres molécules. L'adaptation d'une molécule au pharmacophore est approximée par un score de fitness témoignant de la superposition des atomes du ligand aux éléments de même nature physicochimique du pharmacophore. Comme pour l'arrimage moléculaire, une chimiothèque peut être criblée in silico afin de sélectionner les touches vérifiant le mieux le pharmacophore de référence.

Cette méthode s'adapte très bien aux interfaces protéine-protéine en utilisant une des protéines partenaires en tant que ligand pour créer le pharmacophore, nous obtenons un pharmacophore de l'interface. La taille de l'interface protéine-protéine est très importante ici, le pharmacophore doit être souvent vérifié et sélectionné à la main afin d'obtenir des inhibiteurs d'affinité suffisante, comme ceux récemment identifiés par recherche pharmacophoriques à partir de larges chimiothèques (**Figure 1.11**).

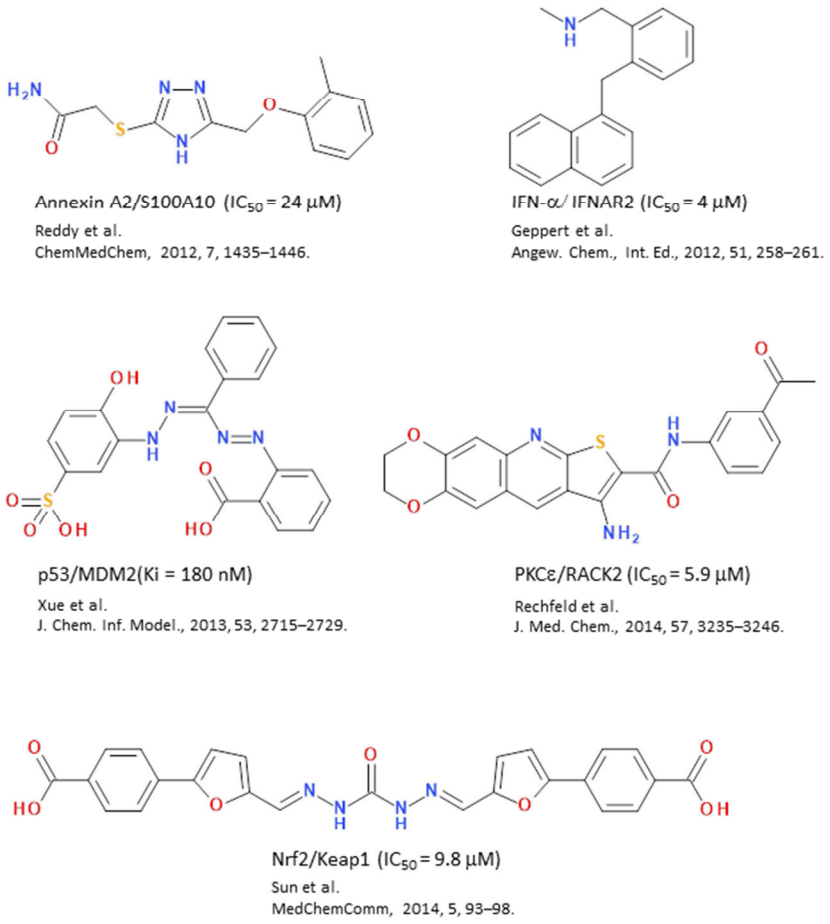


Figure 1.11: *Inhibiteurs d'interfaces protéine-protéine découverts par une recherche de pharmacophore. Pour chaque inhibiteur, l'interface et sa constante de liaison sont précisées.*

1.5.4. Bases de données d'inhibiteurs d'interfaces protéine-protéine

Initialement limité à des séries chimiques similaires ciblant un très petit nombre d'interfaces (ex: p53-MDM2, BclXl-Bak, IL2-IL2R), le répertoire d'inhibiteurs connus d'interfaces protéine-protéine ne cesse de croître et accentue notre connaissance encore parcellaire de l'espace chimique qui lui est associé[61]. Trois bases de données stockent l'essentiel de l'information disponible sur les interfaces et leurs ligands.

2P2I

2P2Idb[73,74] est une base de données structurale (http://2p2idb.cnrs-mrs.fr/) dédiée aux interactions protéine-protéine pour lesquelles des modulateurs sous forme de petite molécules sont connus (**Figure 1.12**).

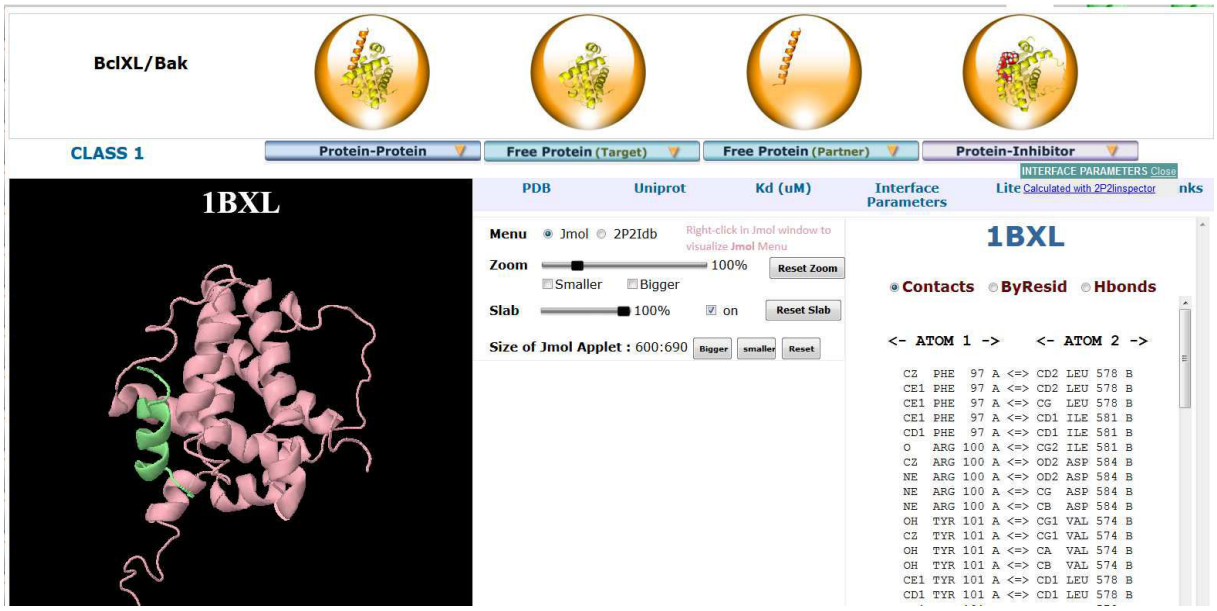


Figure 1.12: *Capture d'écran de la base de données 2P2I illustrant les données structurales connues sur le complexe BClXL-Bak.*

Elle compile manuellement les informations structurales relatives aux inhibiteurs orthostériques, à leur cible (structure issues de la PDB) et fournit des liens vers d'autres bases de données utiles. 2P2Idb comprend toutes les interactions pour lequel à la fois le complexe protéine-protéine, les protéines dissociées et des complexes protéine-inhibiteur ont été structurellement caractérisés. Depuis sa première version en 2010, la base de données n'a cessé de croître et la version actuelle contient 27 complexes protéine-protéine et 274 complexes protéine-inhibiteur correspondant à 242 petites molécules inhibitrices uniques. Des

outils cheminformatiques compagnons (2P2I inspector, 2P2I score, 2P2I hunter) ont été développés autour de cette base de données afin de caractériser ces interfaces par propriétés physicochimiques[74], prédire leur droguabilité structurale[73] et concevoir des chimiothèques d'inhibiteurs d'interfaces[75].

iPPI-DB

La base de données iPPI-DB[76] (http://www.ippidb.cdithem.fr) recense 1756 inhibiteurs d'interfaces protéine-protéine (18 familles différentes) décrits dans la littérature selon 4 critères distincts: (i) connaissance de la fonction de l'interface, (ii) ligands non-peptidiques, (iii) activité in vitro définie (IC_{50}, EC_{50}, K_d, K_i), (iv) activité in vitro en dessous d'un seuil de concentration de 30 µM (**Figure 1.13**)

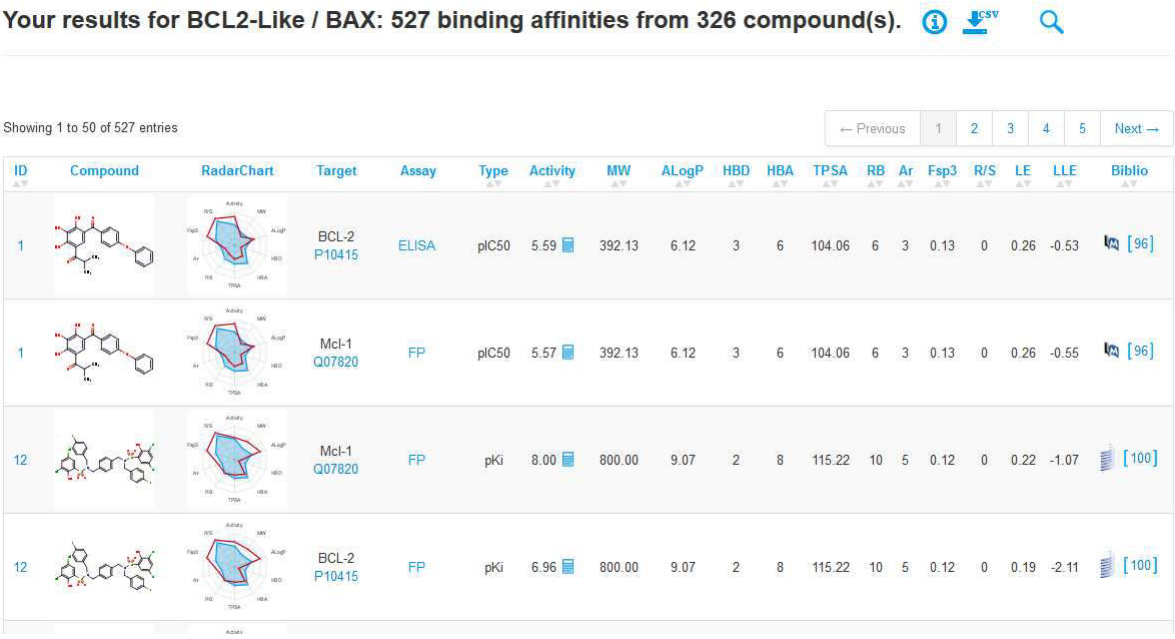


Figure 1.13: *Capture d'écran de la base de données 2P2I illustrant les données structurales connues sur la famille de complexes impliquant la protéine BCl2 et ses homologues*

Pour chaque inhibiteur, outre son activité in vitro et le type d'essai utilisé, sont répertoriées diverses propriétés physicochimiques calculées (divers comptes moléculaires et indices) ainsi qu'une comparaison graphique (plot en étoile) des propriétés du ligand vis-à-vis de la totalité des inhibiteurs répertoriés.

TIMBAL

Timbal[77] est une base de données (http://mordred.bioc.cam.ac.uk/timbal) recensant de manière automatisée 8889 inhibiteurs et stabilisateurs d'interfaces protéine-protéine, à partir de requêtes dans la base de données ChEMBL[78] suivant une liste pré-établie de 50 cibles et un indice de confiance sur le type d'essai réalisé (**Figure 1.14**). Il est toutefois à noter qu'une grande partie des données présentées (environ 50%) concerne une seule famille d'interface (intégrines) pas toujours clairement identifiée au niveau moléculaire.

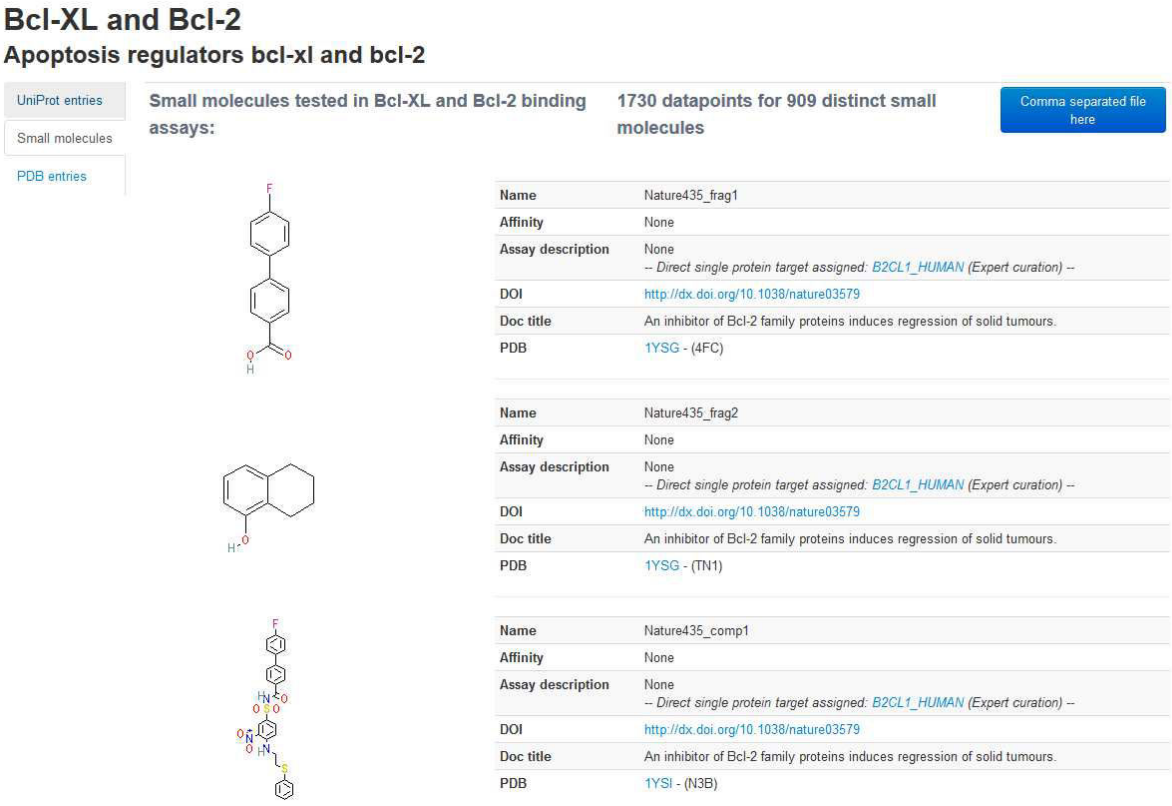


Figure 1.14: *Capture d'écran de la base de données TIMBAL illustrant les données structurales connues sur la famille de complexes impliquant la protéine BCl2 et ses homologues*

1.6. Conclusion

La résolution des structures de protéines et notamment par diffraction des rayons X a révolutionné l'étude des interactions protéine-protéine. L'afflux et la complexité grandissante des données à amener son lot de problèmes. Les structures présentes dans les cristaux ne sont pas toujours similaires à celles effectivement mises en jeu in vivo. La résolution partielle de ce problème passe par la création de jeux de données permettant la validation d'outils de distinction entre les interfaces biologiquement pertinentes et les contacts cristallins. La qualité des jeux a toujours été variable et a évolué avec la qualité des structures résolues. Les outils basés sur des méthodes d'apprentissage très rapides méritent une attention particulière et pourraient être encore améliorés.

Aujourd'hui de nombreux classifieurs d'interaction protéine-protéine existent mais peu d'outils présentent des informations complémentaires pour l'analyse des interfaces protéine-protéine. Leur étude est devenue une discipline à part entière et apporte de grands espoirs pour l'avenir de la pharmacologie. Les méthodologies sont variées et commencent à montrer des succès cliniques. Chaque interface reste néanmoins un cas à part et nécessite une actualisation de la stratégie la plus adaptée afin d'identifier des petites molécules inhibitrices. Malgré tout, les interfaces protéine-protéine représentent un potentiel encore très faiblement exploité, notamment de par un manque d'informations structurales et fonctionnelles à l'échelle protéomique. Le nombre grandissant de structures cristallographiques pertinentes et de données expérimentales contribuera dans un futur très proche au développement d'outils bio et cheminformatiques de qualité permettant l'élargissement considérable de l'espace chimique associé aux petites molécules inhibitrices et à de meilleures prédictions à la fois qualitatives et quantitatives.

1.7. Bibliographie

1. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66,** 334–395 (2014).
2. Andreani, J. & Guerois, R. Evolution of protein interactions: From interactomes to interfaces. *Arch. Biochem. Biophys.* **554,** 65–75 (2014).
3. Stumpf, M. P. H. et al. Estimating the size of the human interactome. *Proc. Natl. Acad. Sci.* **105,** 6959–6964 (2008).
4. Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.* **3,** 301–317 (2004).
5. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. **28,** 235–242 (2000).
6. Vickery, H. B. The Origin of the Word Protein. Yale *J. Biol. Med.* **22,** 387–393 (1950).
7. Lee, M. H. et al. Purification and characterization of Klebsiella aerogenes UreE protein: a nickel-binding protein that functions in urease metallocenter assembly. *Protein Sci. Publ. Protein Soc.* **2,** 1042–1052 (1993).
8. Zheng, H. et al. X-ray crystallography over the past decade for novel drug discovery – where are we heading next? *Expert Opin. Drug Discov.* **10,** 975–989 (2015).
9. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181,** 223–230 (1973).
10. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42,** D304–D309 (2014).
11. Sillitoe, I., Dawson, N., Thornton, J. & Orengo, C. The history of the CATH structural classification of protein domains. *Biochimie* **119,** 209–217 (2015).
12. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
13. Nooren, I. M. A. & Thornton, J. M. NEW EMBO MEMBER'S REVIEW. E*MBO J.* **22,** 3486–3492 (2003).
14. Grünberg, R., Leckner, J. & Nilges, M. Complementarity of structure ensembles in protein-protein binding. *Struct. Lond. Engl.* 1993 **12,** 2125–2136 (2004).
15. Bosshard, H. R. Molecular Recognition by Induced Fit: How Fit is the Concept? *Physiology* **16,** 171–173 (2001).

16. Camacho, C. J., Kimura, S. R., DeLisi, C. & Vajda, S. Kinetics of desolvation-mediated protein-protein binding. *Biophys. J.* **78,** 1094–1105 (2000).
17. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).
18. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254,** 1598–1603 (1991).
19. Zhang, L. & Buck, M. Molecular Simulations of a Dynamic Protein Complex: Role of Salt-Bridges and Polar Interactions in Configurational Transitions. *Biophys. J.* **105,** 2412–2417 (2013).
20. Holding, A. N. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* **89,** 54–63 (2015).
21. Schuck, P. Analytical Ultracentrifugation as a Tool for Studying Protein Interactions. *Biophys. Rev.* **5,** 159-171 (2013).
22. Fekete, S., Beck, A., Veuthey, J.-L. & Guillarme, D. Theory and practice of size exclusion chromatography for the analysis of protein aggregates. *J. Pharm. Biomed. Anal.* **101,** 161–173 (2014).
23. Hong, P., Koza, S. & Bouvier, E. S. P. Size-Exclusion Chromatography for the Analysis of Protein Biotherapeutics and their Aggregates. *J. Liq. Chromatogr. Relat. Technol.* **35,** 2923–2950 (2012).
24. Folta-Stogniew, E. Oligomeric states of proteins determined by size-exclusion chromatography coupled with light scattering, absorbance, and refractive index detectors. *Methods Mol. Biol.* **328,** 97–112 (2006).
25. Petoukhov, M. V. & Svergun, D. I. Applications of small-angle X-ray scattering to biomacromolecular solutions. *Int. J. Biochem. Cell Biol.* **45,** 429–437 (2013).
26. Lorber, B., Fischer, F., Bailly, M., Roy, H. & Kern, D. Protein analysis by dynamic light scattering: methods and techniques for students. *Biochem. Mol.* **40,** 372–382 (2012).
27. Böttcher, B. & Hipp, K. Single-particle applications at intermediate resolution. *Adv. Protein Chem. Struct. Biol.* **81,** 61–88 (2010).
28. Kiselar, J. G. & Chance, M. R. Future directions of structural mass spectrometry using hydroxyl radical footprinting. *J. Mass Spectrom.* **45,** 1373–1382 (2010).
29. Soares da Costa, T. P. et al. Quaternary Structure Analyses of an Essential Oligomeric Enzyme. *Methods Enzymol.* **562,** 205–223 (2015).

30. Rose, P. W. et al. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **39,** D392–D401 (2011).
31. Levy, E. D. PiQSi: Protein Quaternary Structure Investigation. *Structure* **15,** 1364–1367 (2007).
32. Krissinel, E. Crystal contacts as nature's docking solutions. *J. Comput. Chem.* **31,** 133–143 (2010).
33. Stehr, H., Duarte, J. M., Lappe, M., Bhak, J. & Bolser, D. M. PDBWiki: added value through community annotation of the Protein Data Bank. Database *J. Biol. Databases Curation* **2010,** (2010).
34. Janin, J. Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.* **4,** 973–974 (1997).
35. Henrick, K. & Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23,** 358–361 (1998).
36. Ponstingl, H., Kabir, T. & Thornton, J. M. Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr.* **36,** 1116–1122 (2003).
37. Krissinel, E. Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res.* **43,** W314–319 (2015).
38. Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* **7,** 27 (2006).
39. Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J. & Poupon, A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics* **24,** 652–658 (2008).
40. Tsuchiya, Y., Nakamura, H. & Kinoshita, K. Discrimination between biological interfaces and crystal-packing contacts. *Adv. Appl. Bioinform. Chem.* **1,** 99–113 (2008).
41. Schärer, M. A., Grütter, M. G. & Capitani, G. CRK: An evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins Struct. Funct. Bioinform.* **78,** 2707–2713 (2010).
42. Mitra, P. & Pal, D. Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Struct*ure. 1993 **19,** 304–312 (2011).
43. Duarte, J. M., Srebniak, A., Schärer, M. A. & Capitani, G. Protein interface classification by evolutionary analysis. *BMC Bioinformatics* **13,** 334 (2012).

44. Sudarshan, S., Kodathala, S. B., Mahadik, A. C., Mehta, I. & Beck, B. W. Protein-protein interface detection using the energy centrality relationship (ECR) characteristic of proteins. *PloS One* **9,** e97115 (2014).
45. Luo, J. et al. Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins* **82,** 3090–3100 (2014).
46. Liu, Q., Li, Z. & Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics* **15 Suppl 16,** S3 (2014).
47. Xu, Q. & Dunbrack, R. L. The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.* **39,** D761–770 (2011).
48. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372,** 774–797 (2007).
49. Lo Conte, L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285,** 2177–2198 (1999).
50. Chakrabarti, P. & Janin, J. Dissecting protein-protein recognition sites. *Proteins Struct. Funct. Genet.* **47,** 334–343 (2002).
51. Levy, E. D. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* **403,** 660–670 (2010).
52. Ponstingl, H., Henrick, K. & Thornton, J. M. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41,** 47–57 (2000).
53. Prasad Bahadur, R., Chakrabarti, P., Rodier, F. & Janin, J. A Dissection of Specific and Non-specific Protein–Protein Interfaces. *J. Mol. Biol.* **336,** 943–955 (2004).
54. Da Silva, F., Desaphy, J., Bret, G. & Rognan, D. IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces. *J. Chem. Inf. Model.* **55,** 2005–2014 (2015).
55. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934 (2001).
56. Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S. & Capitani, G. A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Struct. Biol.* **14,** 22 (2014).
57. Hwang, H., Vreven, T., Janin, J. & Weng, Z. Protein-Protein Docking Benchmark Version 4.0. *Proteins* **78,** 3111–3114 (2010).

58. Ryffel, B., Woerly, G., Hofmann, G., Arnold, J., Greiner, B. & Foxwell B.M. Discovery of novel immunosuppressants in vitro. *ALTEX* **8,** 31–39 (1991).
59. Arkin, M. R., Tang, Y. & Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. *Chem. Biol.* **21,** 1102–1114 (2014).
60. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267,** 383–386 (1995).
61. Kuenemann, M. A., Bourbon, L. M. L., Labbé, C. M., Villoutreix, B. O. & Sperandio, O. Which Three-Dimensional Characteristics Make Efficient Inhibitors of Protein–Protein Interactions? *J. Chem. Inf. Model.* **54,** 3067–3079 (2014).
62. Azzarito, V., Long, K., Murphy, N. S. & Wilson, A. J. Inhibition of α-helix-mediated protein-protein interactions using designed molecules. *Nat. Chem.* **5,** 161–173 (2013).
63. Bernal, F., Wade, M., Godes, M., Davis, T.N., Whitehead, D.G., Kung, A.L., Wahl, G.M. & Walensky, L.D. A stapled p53 helix overcomes HDMX-mediated suppression of p53. *Cancer Cell* **18,** 411–422 (2010).
64. Gavenonis, J., Sheneman, B. A., Siegert, T. R., Eshelman, M. R. & Kritzer, J. A. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.* **10,** 716–722 (2014).
65. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des*. **16,** 151–166 (2002).
66. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D. & Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **60,** 325–332 (2005).
67. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153,** S7–S26 (2008).
68. Morris, G. M. et al. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).
69. Friesner, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).
70. Krüger, D. M., Jessen, G. & Gohlke, H. How good are state-of-the-art docking tools in predicting ligand binding modes in protein-protein interfaces? *J. Chem. Inf. Model.* **52,** 2807–2811 (2012).
71. Leach, A. R., Gillet, V. J., Lewis, R. A. & Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **53,** 539–558 (2010).

72. Güner, O. F. & Bowen, J. P. Setting the record straight: the origin of the pharmacophore concept. *J. Chem. Inf. Model.* **54,** 1269–1283 (2014).
73. Basse, M. J. et al. 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res.* **41,** D824–827 (2013).
74. Basse, M.-J., Betzi, S., Morelli, X. & Roche, P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database J. Biol. Databases Curation* **2016,** (2016).
75. Hamon, V. et al. 2P2I HUNTER: a tool for filtering orthosteric protein-protein interaction modulators via a dedicated support vector machine. *J. R. Soc. Interface.* **11,** 20130860 (2014).
76. Labbé, C. M., Laconde, G., Kuenemann, M. A., Villoutreix, B. O. & Sperandio, O. iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. *Drug Discov. Today* **18,** 958–968 (2013).
77. Higueruelo, A. P., Jubb, H. & Blundell, T. L. TIMBAL v2: update of a database holding small molecules modulating protein–protein interactions. *Database J. Biol. Databases Curation* (2013). doi:10.1093/database/bat039
78. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40,** D1100–1107 (2012).

Chapitre 2

Détermination de la pertinence biologique d'une interface protéine-protéine

2.1. Mise en contexte

La forme et la composition des interfaces protéine-protéine (PPI) sont caractéristiques de la pertinence biologique de leurs structures. Les PPIs sont essentielles au monde du vivant notamment pour le transfert d'information et le maintien structural au sein des cellules. Analyser des interfaces et émettre des hypothèses basées sur des structures qui ne représentent pas l'état biologique des protéines est dangereux. Les outils permettant d'établir cette véracité biologique n'étaient pas applicables aux données nous intéressant. De plus, le phénomène d'interactions entre deux protéines est encore méconnu et il est important de définir si l'interface est biologiquement viable, de la détecter et de la décrire. La caractérisation des interfaces peut nous en apprendre plus sur l'oligomérisation et la dissociation des complexes protéiques.

Les premiers exemples de régulation de PPIs par des inhibiteurs de faibles poids moléculaires montrent que de nombreuses interfaces, même de petite taille, peuvent être biologiquement pertinentes, et donc représenter des cibles privilégiées à des modulateurs de faible poids moléculaire.

Dans ce chapitre, nous allons aborder la pertinence biologique des structures de PPIs. Lors de nos premières recherches, nous nous sommes aperçus qu'il était usuel de trouver des structures que l'on peut définir comme étranges. Il est courant, par exemple, de trouver des molécules de détergents ou d'agents précipitants à l'interface. De même, certaines structures de PPIs ont pu être déterminées dans des conditions expérimentales non physiologiques (ex: haute pression). Nous nous sommes posés la question de savoir comment mettre ces structures de côté, car il n'existait que très peu de méthodes de détermination de la pertinence biologique de PPIs dont le domaine d'applicabilité couvrait l'ensemble de la PDB.

Les classifieurs les plus utilisés sont performants mais ils ont été entraînés sur des jeux de données très anciens et de mauvaise qualité. Ce sont des calculs longs souvent basés sur des calculs énergétiques fortement influencés par la taille de l'interface. Or, nous voulions précisément être en mesure de prédire à la fois des interfaces biologiques de petite taille (< 1000 $Å^2$) et des contacts cristallins de grande taille (> 2000 $Å^2$). Les interactions biologiques de petites tailles concernent souvent des hélices-α qui se lient à une cavité du partenaire. Ces

structures, incomplètes la plupart du temps, sont systématiquement rejetées par les classifieurs connus, car la taille de la zone d'interaction est trop faible.

Afin de corriger ce problème, nous avons mis au point un jeu d'apprentissage complet basé sur les jeux existants auxquels nous avons rajouté des PPIs biologiquement pertinentes de petite taille et des PPIs non pertinentes de grande taille. Nous avons aussi développé une méthode de détection et de caractérisation des interfaces protéine-protéine rapide (IChemPIC), basée sur des descripteurs topologiques des interactions moléculaires mises en jeu.

Ce chapitre a fait l'objet d'une publication qui décrit le fonctionnement d'IChemPIC, notre classifieur d'interfaces protéine-protéines.

IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces

Franck Da Silva[1], Jérémy Desaphy[1#], Guillaume Bret[1] and Didier Rognan[1*]

[1] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France.

*To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

Current address: Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, United States

Publié dans Journal of Chemical Information and Modelling, 2015, 55, 2005-2014

2.2. Introduction

Les interactions protéine-protéine (PPI) sont au cœur de la plupart des situations pathologiques au sein des cellules vivantes, et attirent donc de plus en plus la recherche pharmaceutique.[1-3] Parmi les nombreuses stratégies pour identifier les modulateurs d'interface de faible poids moléculaire, l'approche fondée sur la structure rationnelle a historiquement joué un rôle important, notamment en raison de l'intégration possible du criblage biophysique des banques de fragments (résonance plasmonique de surface, calorimétrie de titration isotherme, spectroscopie par résonance magnétique nucléaire, spectrométrie de masse) en déterminant la structure cristallographique.[4] Pour exploiter pleinement les connaissances structurales des cibles droguables d'intérêt, il est souhaitable de connaître leur véritable état d'oligomérisation ainsi que leur pertinence biologique. Tout au long de ce manuscrit, nous allons considérer comme complexe protéine-protéine «biologique» toute entrée possédant une vraie pertinence biologique ainsi qu'une fonction physiologique (par exemple l'adhésion cellulaire, la signalisation cellulaire, la reconnaissance immunitaire, la transcription). Les complexes homo ou hétéro-oligomériques résultant soit de la formation du cristal ou dépourvus de toute fonction biologique connue seront considérés comme «cristallins» où non-pertinents. Malheureusement, déduire la structure quaternaire et la pertinence biologique à partir des coordonnées atomiques dans la Protein Data Bank (PDB) [5] n'est pas simple. Par exemple, le contenu de l'unité asymétrique (ASU) déposée dans la PDB (la fraction de la cellule cristallographique unitaire qui n'a pas de symétrie cristallographique) peut décrire une ou plusieurs copies d'une macromolécule, mais sans indication particulière sur quel état d'oligomérisation (par exemple, monomère, dimère) est le plus pertinent. De même, l'ASU peut avoir besoin d'opérations de symétrie cristallographique à appliquer avant de reconstituer l'assemblage macromoléculaire biologiquement pertinent (unité biologique). Des procédures automatisées, de discrimination de structures 3D, de différenciation des contacts cristallins et d'interfaces biologiquement pertinentes stables sont donc nécessaires pour éviter des expériences biochimiques longues et coûteuses telles que la filtration sur gel, diffusion de la lumière ou la sédimentation à l'équilibre

En règle générale, l'aire des interfaces cristallographiques est beaucoup plus faible (<1000 Å²) que celles des interfaces biologiquement pertinentes[6]. Toutefois, cette règle simple souffre de nombreuses exceptions, car certaines interfaces très importantes comme celles impliquant des sites de reconnaissance d'hélice α peuvent être de très petite taille (par

exemple 780 Å² pour le complexe p53-mdm2). De nombreuses méthodes de classification ont donc été conçues pour prédire directement l'état d'oligomérisation des complexes protéiques à partir des simples coordonnées atomiques. [7] La toute première approche, reportée en 1998 est PQS (protein quaternary structure file server) [8] utilise une fonction de score empirique basée sur des descripteurs (aire de l'interface ; le nombre de résidus d'interfaces enfouis, les ponts salins et liaison disulfure ; l'énergie de solvatation de la structure quaternaire). Bien qu'imparfaite (au moins 20% des erreurs de classification ont été signalés par les auteurs eux-mêmes), le serveur PQS a ouvert la voie à de nombreuses méthodes qui peuvent être regroupées en deux catégories.

Un premier type d'approches, dont PISA[9] est le principal représentant, repose sur les premiers principes de la physique pour prédire la stabilité des assemblages de protéines en solution. Par exemple, PISA calcule explicitement l'énergie libre de dissociation de Gibbs afin de prédire la pertinence biologique d'un assemblage macromoléculaire. Appliqué à un ensemble de données de 218 structures PDB, cette méthode a atteint un taux de réussite remarquable de 90% dans la prédiction de véritables interfaces biologiques[9]. PISA peut être considéré comme une méthode de référence, et est actuellement utilisé pour prédire les structures quaternaires de chaque entrée du site RCSB PDB. Un deuxième groupe de méthodes[7, 10-17], génère des modèles de régression/classification linéaire ou non-linéaire sur un jeu d'entrainement prédéfini (contacts cristallins et interfaces pertinentes) afin de prédire la structure quaternaire de jeux externes. Plusieurs descripteurs géométriques et descripteurs de complémentarité chimique peuvent être utilisés pour discriminer, avec une précision similaire (ca. 85-90%), les contacts cristallins des interfaces biologiquement pertinentes. Très souvent, ces méthodes (e.g. IPAC,[7] DiMoVo,[12] or NOXClass[13]) utilisent des machines d'apprentissages (séparateurs à vaste marge, arbres de décision, inférence Bayésienne) sur les vecteurs de contacts atomiques ou les vecteurs inter-résidus pour décider quel jeu de paramètres est le plus adéquat pour une classification optimale. La conservation des résidus composant le cœur de l'interface[18, 19] peut être ajoutée aux descripteurs précédemment cités, comme par exemple dans EPPIC,[20] pour mettre en avant l'importance des résidus du cœur de l'interface au sein des interfaces biologiquement pertinentes.

Dans le but de pouvoir être comparé aux autres, beaucoup d'études ont été réalisées sur un nombre limité de jeux d'analyse comparative[10, 21, 22] qui sont pourtant connus pour être biaisés par la taille des entrées les composant, des petits contacts cristallins opposés à de grandes interfaces biologiquement pertinentes. [7, 12, 20] En conséquence, la plus part des classifieurs ont

une précision beaucoup plus faible quand on les applique à un jeu de données dont les tailles des interfaces sont équilibrées. Un jeu de données récent[20] a été créé en faisant attention de sélectionner des interfaces pertinentes et des contacts cristallins ayant une aire comparable et très stricte (autour de 1000Å²). Notre but final étant d'identifier les interfaces de petites tailles potentiellement droguables[23], aucun des jeux de données existant ne nous apparaissait satisfaisant. Nous avons donc conçu manuellement un jeu de données (FDS set) de 200 complexes protéine-protéine biologiquement pertinents non redondants dont les structures cristallographiques sont connues, qui a été complété par un nombre équivalent de 200 interfaces cristallines filtrées pour couvrir une aire d'interface comparable. Nous avons ensuite utilisé un algorithme d'apprentissage automatique (Random Forest) à l'aide de 45 descripteurs d'interactions moléculaires, pour former un modèle qui, lorsqu'il est appliqué à plusieurs ensembles de test externes, permet d'obtenir une bonne précision et une robustesse stable dans la distinction entre contacts cristallins et interfaces biologiquement pertinentes, quelles que soient leurs aires d'interfaces.

2.3. Méthodes informatiques

2.3.1. Jeux de données

FDS dataset.

Les contacts cristallins ont été récupérés à partir de deux ensembles de données précédemment rapportés.[20, 21] Tout d'abord, 141 protéines monomèriques connus à partir du jeu de données Bahadur avec une zone d'interface cristalline dans l'intervalle 400-1200 Å², ont été récupérées de la manière suivante. Les coordonnées atomiques de l'unité asymétrique ont été récupérées à partir de la « RCSB Protein Data Bank », pour chacune, une cellule unitaire est reconstruite à l'aide AmberTools14[24]. Pour chaque structure et toutes les paires de chaînes possibles, l'aire d'interface IA (Eqn1) a été mesurée à l'aide de MSMS[25] après le retrait des atomes n'appartenant pas aux protéines (solvant, des ligands, des ions) et en utilisant un rayon de sonde et une densité de sommet de 1,4 Å et 2,0 / Å², respectivement.

$$IA_{A,B} = \frac{(ASA_A + ASA_B) - ASA_{AB}}{2} \qquad \textbf{(Eqn. 1)}$$

Où $IA_{A,B}$ est l'aire d'interface entre les chaines A et B, ASA_A est la surface accessible au solvant de la chaine A isolée, ASA_B est la surface accessible au solvant de la chaine B et ASA_{AB} est la surface accessible au solvant du complexe AB.
L'interface possédant la plus grande aire est conservée pour chaque entrée PDB. Le jeu de Bahadur est ensuite complété par 82 interfaces recrées sur des protéines connues pour être monomériques en solution provenant du jeu de données DCXtal[20] et sélectionnées sur la base de leurs aires d'interface, de manière à ce qu'elle soit comprise dans la gamme : 1000-1500Å². Les structures PDB correspondantes ont été directement récupérées sur le site EPPIC website (http://www.eppic-web.org/ewui/#downloads). La redondance entre les protéines a été supprimée en ne conservant que les protéines ayant une identité de séquence inférieure à 70%, en se basant sur les règles de redondance de la RCSB[26]. Le jeu final composé de 200 contacts cristallins non-redondant (PDB id, nom de la protéine, chaines, aire d'interfaces, résolution, classification) est donné dans le **Tableau supplémentaire 2.1**.

Un ensemble de 200 interfaces biologiquement pertinentes et non redondantes (<70% d'identité de séquences paire par paire entre deux chaines) a été créé à la main depuis la littérature en accord avec les sources suivantes : (i) le jeu récent DCbio[20] d'interfaces biologiques pertinentes d'homodimères (74 PPIs) ; (ii) la base de données 2P2I[23] archivant les hétérodimères pour lesquels est connue la structure cristallographique, la structure de chaque monomère en forme libre et au moins un des partenaire co-cristallisé avec un inhibiteur de faible poids moléculaire (10 PPIs) ;(iii) des inhibiteurs d'interfaces de faible poids moléculaires[27] connus dont la structure du complexe des interfaces est connue (5 PPIs) ; (iv) quelques structures connues d'interface protéine-protéine autour du cancer[1] (8 PPIs); (v) la base de données "PPIAffinity[28]" contenant les structures d'interface protéine-protéine pertinente (complexe et forme libre de chaque partenaire) ainsi que des données expérimentales comme la constante de liaison (54 PPis); (vi) la base de données appelée « hot loops[29]» (20 PPIs) ; (vii) des interfaces biologiquement pertinentes d'hétéromères (18 PPIs). La pertinence des 200 complexes a été vérifié à la main au vu de la litérature[20, 23, 27-29]

Les structures correspondantes ont été téléchargées depuis la PDB. Les chaines participant à l'interface ont été sélectionnées manuellement en se référant aux sources précédemment citées. La surface enfouie de l'interface a été calculé sur les structures sans les atomes n'appartenant pas à la protéine, en suivant l'Eqn 1 comme précédemment. Le jeu

complet des 200 interfaces biologiquement pertinentes (PDB id, nom de la protéine, chaines, aire d'interfaces, résolution, classification) est donné dans le **Tableau supplémentaire 2.2**.

Les 400 interfaces décrites ci-dessus (cristallographiques, biologiques) ont été réparties au hasard en deux groupes (75% dans le jeu d'apprentissage, 25% dans le jeu de test) conservant une proportion égale d'interfaces cristallographiques et biologiques dans chaque sous-ensemble. Une attention particulière a également été donnée pour déterminer une distribution équivalente des aires de la zone d'interface dans les deux ensembles. En suivant la procédure décrite ci-dessus, réaliser plusieurs répartitions aléatoires (75/25) n'a pas influencé les résultats obtenus (meilleurs paramètres RF, F-mesure des meilleurs modèles RF sur la validation et des ensembles de tests externes, données non présentées). Le jeu d'apprentissage ainsi que celui de test sont donnés dans les **Tableaux supplémentaires 2.1 et 2.2.**
Les entrées des jeux de données IPAC[7], Ponstigl[10] and Bahadur[21] ont été extraites de la PDB selon les identifiants PDB et les noms de chaîne décrits dans la littérature.

Coordonnées atomiques. Pour chaque entrée PDB, les hydrogènes sont ajoutés avec Protoss,[30] une méthode récemment décrite pour le placement des hydrogènes dans des complexes protéine-ligand, qui tient compte des états tautomériques et de protonation des tous les éléments présents. La méthode génère la position la plus probable des hydrogènes et permet de recréer le plus grand réseau de liaison grâce à une fonction de score empirique. Les structures tout atome des entrées du jeu de données FDS peuvent être téléchargées sur : http://bioinfo-pharma.u-strasbg.fr/IChemPIC.

2.3.2. Descripteurs d'interfaces protéine-protéines

Les interfaces entre les chaînes protéiques sont détectées suivant une procédure en trois étapes. Tout d'abord, l'interface est grossièrement définie par comptage des distances par paires entre tous les atomes des différentes chaînes. Ne sont gardés que les patchs pour lesquels au moins 20 distances interatomiques sont plus courtes que 5 Å. Dans une deuxième étape, toutes les interactions intermoléculaires (hydrophobe, aromatique, liaison hydrogène, liaison ionique) entre les deux chaînes sélectionnées sont définies avec précision en utilisant des paramètres par défaut de l'outil IChem développé en interne[31]. L'ensemble des règles topologiques, utilisées pour définir les interactions sur la base des paires d'atomes, de leur

distances et des angles, est explicitement décrit dans le précédent rapport décrivant l'outil[31]. Dans une troisième étape, un pseudoatome d'interaction (IPA) est placé à mi-distance de chaque paire d'atomes en interaction selon IChem. Il est important de noter que les IPAs hydrophobes sont groupés si moins de 1,0 Å les séparent[31]. Si le nombre total d'IPAs est supérieur ou égal à 5, l'interface est conservée; sinon elle est écartée. Enfin, un vecteur de 45 nombres réels est généré pour chaque interface restante décrivant sa taille, sa complémentarité chimique et son enfouissement (**Tableau complémentaire 2.3**). Le vecteur final a la forme suivante:

- Le nombre total de pseudo-atomes (un paramètre);
- Le pourcentage de chaque type d'interaction (quatre paramètres);
- La distribution (en compte), pour chaque type d'interaction, de l'enfouissement des IPAs, divisé en dix intervalles allant de 25 à 100% d'enfouissement (4x10 paramètres). L'enfouissement de chaque IPA a été déduit comme précédemment décrit[32] en projetant 120 rayons régulièrement espacées de 8 Å de long ayant leur origine les coordonnées 3D de l'IPA, et en déterminant le nombre intersectant la surface de la protéine environnante.

Au total, le procédé complet comprenant la génération des atomes d'hydrogène, la détection d'interactions et la génération de descripteurs est suffisamment rapide (quelques secondes par entrée PDB) pour être appliqué à l'ensemble de PDB.

Modèle Random Forest

Les modèles de forêts aléatoire (RF) sont générés à l'aide de la librairie RandomForest 4.6-7[33] au sein de l'outil R cran. Un total de 500 arbres de décision (paramètre ntree) a été entrainé sur l'ensemble des descripteurs du jeu d'apprentissage (n=300), en faisant varier le nombre de variables utilisées à chaque nœud (mtry). Une procédure de validation croisée par cinq est utilisée pour séparer le jeu d'entrainement en cinq sous jeux d'entrainement aléatoire (4/5 du jeu) et un sous jeu de test (1/5 des données) et analyser puissance de prédiction du modèle RF sur le jeu de test interne. Pour chaque valeur du paramètre mtry (nombre entier entre 2 et 10), le modèle de validation croisée correspondante a été évalué en fonction des paramètres statistiques suivants:

$$\text{Sensibilité} = TP/(TP+FN)$$

$$\text{Précision} = TP/(TP+FP)$$

$$\text{Specificité} = TN/(TN+FP)$$

$$\text{Exactitude} = (TP+TN)/(TP+FP+TN+FN)$$

$$\text{F-measure} = 2*(\text{Sensitivity}*\text{Precision})/(\text{Sensitivity}+\text{Precision})$$

Dans lequel TP sont vrais positifs (interfaces biologiques prédites pertinentes), FP sont des faux positifs (interfaces cristallographiques prédites pertinentes), TN sont vrais négatifs (interfaces cristallographiques prédites cristallographiques) et FN sont des faux négatifs (interfaces biologiques prédites cristallographiques).

La meilleur valeur de mtry est utilisée (i) pour générer dix modèles depuis le jeu d'apprentissage complet (300 complexes) en faisant varier la graine aléatoire, (ii) utiliser les dix modèles générées pour prédire la pertinence des 100 entrées présentes dans le jeux de données du test externe.

2.3.3. Comparaison aux autres méthodes

Les prédictions IChemPIC ont été effectuées en utilisant le serveur IChemPIC (http://bioinfo-pharma.u-strasbg.fr/IChemPIC). La prédiction finale se fait avec un consensus des dix modèles, il faut au moins cinq des dix prédictions RF («biologique» ou «cristallographique ») pour annoter une protéine testée. Dans le cas d'un nombre égal de prédictions pour les deux types, l'interface est prédite cristallographique. IChemPIC a été comparé à quatre méthodes de référence (NOXClass[13], PISA[9], DiMoVo[12] et EPPIC [20]) sur trois jeux de test externes. Pour chacun de ces outils, les paramètres standards disponibles dans leur version en ligne ont été choisis, en donnant comme entrée soit le code PDB et le nom des chaines impliquées (interfaces biologiques) ou en soumettant le fichier de structure préalablement préparé (contacts cristallins). Pour la classification SVM multi-niveaux NOXClass (http://noxclass.bioinf.mpi-inf.mpg.de/), les probabilités d'appartenance au différentes classes (biologiques, cristallographiques) ont été retenues pour chaque paire de chaînes de protéines, en conservant trois descripteurs (zone d'interface, ratio entre l'aire d'interface et la surface des protéines, la composition en acides aminés de l'interface). Dans PISA (http://www.ebi.ac.uk/pdbe/pisa/), l'interface a été définie comme biologique si l'interface correspondante a été prédite pour être stable parmi tous les ensembles proposés.

Dans le cas contraire, l'interface a été prédite cristallographique. En utilisant la méthode de prédiction Dimovo (http://albios.saclay.inria.fr/dimovo), un score supérieur à 0,50 a été utilisé pour assigner une fonction biologique potentielle à une interface. Enfin, les prédictions EPPIC (Bio ou Xtal) ont été effectuées sur un serveur web (http://www.eppic-web.org/ewui/) et sur la base du système de vote par consensus (score final) en tenant compte des quatre descripteurs (cœur de l'interface, la géométrie de l'interface, conservation du cœur, surface de l'interface).

2.4. Résultats et discussion

2.4.1. Paramétrisation du jeu FDS d'interfaces protéine-protéine droguable

Aucun des jeux de données de référence n'est adapté à discriminer les contacts cristallins des interfaces protéine-protéine biologiquement pertinentes. D'un côté, les jeux de données historiques[10, 13, 21, 22] sont déséquilibrés par une majorité des entrées cristallographiques possédant une aire faible (500-1000 Å²) et par les véritables entrées biologiquement pertinentes de grande taille (1000-3000 Å²). De l'autre côté, le jeu de donnée DC[20] corrige cette anomalie en sélectionnant les entrées avec une répartition homogène des aires d'interface (1000 à 1500 Å²) qui cependant tombe encore en dehors du domaine d'applicabilité de nombreuses PPI biologiquement pertinentes et importantes (par exemple interface p53-mdm2 de 780 Å², PDB ID: 1YCR) modulées par des inhibiteurs de faible poids moléculaire[35]. Les deux jeux Bahadur et DC qui ont des données qui ne se chevauchent pas beaucoup par rapport à la plage de la zone d'interface ont donc été fusionnés pour élargir le domaine d'applicabilité de nos prochaines prédictions. Nous avons finalement rassemblé manuellement un ensemble supplémentaire de 115 PPIs biologiquement pertinentes pour obtenir un nombre final de 400 interfaces qui a été divisé en un jeu d'entrainement (75% des données) et un jeu de test (25% des données). L'inspection de la distribution respective des tailles de l'aire d'interface dans les deux ensembles ne révèle aucun biais majeur, bien que les interfaces biologiques restent en moyenne un peu plus grandes que les contacts cristallins (figure 2.1). Nous démontrerons plus tard que la taille de l'interface n'a pas une influence majeure dans la discrimination cristallographique à partir d'interfaces biologiques.

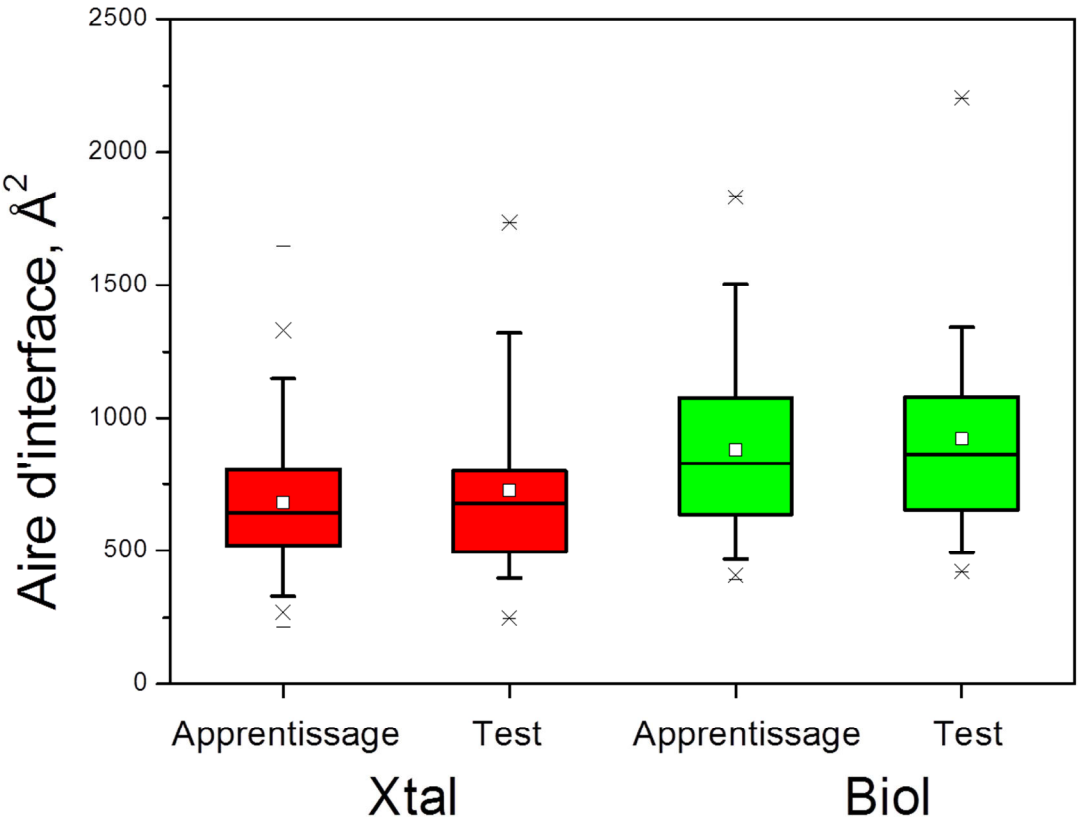


Figure 2.1 *Distribution des aires d'interface dans les jeux FDS apprentissage et de test (Xtal, l'interface est cristallographique; Biol, l'interface est biologiquement pertinente). Les boîtes délimitent les 25ème et 75ème percentiles, les moustaches délimitent les 5ème et 95ème percentiles. Les valeurs médianes et moyennes sont indiquées par une ligne horizontale et un carré dans la boîte. Les croix délimitent les 1% et 99ème percentiles, respectivement. Les valeurs minimales et maximales sont indiquées par un tiret.*

Environ 80% des contacts cristallins (jeu d'entrainement et de test) concerne des enzymes dont le site catalytique est très connu, le reste étant composé par des protéines de transport. La proportion d'enzymes dans les interfaces pertinentes est plus faible (environ 55%), l'ensemble biologiquement pertinent présente plusieurs exemples de complexes de reconnaissance impliqués dans des processus biologiques importants (par exemple la reconnaissance immunitaire, la signalisation cellulaire, adhésion cellulaire, la transcription).

La résolution moyenne des structures cristallographiques est de 1.84 ± 0.35 Å pour les contacts cristallins et 2.10 ± 0.58 Å pour les interfaces biologiquement pertinentes. Une grande majorité des structures a été résolue à haute résolution (<2.5 Å). Durant la préparation de la protéine, nous avons vérifié que les chaines impliquées dans l'interface étaient complètes. Aucun des 400 complexes PDB décrits ici ne présente une chaîne latérale

incomplète à l'interface protéine-protéine sélectionnée. Une vérification des ions présents aux interfaces a aussi été effectuée et a montré qu'il n'y avait pas d'ions importants à proximité des interfaces sélectionnées. Conserver les molécules d'eau présente à l'interface aurait été un plus mais il n'y a pas de molécules d'eau dans 184 des 400 entrées ce qui nous a imposé de les retirer toutes. Nous avons un protocole unique pour le traitement de l'ensemble des structures utilisées. En observant les 400 structures utilisées, nous avons observé que des molécules d'eau n'étaient utilisées à l'interface que dans 30% des structures en contenant (30% des 216 entrées contenant de l'eau). Les liaisons réalisées par l'eau sont généralement une liaison hydrogène unique. Ces observations nous confortent dans dans le fait que supprimer les molécules d'eau n'a pas de conséquences majeures sur ces travaux.

2.4.2. Détection des interfaces et génération des descripteurs

Nous avons premièrement détecté les interfaces entre les chaines, puis déterminé explicitement toutes les interactions non liées (contacts hydrophobes, interactions aromatiques, liaisons hydrogènes et ioniques) et généré des pseudo-atomes d'interface (IPA) pour décrire chaque interaction (**Figure 2.2**).

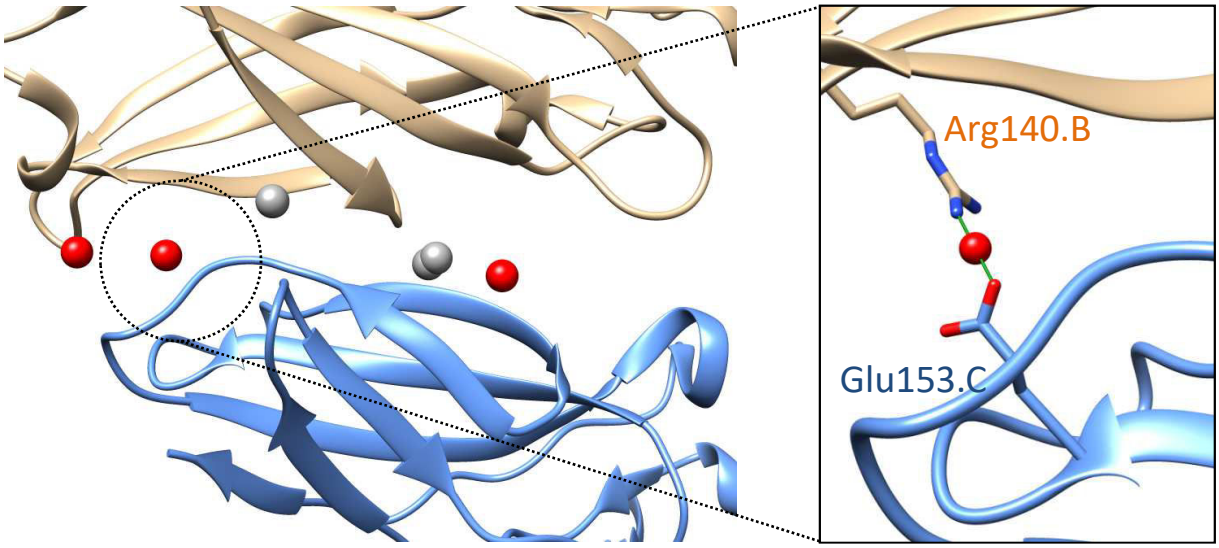


Figure 2.2 *Interface (PDB ID: 4NNY) entre la sous-unité alpha du récepteur à l'interleukine-7 (bronze, chaîne C) et la cytokine « receptor-like » facteur 2 (bleu, chaîne C). Six pseudoatomes d'interaction (sphères) sont placés à mi-distance de chaque paire d'atomes en interaction et se voient attribuer une propriété correspondant au type d'interaction (hydrophobe, aromatique, liaison hydrogène, une liaison ionique). L'image de droite est un zoom sur une liaison ionique unique avec affichage explicite des chaînes latérales.*

Un assemblage moléculaire complexe des plusieurs milliers d'atomes peut être représenté par un ensemble d'IPA beaucoup plus simple (60-70 points en moyenne) décrivant à la fois la nature et l'enfouissement des interactions correspondantes. Puisque nous considérons explicitement des liaisons hydrogène, il convient de noter que tous les atomes d'hydrogène sont ajoutés aux fichiers PDB natifs, tout en optimisant les états de tautomérie et de protonation des acides aminés[30].

Bien que les interfaces biologiquement pertinentes aient un nombre moyen d'IPA plus grand (86 ±40) que les contacts cristallins (50 ± 30), le pourcentage moyen des différents types d'interactions reste similaire dans les deux ensembles (**Tableau 2.1**)

Table 2.1. *Pourcentage moyen pour chaque type d'interactions présente aux interfaces protéine-protéine.*

Type d'interactions	Interface protéine-protéine[a]	
	Cristallographique	Biologique
Hydrophobe	78.06 ± 15.70	83.32 ± 9.71
Aromatique	0.24 ± 1.14	0.10 ± 0.32
Liaison Hydroène	17.97 ± 12.11	13.51 ± 7.24
Liaison Ionique	3.65 ± 5.80	3.00 ± 3.87

[a] statistiques établies sur 27186 interactions protéine-protéine (200 contacts cristallins et 200 interfaces pertinentes du jeu FDS), détectées avec IChem.[31]

Comme prévu, les contacts hydrophobes sont majoritaires et représentent près de 80% des interactions. De plus, ils sont plus présents dans les interfaces biologiquement pertinentes (**Tableau 2.1**). Les interactions aromatiques (face/face et face contre arête) sont rares mais elles existent et sont un peu plus représentées dans les contacts cristallins, comme le montraient de précédentes observations[21]. Les liaisons hydrogènes sont plus fréquentes dans les contacts cristallins que dans les assemblées biologiques. Cependant, la qualité des interactions hydrogènes (force, accessibilité) n'est pas prise en compte dans cette analyse. Finalement les interactions ioniques sont également représentées dans les deux ensembles. Même si les interactions métalliques ont été retrouvées quelques fois, cette valeur n'a pas été transmise aux modèles de forêts aléatoires afin de générer le descripteur d'interface le plus simple possible.

2.4.3. Modèle de classification binaire en forêt aléatoire

La Forêt Aléatoire (RF) est une méthode très polyvalente d'apprentissage automatique pour la classification et la régression qui repose sur un grand nombre d'arbre de décision indépendants[36]. Chaque arbre est créé par « bootstrap » des données d'origine en utilisant un sous-ensemble aléatoire de caractéristiques. Ensuite, les arbres individuels sont combinés à travers un processus de vote pour fournir une prédiction non biaisée. En contraste avec les arbres de décision unique, les forêts aléatoires ont une variance faible et très peu de biais. Considérant que les forêts aléatoires ont peu de paramètres à régler (nombre d'arbres, nombre de variables à chaque division), la méthode est facile à utiliser afin de produire un modèle raisonnablement rapide et efficace. Parmi les nombreuses applications potentielles, le RF est de plus en plus utilisé dans les sciences de la vie que ce soit en tant que classifieur ou comme méthode de régression non-linéaire[37].

Dans notre application, le nombre d'arbres (paramètre ntree) a été fixé à 500. Outre une nette influence sur le temps global de calcul, les variations de ce paramètre n'ont pas influencé les résultats présentés. Le nombre de variables échantillonnées au hasard en tant que candidats à chaque division (paramètre mtry) a été systématiquement varié de deux à dix variables utiles, et chaque modèle a été répété cinq fois en faisant varier la graine aléatoire de départ. En utilisant une valeur de mtry égal à quatre, la modélisation des forêts aléatoires conduit à un modèle stable et robuste avec validation croisée par 5 (F-mesure = 0,776 ± 0,09) lorsqu'elle est appliquée à l'ensemble de la formation FDS (**Tableau 2.2**).

Table 2.2 *Statistique du meilleur modèle RF généré avec le jeu FDS d'entrainement*

Paramètre	Apprentissage (n=300)[a]	Test (n=100)[b]
Sensibilité	0.794 ± 0.017	0.728 ± 0.014
Precision	0.759 ± 0.010	0.745 ± 0.018
Specificité	0.747 ± 0.014	0.750 ± 0.025
Exactitude	0.771 ±0.009	0.739 ± 0.012
F-measure	0.776 ± 0.009	0.736 ± 0.010

[a] Moyenne et déviation standard des meilleurs modèles avec validation croisée (*ntree*=500, *mtry* = 4), répétés 5 fois avec différentes graines aléatoire.

[b] moyenne et déviation standard (*ntree=500*, *mtry* = 4) sur la prédiction des 100 interfaces du jeu FDS de test.

Le modèle est également bon pour prédire aussi bien les interfaces biologiques (sensibilité) que les interfaces cristallographiques (spécificité). Lorsqu'il est appliqué au jeu externe FDS de 100 PPIs, une baisse modérée de la précision (0,739 ± 0,012) et de la F-mesure (0,736 ± 0,010) est observée, mais le modèle est toujours robuste et prédit aussi bien les deux catégories de PPIs (sensibilité = 0,728 ± 0,014; spécificité = 0,750 ± 0,025; **Tableau 2.2**).

Pour être sûr que les données observées ne sont ni le résultat de surentraînement, ni de corrélation chanceuse, nous avons d'abord effectué un test de y-scrambling par l'assignation aléatoire de la variable dépendante (cristallographique ou biologique) à chacune des 400 interfaces protéine-protéine du jeu de données FDS. Comme prévu, la F-mesure des modèles RF correspondantes (mêmes paramètres que ci-dessus) à sensiblement chuté à une valeur moyenne de 0,515 et 0,502, lorsqu'il est appliqué au jeu d'entrainement et au jeu de test externe. Nous avons ensuite calculé 45 modèles RF (dix essais/modèle) dans lequel les valeurs des 45 descripteurs ont été itérativement permutées pour chaque entrée de l'ensemble de la formation. Pour l'ensemble des 45 descripteurs, les 300 valeurs de descripteur calculées précédemment ont été assignées au hasard (apprentissage). L'analyse des variations de la F-mesure moyenne pour l'ensemble du jeu d'apprentissage permet d'identifier les paramètres les plus importants parmi nos 45 descripteurs (**Figure 2.3**).

Sur les 45 descripteurs, 11 ont une réelle contribution au modèle général (> 1% diminution de la F-mesure) lorsque leurs valeurs respectives sont interverties. Les paramètres les plus importants sont clairement le nombre de pseudoatoms d'interaction (nPTS) et le pourcentage de contact hydrophobes très enfouis (descripteurs Hydro7-hydro10, **Tableau annexe S2.3**).

Permuter les valeurs prises par le nombre total d'IPAs (nPTS) diminue la F-mesure globale du modèle de 1,6% (Figure 2.3). Alors que des contacts hydrophobes accessibles (paramètres Hydro1-Hydro6) ne contribuent pas vraiment à la F-mesure globale, les interactions hydrophobes plus enfouies (Hydro7, Hydro8, Hydro9, paramètres Hydro10) sont vraiment critiques. De manière remarquable, la permutation de valeur du paramètre Hydro10 (pourcentage des contacts hydrophobes 100% enfouis) diminue la F-mesure du modèle RF de près de 3% (**Figure 2.3**). En conséquence, les résidus du cœur de l'interface hydrophobes, définis comme enfouis d'au moins 95% ont récemment été décrits comme les principaux

déterminants de interfaces[20]. Des paramètres de moindre importance, mais toutefois encore utile, sont les pourcentages des autres interactions (les liaisons hydrogène, liaisons ioniques) très enfouies qui tendent à être plus élevés dans les interfaces biologiques que dans les contacts cristallins (**Figure 2.3**).

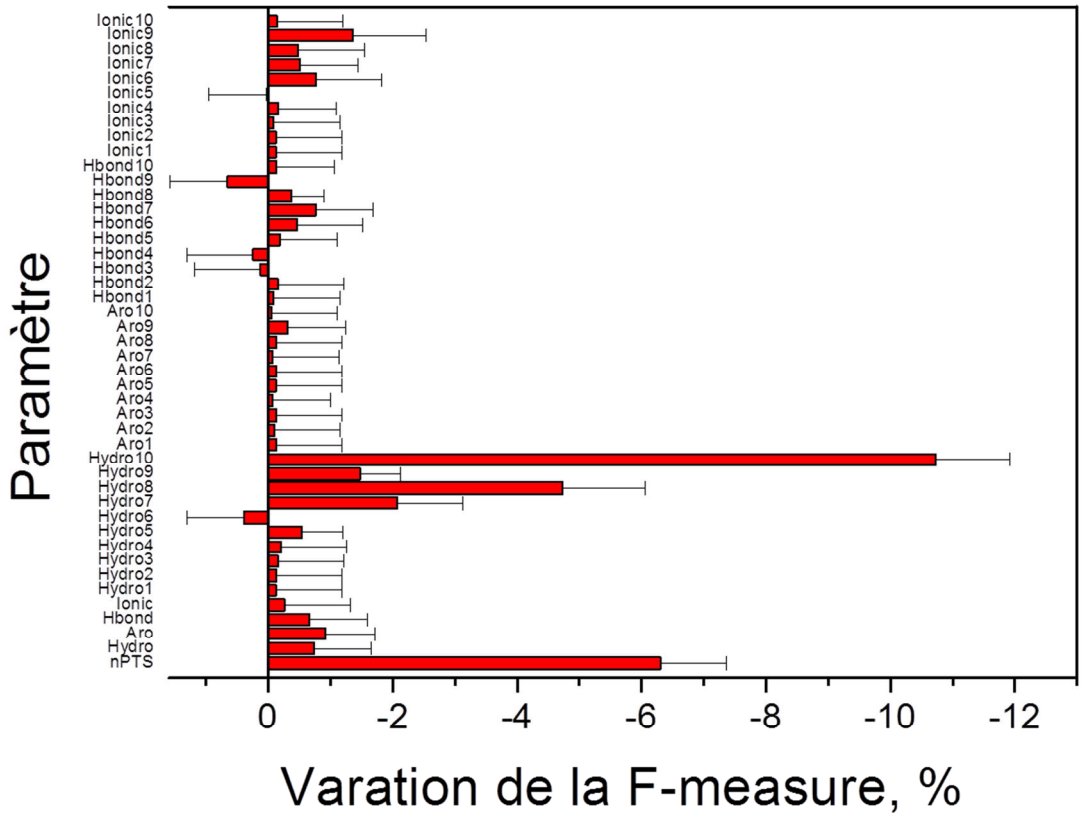


Figure 2.3 *Influence de la permutation des valeurs de descripteur sur la F-mesure moyenne de dix modèles RF obtenus avec les meilleurs paramètres de validation croisée (ntree = 500, mtry = 4) et entrainés sur le jeu d'apprentissage FDS.*

Permuter les valeurs de quatre des 45 paramètres (Hydro5, Aro8, Aro9, Hbond7) conduit à de légèrement meilleurs modèles RF. La plus forte baisse observée en F-mesure (mélange des valeurs de paramètre Hydro8) est seulement de 5% et est probablement expliquée par des effets compensatoires sur l'élimination du descripteur le plus critique. Pour démontrer cette hypothèse, nous avons supprimé le descripteur Hydro8 à partir du vecteur initial, recalculé un modèle RF sur l'ensemble des n-1 descripteurs (F-mesure de 0,705 sur l'ensemble de la formation) et permuté à nouveau itérativement les valeurs des descripteurs. Cette fois, le descripteur le plus critique est Hydro10 (ancien second descripteur le plus important) avec une diminution beaucoup plus forte de la F-mesure (11,3 ± 3,3%). Cette observation illustre parfaitement notre hypothèse et l'effet compensatoire du paramètre Hydro10 lors du retrait de l'influence du descripteur Hydro8.

La contribution plus faible du paramètre Hydro9 (nombre de IPAs hydrophobes enfouis entre 91,6% et 100%) par rapport à celle de Hydro8 (compte des IPAs hydrophobes enterrée entre 83,3% et 91,6%) et Hydro10 (nombre de 100% enterré hydrophobe IPAs) est intrigante et peut être expliquée par une distribution particulière des valeurs des paramètres lorsque l'on compare les contacts cristallins et les interfaces biologiquement pertinentes (**Figure 2.4**). Par conséquent, les distributions de Hydro8 et Hydro10 sont clairement différentes lors de l'examen des deux sous-ensembles d'interfaces (valeurs plus élevées du paramètre Hydro8 dans les contacts cristallographiques, valeurs plus élevées du paramètre Hydro10 dans les interfaces biologiques). Curieusement, les valeurs des paramètres Hydro9 sont distribuées de manière similaire (**Figure 2.4**), ce qui explique pourquoi ce paramètre contribue moins au modèle de validation croisée RF.

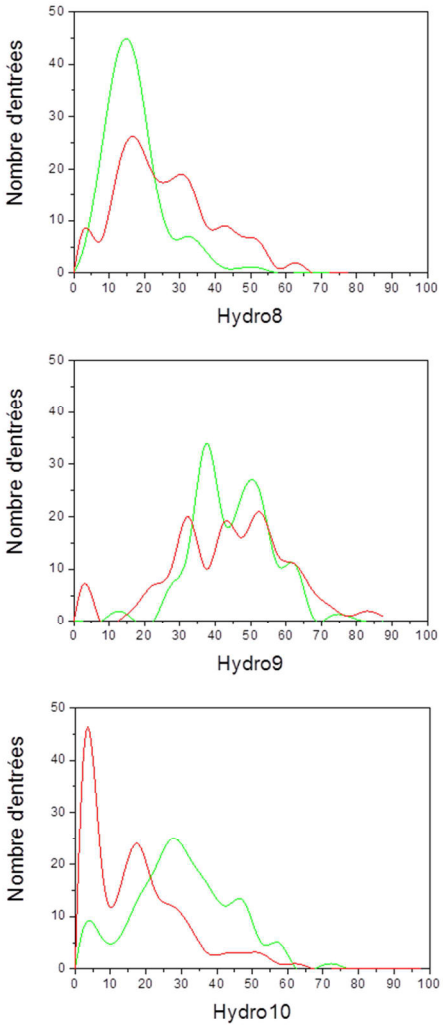


Figure 2.4 *Distribution des paramètres Hydro8, Hydro9 et Hydro10 au sein du jeu d'apprentissage FDS (vert : interfaces biologiquement pertinentes, rouge : contacts cristallins)*

Pour confirmer l'importance suggérée ci-dessus de certains paramètres d'interface (NPTs, Hydro7, Hydro8, Hydro9, Hydro10), nous avons classé les 300 interfaces d'apprentissage par valeur décroissante de chaque descripteur (45 listes d'entrées PDB classés par ordre décroissant pour le descripteur étudié). Nous avons ensuite procédé à une classification binaire des 300 entrées (cristallographiques, biologiques) dans les rangs obtenus dans ces 45 listes. Un descripteur parfait conduirait à une classification (ROC AUC = 1) dans laquelle toutes les 150 interfaces biologiques sont classés avant la première interface cristalline. En utilisant la classification ROC, on peut donc estimer l'importance relative de chaque descripteur pour discriminer les deux catégories. Toute classification à base de descripteur unique avec un AUC plus élevé que 0,7 (**Figure 2.5**) indique que ce descripteur est particulièrement efficace. Cette analyse confirme le rôle crucial des deux paramètres (nPTS, Hydro10) sur la discrimination des deux sous-ensembles d'interface.

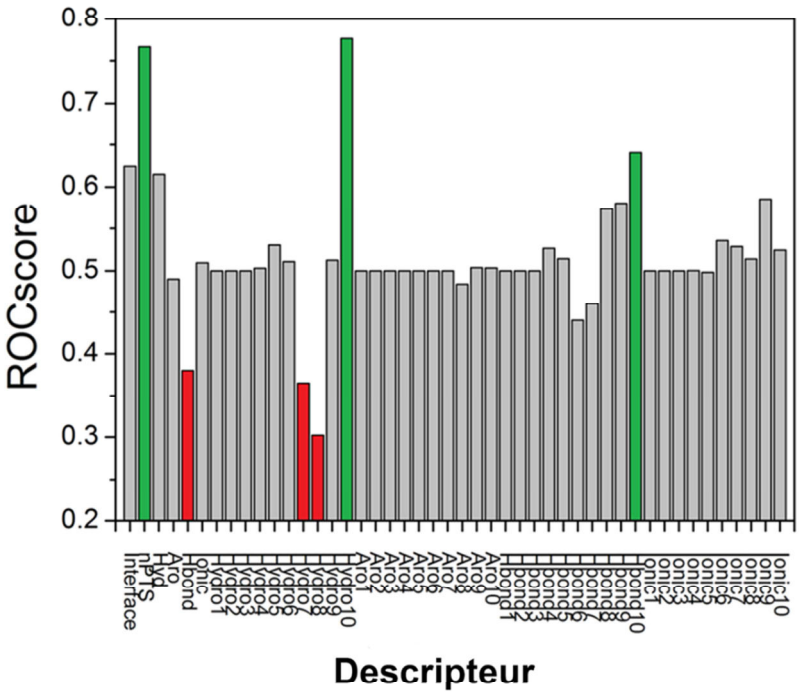


Figure 2.5 *Aire sous la courbe ROC pour une classification binaire (contact cristallin, interface pertinente) des 300 interfaces (FDS apprentissage) lors d'un classement décroissant de chacun des 45 descripteurs d'IChemPIC*

Cette analyse complémentaire montre également que les valeurs observées pour trois descripteurs (Hbond, Hydro7 et Hydro8) sont en effet plus élevés pour les contacts cristallins (ROCscore <0,50) et contribuent donc également à discriminer les deux ensembles d'entrées PDB. Il est important de noter, qu'en utiliser l'aire d'interface comme descripteur ne conduit pas à une bonne classification binaire (ROCscore = 0,59) qui confirme que le jeu d'apprentissage FDS est très bien équilibrée par rapport à ce critère important qui a été négligé par le passé.

2.4.4. Comparaison d'IChemPIC aux méthodes existantes

IChemPIC a été comparée à quatre méthodes reconnues (NOXClass[13], PISA[9], DiMoVo[12] et EPPIC[20]) pour prédire la nature des PPI provenant de trois jeux de test externes différents.

Au vu de la diversité des interfaces dans le jeu de données FDS, il est peu surprenant que la précision observée des méthodes existantes est nettement inférieure à celle rapportée dans les publications qui les décrivent[9, 13 12, 20]. NOXClass est remarquablement sensible (bon taux de vrais positifs), mais au prix d'une spécificité beaucoup plus faible (faible taux de vrai négatif). En revanche, EPPIC et dans une moindre mesure Dimovo sont spécifiques dans la détection des contacts cristallins (spécificité = 0,949), mais sont moins performants dans la reconnaissance des interfaces biologiquement pertinentes (faible sensibilité), notamment lorsque la zone d'interface est faible (<750 Å², **Tableau supplémentaire S2.4**). PISA, la méthode actuellement utilisée pour prédire les assemblages macromoléculaires dans la RCSB PDB, est le plus stable par rapport à tous les paramètres statistiques pris en compte (**Tableau 3**). Au final, IChemPIC apparaît toujours comme la méthode de choix pour une classification binaire des interfaces protéine-protéine, car elle offre une performance constante et robuste pour prédire les interfaces biologiquement pertinentes et les contacts cristallins, quelle que soit leurs tailles (voir les prédictions complètes dans le **Tableau complémentaire S2.4**)

D'une part, il est inférieur à NoxClass et PISA pour la détection d'interfaces biologiques, mais beaucoup plus précis pour prédire les contacts cristallins. D'autre part, IChemPIC est moins précis que les prédictions des contacts cristallins d'EPPIC mais nettement meilleur dans la prédiction des interfaces biologiquement pertinentes (**Tableau 2.3**). Sur les cinq méthodes testées ici, IChemPIC est la seule méthode capable de prédire

avec une bonne précision l'ensemble des PPIs, quel que soit le statut (contacts cristallins, biologiquement pertinent) ou la taille de l'interface.

Table 2.3 *Comparaison d'IChemPIC à quatre méthodes de référence pour prédire le statut (biologique, cristallin) du jeu de données externe FDS (n=100)*

Statistiques	Méthodes				
	IChemPIC[a]	NOXClass	DiMoVo[b]	PISA[c]	EPPIC[d]
Sensibilité	0.740	0.878	0.480	0.771	0.667
Précision	0.755	0.694	0.857	0.725	0.909
Spécificité	0.760	0.525	0.733	0.674	0.949
Exactitude	0.750	0.719	0.538	0.725	0.826
F-measure	0.747	0.775	0.615	0.747	0.769

[a] prévisions consensuelles (biologiques ou cristallographiques) sur dix modèles RF indépendants. Dans le cas d'un nombre égal de prédictions pour les deux propriétés, l'interface est prédite cristallographique. Les prévisions ont été obtenues à l'aide du serveur IChemPIC (http://bioinfo-pharma.u-strasbg.fr/IChemPIC).
[b] 35 entrées commune au jeu d'entrainement de DiMoVo ont été supprimées.
[c] Deux entrées (1i5h, 1y7q) n'ont pu être prédites par PISA (données cristallographiques absentes); 7 entrée présentes dans le jeu d'entrainement de PISA ont été supprimées.
[b] 29 entrées communes aux jeux d'entrainements de EPPIC et FDS ont été supprimées.

Du fait qu'IChemPIC a été entrainé sur le jeu de données FDS, il est juste de comparer ses performances sur des jeux de tests externes totalement indépendants. Nous avons donc choisi trois jeux de données externes supplémentaires (IPAC[7], Ponstigl[10] and Bahadur[21]) contenant des entrées PDB non présentes dans le jeu d'apprentissage FDS. Les deux premiers jeux ont notamment été utilisés pour l'analyse comparative de la plupart des outils similaires à IChemPIC. Comme indiqué précédemment,[12, 20] les jeux de donées Bahadur et Ponstingl ne sont pas très instructifs à cause d'une fort biais vers de petits contacts cristallins et des grandes interfaces biologiquement pertinentes. En conséquence, tous les programmes, y compris IChemPIC obtenir une excellente précision (0,85 à 0,95) pour prédire la nature de ces entrées (**Tableau 2.4**). IChemPIC présente notamment la plus haute F-mesure (0,932 et 0,870, respectivement) sur ces deux ensembles externes, ce qui indique sa robustesse

dans la prédiction des interfaces aussi bien biologiques et cristallographiques (voir les prédictions complètes des **Tableaux supplémentaires 2.5 et 2.6**).

Le dernier jeu externe (jeu IPAC3)[7] est composé de 66 protéines hétérodimériques dont la structure cristalline est connue et dont des constantes de liaison sont expérimentalement déterminées. Il permet notamment d'évaluer la sensibilité de la méthode pour prédire les interfaces biologiques très différentes. Sur les cinq méthodes, NOXClass présente les meilleures performances (seulement la sensibilité est rapportée) lorsqu'il est appliqué à cet ensemble de données (**Tableau 2.4**). Étonnamment, cette méthode ne manque aucune entrée même appliquée aux complexes de plus basse d'affinité (Kd $<10^{-5}$ M, **Tableau supplémentaire 2.7**). Compte tenu de la propension de NoxClass à surestimer les interfaces biologiques dans les jeux de tests externes examinés précédemment (sensibilité >> précision; **Tableaux 2.3 et 2.4**), son excellente performance devrait donc être considérée avec une extrême prudence. D'autres méthodes sont en effet sensibles à la force des complexes correspondants et ont logiquement échoué à prédire comme biologique les complexes de faible affinité (**Tableau 2.7**). Parmi ces méthodes, IChemPIC présente clairement la plus grande exactitude (**Tableau 2.4**).

Table 2.4 *Performance de IChemPIC par rapport aux méthodes de l'état de l'art dans la prédiction du statut (cristallographique, biologique) de trois jeux de référence indépendants.*

Jeux	Nombre d'interfaces		Statistiques	Méthode				
	Crystallographic	Biological		IChemPIC[a]	NOXClass	DiMoVo	PISA	EPPIC
Bahadur[b]	20	122	Sensibilité	0.902	0.938	n.a.[c]	0.918	0.885
			Précision	0.965	0.892	n.a.	0.875	0.973
			Spécificité	0.800	0.450	n.a.	0.556	0.850
			Exactitude	0.887	0.855	n.a.	0.835	0.880
			F-measure	0.932	0.915	n.a.	0.896	0.927
Ponstingl[d]	67	76	Sensibilité	0.882	0.919	0.714	n.a.[e]	0.895
			Précision	0.859	0.760	0.714	n.a.	0.840
			Spécificité	0.831	0.731	0.930	n.a.	0.806
			Exactitude	0.858	0.822	0.887	n.a.	0.853
			F-measure	0.870	0.832	0.714	n.a.	0.866
IPACdb[f]	0	66	Sensibilité	0.706	0.946	0.394	0.682	0.636

[a] prédictions consensuelles (biologiques ou cristallographiques) sur dix modèles RF indépendants. Dans le cas d'un nombre égal de prédictions pour les deux propriétés, l'interface est prédite cristallographique. Les prévisions ont été obtenues à l'aide du serveur IChemPIC (http://bioinfo-pharma.u-strasbg.fr/IChemPIC).
[b] 142 structures PDB (122 biologiques, 20 cristallines) ne sont pas présentes dans le jeu d'entrainement de IChemPIC. Les entrées présentes dans les jeux d'entrainement de NoxClass (n = 25), Dimovo (n = 142) et PISA (n = 63) ont été supprimées lorsque la méthode correspondante a été utilisée pour la prédiction.
[c] pas applicable car DiMoVo a été entrainé sur le jeu de données Bahadur.
[d] 143 structures PDB (76 biologiques, 67 cristallines) ne sont pas présentes dans le jeu d'entrainement de IChemPIC. Les entrées présentes dans les jeux d'entrainement de NoxClass (n = 14), Dimovo (n = 72), et PISA (n = 109) ont été supprimées lorsque la méthode correspondante a été utilisée pour la prédiction.
[e] non applicable car PISA a été entrainé sur le jeu de données Ponstingl.
[f] 66 PDB structures hétérodimères (Validation Set 3) de constantes de liaison connues. Les entrées présentes dans les jeux d'entrainement de IChemPic (n = 15) et NoxClass (n = 10) ont été supprimées lorsque la méthode correspondante a été utilisée pour la prédiction.

2.4.5. Application pratique d'IChemPIC à l'ensemble de la PDB et explication d'erreurs

IChemPIC a ensuite été appliqué pour classer 4950 structures d'interfaces non redondantes extraites de Dockground[38]. Toutes ces structures sont basées sur le fichier de l'unité biologique (Biounit) déduit des prédictions PISA et fourni en ligne par le RCSB PDB. Environ 30% (1493 au total) de ces interfaces sont néanmoins prédites comme contacts cristallins par IChemPIC (**Tableau S2.8).** Ces écarts résultent de trois causes principales (**Figure 2.6**).

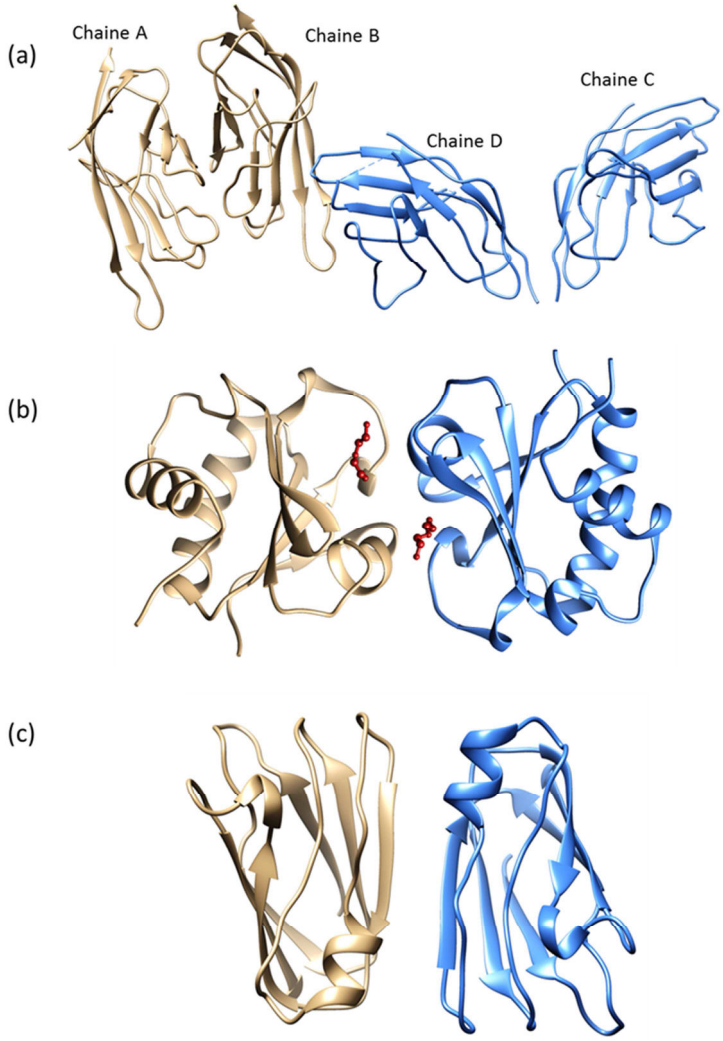


Figure 2.6 *Exemple d'interfaces prédites pertinente par DockGround qu'IChemPIC rejette.* ***A)*** *complexe CTLA-4 (Chaines A, B) / B7-2 (Chaines C,D; PDB ID 1i85),* ***B)*** *Phosducine-like protéine humaine 2 avec des molécules de PEG liées (sphère rouge) à l'interface (PDB ID 3evi);* ***C)*** *plastocyanine de la cyanobactérie Synechocystis sp. PCC 6803 (PDB ID 1pcs).*

Premièrement, notre méthode comme tout autre est loin d'être parfaite et ne parvient pas à prédire correctement 25% des entrées du jeu de test (**Tables 2.3 et 2.4**). Dans de nombreux cas, l'erreur se produit parce qu'IChemPIC ne prédit pas les interfaces des structures quaternaires. Par conséquent, deux chaînes peuvent former une interface stable en fonction du contexte précis ou d'un état d'oligomérisation beaucoup plus grand. Par exemple, l'interface isolée entre CTLA-4 (chaîne B) et B7-2 (chaîne D) est prédite non pertinente (PDB ID 1I85, interface = 621 A^2, NPTS = 42), car elle existe seulement dans un plus grand réseau (**Figure 2.6a**) expliquant l'organisation périodique de ces molécules dans la synapse immunologique à la surface cellulaire[39]. Deuxièmement, un grand nombre d'interfaces de petites dimensions (149 sont plus petites que 500 A^2) sont des conséquences évidentes des conditions de cristallisation, facilitée par la présence d'un sel ou d'un agent de précipitation. Ce cas est bien illustré par la structure cristallographique de la protéine phosducine humaine 2[40] (PDB ID 3EVI, interface = 422 A^2, NPTS = 21) qui présente deux molécules de diéthylène glycol stabilisant une interface homodimèrique artefactuelle (**Figure 2.6b**). Enfin, un empilement cristallin énergétiquement fort peut produire des interfaces artificielles, comme illustré ici par l'assemblée prédite biologiquement pertinente d'une plastocyanine de cyanobactérie (**Figure 2.6c**) avec une parfaite symétrie C2 (PDB ID 1PCS, interface = 395 A^2, NPTS = 6), mais pas de pertinence biologique[41].

De l'exercice actuel, nous estimons à 15% le pourcentage d'unités biologiques PDB pour lesquels l'état d'oligomérisation proposé est susceptible d'être incorrect. Nous suggérons donc fortement l'utilisation d'un classificateur précis comme IChemPIC pour réduire le nombre de ces ensembles biologiques erronés et permettre la conception d'inhibiteurs de PPI sur les interfaces biologiquement pertinentes.

2.5. Conclusions

Nous présentons ici une nouvelle approche informatisée (IChemPIC) pour distinguer les interfaces protéine-protéine biologiquement pertinentes et les contacts cristallins. Étant donné qu'aucun des jeux de données de référence existants n'est satisfaisant, notamment pour prédire de petites interfaces biologiques droguables; nous avons défini manuellement de nouveaux jeux d'apprentissage et de tests externes (FDSdataset) pour permettre: (i) une couverture comparable des zones d'interface pour les contacts cristallins et les interfaces biologiques pertinentes, (ii) l'application aux interfaces protéine-protéine de petite taille connues pour être modulables par des molécules de faibles poids moléculaire.

En décrivant l'interface par un simple vecteur de 45 réels se concentrant sur les interactions intermoléculaires, les machines d'apprentissages peuvent être utilisées pour classer les interfaces selon leur statut (contacts cristallins ou interface biologiquement pertinente). En raison de son niveau de simplicité et de sa faible paramétrisation, la méthode d'apprentissage forêt aléatoire (Random Forest) a été choisie pour obtenir un modèle qui distingue les contacts cristallins des interfaces biologiques avec une précision robuste de 80%. En ce qui concerne les autres méthodes actuelles, IChemPIC est la seule approche capable de prédire avec la même précision les deux catégories d'interfaces protéine-protéine, quel que soit le jeu de test externe. Il existe de nombreux avantages à utiliser IChemPIC par rapport à d'autres méthodes: (i) L'ajout explicite des atomes d'hydrogène permet d'utiliser des liaisons hydrogène comme descripteurs pour le développement de modèles; (Ii) la méthode peut être appliquée à des interfaces présentant des modifications post-traductionnelles; (Iii) la performance est indépendante de la taille de l'interface, (iv) le domaine d'applicabilité est vaste allant des petites interfaces protéine-protéine biologiques (500 Å²) à des contact cristallographique de grande taille (1500 Å²).

Il faut toutefois reconnaître qu'IChemPIC est actuellement paramétré pour traiter les interfaces entre deux chaînes de protéines. Par exemple, les trois interfaces possibles (AB, BC, AC) d'un hétérotrimère ABC seront prédites soit cristallographique ou biologique, mais aucune prédiction ne sera effectuée pour les interfaces triples entre une chaîne et les deux autres. En d'autres termes, aucune prédiction n'est faite pour la totalité de l'ensemble comme avec PISA par exemple. Cet inconvénient explique certains des faux négatifs observés par IChemPIC et pourrait être facilement corrigé en permettant la détection de toutes les interactions possibles entre une chaîne unique et son environnement natif de protéine

complète. Cependant, étant donné que notre méthode est principalement destiné à détecter toutes les interfaces biologiquement pertinentes de la PDB pouvant être modulées (inhibition ou stabilisation) par des molécules de faibles poids moléculaires, nous préférons limiter IChemPIC au traitement de deux chaînes afin de localiser l'interface ciblée par un modulateur potentiel de PPI. IChemPIC peut être utilisé en ligne (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) à partir d'un identifiant PDB ou un fichier d'entrée PDB fourni par l'utilisateur.

2.6. Remerciements

Nous remercions le Centre National de la Recherche Scientifique (CNRS, Institut de Chimie) et la Région Alsace pour la bourse de doctorat. Le centre de calcul HPC (Université de Strasbourg, France) et le Centre de l'IN2P3 (CNRS, Villeurbanne, France) sont remerciés pour l'attribution de temps de calcul et pout l'excellent support informatique. Nous remercions sincèrement le Professeur M. Rarey (Université de Hambourg, Allemagne) pour nous avoir fourni une version exécutable de Protoss. O. Sperandio, X. Morelli, P. Roche et E. Kellenberger sont remerciés pour la lecture critique du manuscrit et leurs remarques utiles.

2.7. References

1. Ivanov, A. A., Khuri, F. R., Fu, H. Targeting Protein-Protein Interactions as an Anticancer Strategy. *Trends Pharmacol. Sci.* **34**, 393-400 (2013).
2. Villoutreix, B. O., Kuenemann, M. A., Poyet, J.-L., Bruzzoni-Giovanelli, H., Labbé, C., Lagorce, D., Sperandio, O., Miteva, M. A. Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology. *Mol. Inf.* **33**, 414-437 (2014).
3. Wells, J. A., McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **450**, 1001-1009 (2007).
4. Murray, C. W., Verdonk, M. L., Rees, D. C., Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **33**, 224-232 (2012).
5. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).
6. Janin, J. Specific Versus Non-Specific Contacts in Protein Crystals. *Nat. Struct. Biol.* **4**, 973-974 (1997).
7. Mitra, P., Pal, D. Combining Bayes Classification and Point Group Symmetry under Boolean Framework for Enhanced Protein Quaternary Structure Inference. *Structure* **19**, 304-312 (2011).
8. Henrick, K., Thornton, J. M. PQS: A Protein Quaternary Structure File Server. *Trends Biochem. Sci.* **23**, 358-361 (1998).
9. Krissinel, E., Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774-797 (2007).
10. Ponstingl, H., Henrick, K., Thornton, J. M. Discriminating between Homodimeric and Monomeric Proteins in the Crystalline State. *Proteins* **41**, 47-57 (2000).
11. Ponstingl, H., Kabir, T., Thornton, J. M. Discriminating Between Homodimeric and Monomeric Proteins in the Cristalline State. *J. Appl. Crystallogr.* **36**, 1116-1122 (2003).
12. Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., Poupon, A. DiMoVo: A Voronoi Tessellation-Based Method For Discriminating Crystallographic and Biological Protein-Protein Interactions. *Bioinformatics* **24**, 652-658 (2008).
13. Zhu, H., Domingues, F. S., Sommer, I., Lengauer, T. Noxclass: Prediction of Protein-Protein Interaction Types. *BMC Bioinformatics* **7**, 27 (2006).

14. Liu, Q., Kwoh, C. K., Li, J. Binding Affinity Prediction for Protein-Ligand Complexes Based on Beta Contacts and B Factor. *J. Chem. Inf. Model.* **53**, 3076-3085 (2013).
15. Tsuchiya, Y., Kinoshita, K., Naiamura, H. Analyses of Homo-Oligomer Interfaces of Proteins from the Complementarity of Molecular Surface, Electrostatic Potential and Hydrophobicity. *Prot.Eng. Des. Sel.* **19**, 421-429 (2006).
16. Block, P., Paern, J., Hullermeier, E., Sanschagrin, P., Sotriffer, C. A., Klebe, G. Physicochemical Descriptors to Discriminate Protein-Protein Interactions in Permanent and Transient Complexes Selected by Means of Machine Learning Algorithms. *Proteins* **65**, 607-622 (2006).
17. Mintseris, J., Weng, Z. Atomic Contact Vectors in Protein-Protein Recognition. *Proteins* **53**, 629-639 (2003).
18. Valdar, W. S., Thornton, J. M. Protein-Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers. *Proteins* **42**, 108-124 (2001).
19. Elcock, A. H., McCammon, J. A. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2990-2994 (2001).
20. Duarte, J. M., Srebniak, A., Scharer, M. A., Capitani, G. Protein Interface Classification by Evolutionary Analysis. *BMC Bioinformatics* **13**, 334 (2012).
21. Bahadur, R. P., Chakrabarti, P., Rodier, F., Janin, J. A Dissection of Specific and Non-Specific Protein-Protein Interfaces. *J. Mol. Biol.* **336**, 943-955 (2004).
22. Chakrabarti, P., Janin, J. Dissecting Protein-Protein Recognition Sites. *Proteins* **47**, 334-343 (2002).
23. Bourgeas, R., Basse, M. J., Morelli, X., Roche, P. Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2p2i Database. *PloS One* **5**, e9598 (2010).
24. Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., Cheatham, I., T.E., Darden, T. A., Duke, R. E., Gohlke, H., Goetz, A. W., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T. S., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, K. M., Paesani, F., Roe, D. R., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C. L., Smith, W., Swails, J., Walker, R. C., Wang, W., Wolf, R. M., X., W., Kollman, P. A. Amber, version 14, http:// http://ambermd.org/ (accessed Jul 15, 2015)
25. Sanner, M. F., Olson, A. J., Spehner, J. C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **38**, 305-320 (1996).

26. Redundancy in the Protein Data Bank , http://www.rcsb.org/pdb/statistics/clusterStatistics.do (accessed Jul 17, 2015)
27. Rognan, D. Rational Design of Protein-Protein Interaction Inhibitors. *MedChemComm*, **6**, 51-60 (2015).
28. Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M., Janin, J. A Structure-Based Benchmark for Protein-Protein Binding Affinity. *Prot. Sci.* **20**, 482-491 (2011).
29. Gavenonis, J., Sheneman, B. A., Siegert, T. R., Eshelman, M. R., Kritzer, J. A. Comprehensive Analysis of Loops at Protein-Protein Interfaces for Macrocycle Design. *Nat. Chem. Biol.* **10**, 716-722 (2014).
30. Bietz, S., Urbaczek, S., Schulz, B., Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminform.* **6**, 12 (2014).
31. Desaphy, J., Raimbaud, E., Ducrot, P., Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **53**, 623-637 (2013).
32. Desaphy, J., Azdimousa, K., Kellenberger, E., Rognan, D., Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **52**, 2287-2299 (2012).
33. Liaw, A., Wiener, M. Classification and Regression by RandomForest. *R news* **2**, 18-22 (2002).
34. R Development Core Team, R statistical computing, version 3.2.0, http://www.r-project.org/ (accessed Jul 15, 2015)
35. Chang, Y. S., Graves, B., Guerlavais, V., Tovar, C., Packman, K., To, K. H., Olson, K. A., Kesavan, K., Gangurde, P., Mukherjee, A., Baker, T., Darlak, K., Elkin, C., Filipovic, Z., Qureshi, F. Z., Cai, H., Berry, P., Feyfant, E., Shi, X. E., Horstick, J., Annis, D. A., Manning, A. M., Fotouhi, N., Nash, H., Vassilev, L. T., Sawyer, T. K. Stapled Alpha-Helical Peptide Drug Development: A Potent Dual Inhibitor of Mdm2 and Mdmx for P53-Dependent Cancer Therapy. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3445-3454 (2013).
36. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5-32 (2001).
37. Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, S. A. F. T. Data Mining In the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle? *Brief. Bioinform.* **14**, 315-326 (2013).

38. Douguet, D., Chen, H. C., Tovchigrechko, A., Vakser, I. A. DOCKGROUND Resource for Studying Protein-Protein Interfaces. *Bioinformatics* **22**, 2612-2618 (2006).
39. Schwartz, J. C., Zhang, X., Fedorov, A. A., Nathenson, S. G., Almo, S. C. Structural Basis for Co-Stimulation by The Human Ctla-4/B7-2 Complex. *Nature* **410**, 604-608 (2001).
40. Lou, X., Bao, R., Zhou, C. Z., Chen, Y. Structure of the Thioredoxin-Fold Domain of Human Phosducin-Like Protein 2. *Acta Cryst.Sect. F*, **65**, 67-70 (2009).
41. Romero, A., De la Cerda, B., Varela, P. F., Navarro, J. A., Hervas, M., De la Rosa, M. A. The 2.15 Å Crystal Structure of a Triple Mutant Plastocyanin from the Cyanobacterium Synechocystis sp. pcc 6803. *J. Mol. Biol.* **275**, 327-336 (1998).

2.8. Conclusion globale

Dans ce chapitre, nous avons développé un nouvel outil (IChemPIC) de caractérisation de PPIs à partir de fichiers PDB. Cette méthode basée sur une caractérisation topologique des interfaces entre les protéines permet la détermination des interfaces biologiquement viables de celles ne présentant pas les caractéristiques pour être stables in vivo.

L'avantage de nos travaux réside dans le fait qu'IChemPIC est la seule méthode réalisant des prédictions stables sur les interfaces de toutes les tailles comprises entre 200 et 4000Å^2. Les prédictions sont d'une précision supérieure à 80% et également réparties sur les différentes tailles d'interface ainsi que sur les interfaces pertinentes ou non. Les descripteurs topologiques sont un moyen rapide et efficace de discriminer les interfaces sans réaliser de lourds calculs d'énergie.

La stabilité de l'outil a été testée à l'aide de moyens non décrits dans l'article, nous avons notamment réalisé des simulations par dynamique moléculaire de complexes prédits ou non biologiquement viable. La prédiction sur les complexes prédits non viables ne change pas au cours de la dynamique, cela montre que l'interaction n'est pas stable même après une minimisation et une courte période de dynamique moléculaire (10 ns). Lorsque que l'on regarde des dynamiques de complexes prédits biologiquement viables, on observe qu'entre 30 et 40% des structures sauvegardées sont prédites viables; donc que la majorité du temps le complexe est prédit non pertinent. Le modèle d'apprentissage n'a en effet été entrainé que sur des structures issues de diffraction des rayons X dans des cristaux. Les résultats des simulations de dynamique moléculaire nous poussent à croire que le modèle a besoin de structures très compactes (cristallines) pour les prédire biologiquement viable. Les poses de la dynamique où les protéines sont plus relâchées ne sont pas prédite pertinentes. Pour la suite de nos études issues de dynamique moléculaire, la simple présence de poses prédites viable suffit à décrire une interface dont la structure cristalline n'est pas connue (modèles, docking protéine-protéine). En parallèle de cette étude nous avons aussi réalisé des dynamiques sous contraintes d'interface protéine-protéine, en forçant un des partenaires à se dissocier progressivment. Les résultats montrent qu'IChemPIC est très sensible à la distance entre les deux protéines. Une déviation atour d'un 1Å de la position des atomes lourds à l'interface suffit à faire passer la prédiction de pertinente à non pertinente dans 100% des poses.

IChemPIC est à l'heure actuelle l'outil le plus fiable pour prédire la véracité biologique d'une interface entre deux protéines, cependant il a été principalement entrainé pour étudier les interfaces entre des protéines globulaires; les résultats sur des protéines membranaires tels que des récepteurs sont prometteurs mais l'interprétation est difficile. Il reste difficile d'obtenir des structures de protéines membranaires dont on est absolument certain du positionnement de la zone d'interaction. La dimérisation de certaines familles de protéines dont les structures sont connues (ex: récepteurs couplés aux protéines G) font débat au sein de la communauté scientifique. Les protéines membranaires ont une surface d'interaction trop différente pour être étudiée avec le même modèle d'apprentissage que les protéines globulaires; un nouveau modèle d'apprentissage pourra être dans le futur dédié aux protéines membranaires dès lors que suffisamment de structures cristallographiques de degré d'oligomérisation incontestable seront disponibles.

Le modèle a une seconde faille au niveau des unités biologiques plus complexes que des dimères. IChemPIC ne prédit pas la structure quaternaire d'une unité biologique mais simplement une interaction entre deux chaines protéiques. Il arrive, au sein de structures de grandes tailles, qu'il y ait des incohérences au sein des prédictions; certaines interfaces étant prédites comme pertinentes tandis que d'autres ne le sont pas. Les études de ces complexes supramoléculaires méritent qu'on y attache de l'importance. Nous avons plusieurs hypothèses à vérifier. La première est de voir s'il est possible de ne plus prédire des dimères mais des unités biologiques par cette même méthode de description topologique. La deuxième est d'observer plus attentivement les unités biologiques dont toutes les prédictions d'interfaces ne vont pas dans la même direction. Il est rare qu'un trimère ou plus s'assemble de manière spontanée en une seule étape. On privilégie le passage par des degrés d'oligomérisation inférieurs comme le dimère. Une structure tétramérique aura tendance à être décrite comme l'assemblage de deux dimères, de ce fait les interactions entre les protéines qui ne font pas parties des dimères originaux peuvent-elles exister sans la présence des autres partenaires ? Une interaction entre deux protéines peut-elle nécessiter la présence d'une troisième protéine pour être stable ? Savoir si IChemPIC peut prédire le degré d'oligomérisation des protéines lors de l'assemblage de structures quaternaires est une bonne perspective à cette étude

2.9. Annexes

Supplementary Table 2.1. jeu de 200 contacts cristallins (FDS dataset)

PDB ID	Protein name	Chain 1	Chain 2	Interface, Å2	Set
1A12	Regulator Of Chromosome Condensation 1	A	C	787	Training
1A3Y	Odorant Binding Protein	A	B	798	Training
1A6Q	Phosphatase 2c	A	E	402	Training
1A7T	Metallo-Beta-Lactamase	A	R	424	Training
1A7V	Cytochrome C'	A	B	536	Training
1AD5	Haematopoetic Cell Kinase Hck	A	B	1073	Training
1AE9	Lambda Integrase	A	B	509	Training
1AF7	Chemotaxis Receptor Methyltransferase Cher	A	L	1131	Training
1AH7	Phospholipase C	A	G	970	Training
1AJK	Circularly Permuted (1-3,1-4)-Beta-D-Glucan 4-Glucanohydrolase	A	Q	535	Training
1AKO	Exonuclease Iii	A	D	968	Training
1AKZ	Uracil-Dna Glycosylase	A	B	583	Training
1AMU	Gramicidin Synthetase 1	A	B	524	Training
1AQZ	Restrictocin	A	B	215	Training
1ATL	Atrolysin C	A	B	396	Training
1AW7	Toxic Shock Syndrome Toxin-1	A	B	1147	Training
1AYI	Colicin E Immunity Protein 7	A	B	483	Training
1AYL	Phosphoenolpyruvate Carboxykinase	A	E	518	Training
1B24	Protein (I-Dmoi)	A	D	1053	Training
1B3J	Mhc Class I Homolog Mic-A	A	B	1645	Training
1B80	Protein (Recombinant Lignin Peroxidase H8)	A	N	589	Training
1B8X	Protein (Aml-1b)	A	B	1319	Training
1BE0	Haloalkane Dehalogenase	A	G	617	Training

1BG0	Arginine Kinase	A	H	596	Training
1BGC	Granulocyte Colony-Stimulating Factor	A	B	713	Training
1BIN	Leghemoglobin A	A	B	573	Training
1BKZ	Galectin-7	A	B	749	Training
1CFY	Cofilin	A	B	370	Training
1CKI	Casein Kinase I Delta	A	B	936	Training
1CLU	Transforming Protein P21/H-Ras-1	A	D	761	Training
1CQX	Flavohaeomglobin	A	B	1110	Training
1D3H	Dihydroorotate Dehydrogenase	A	B	1483	Training
1DXM	H Protein	A	B	575	Training
1DYS	Endoglucanase	A	B	770	Training
1EHY	Protein (Soluble Epoxide Hydrolase)	A	C	700	Training
1EJD	Mura	A	B	1226	Training
1ELP	Gamma-D Crystallin	k	B	517	Training
1EWF	Bactericidal/Permeability-Increasing Protein	A	I	952	Training
1FFR	Chitinase A Mutant Y390f	A	B	1226	Training
1FJM	Protein Serine/Threonine Phosphatase-1 (Alpha Isoform, Type 1)	A	B	416	Training
1FKD	Fk506 Binding Protein	A	B	647	Training
1FMT	Methionyl-Trna Fmet Formyltransferase	A	B	575	Training
1FPO	Hsc20	A	B	1087	Training
1FSU	4-Sulfatase	A	B	527	Training
1FVK	Disulfide Bond Formation Protein	A	B	759	Training
1G6A	Pse-4 Carbenicillinase, R234k Mutant	A	B	1020	Training
1GAR	Glycinamide Ribonucleotide Transformylase	A	B	788	Training
1GPI	Cellobiohydrolase Cel7d	A	B	1042	Training
1HRN	Renin	A	B	734	Training
1HSL	Histidine-Binding Protein	A	B	413	Training
1I4G	Enterotoxin Type A	A	B	758	Training
1ILR	Interleukin-1 Receptor Antagonist Protein	A	B	335	Training

1J96	3alpha-Hsd Type 3	A	B	1218	Training
1JFR	Lipase	A	S	475	Training
1KFS	Protein (Dna Polymerase I Klenow Fragment (E.C.2.7.7.7))	A	B	1329	Training
1KPT	Kp4 Toxin	A	B	512	Training
1KWA	Hcask/Lin-2 Protein	A	B	584	Training
1LQT	Fpra	A	B	1004	Training
1MPG	3-Methyladenine Dna Glycosylase Ii	A	B	364	Training
1MSS	Triosephosphate Isomerase	A	B	1059	Training
1N4G	Cyp121	A	B	1162	Training
1NDB	Carnitine Acetyltransferase	A	B	1390	Training
1NMT	N-Myristoyl Transferase	A	B	894	Training
1NP4	Protein (Nitrophorin 4)	A	B	732	Training
1OME	Beta-Lactamase	A	B	1071	Training
1OVW	Endoglucanase I	A	i	734	Training
1PP3	Thaumatin	A	B	1110	Training
1PPO	Protease Omega	A	B	854	Training
1PVA	Parvalbumin	A	B	442	Training
1QAZ	Protein (Alginate Lyase A1-Iii)	A	B	810	Training
1QCI	Pokeweed Antiviral Protein	A	B	512	Training
1QHA	Protein (Hexokinase)	A	B	1734	Training
1QNT	Methylated-Dna--Protein-Cysteine Methyltransferase	A	j	656	Training
1QTQ	Protein (Glutaminyl-Trna Synthetase)	A	B	728	Training
1RB3	Dihydrofolate Reductase	A	B	622	Training
1RGE	Ribonuclease	B	i	554	Training
1S83	Porcine Trypsin	A	B	935	Training
1SHK	Shikimate Kinase	A	M	542	Training
1SO7	Cytosolic Sialidase Neu2	A	B	1190	Training
1SW6	Regulatory Protein Swi6	A	B	685	Training
1THT	Thioesterase	A	B	469	Training

1TOA	Protein (Periplasmic Binding Protein Troa)	A	B	1316	Training
1TON	Tonin	A	H	707	Training
1VBT	Cyclophilin A	A	f	658	Training
1VJW	Ferredoxin(A)	A	C	478	Training
1VQQ	Penicillin-Binding Protein 2a	A	B	1217	Training
1W9Q	Pdz Of Human Syntenin	A	B	1004	Training
1WLY	2-Haloacrylate Reductase	A	B	1144	Training
1WOQ	Polyphosphate/Atp-Glucomannokinase	A	B	1129	Training
1XGK	N12g,A18g Nmra Mutant	A	B	1020	Training
1XGS	Methionine Aminopeptidase	A	B	1182	Training
1YGH	Protein (Transcriptional Activator Gcn5)	A	C	667	Training
1ZIN	Adenylate Kinase	A	J	564	Training
1ZLQ	Periplasmic Transporter Nika	A	B	1680	Training
256B	Cytochrome B562	A	B	315	Training
256L	Lysozyme	A	E	775	Training
2ACY	Acylphosphatase	A	H	453	Training
2BC2	Metallo Beta-Lactamase Ii	A	z	357	Training
2E1V	Dmat	A	B	1655	Training
2ERC	Rrna Methyl Transferase	A	B	434	Training
2EYI	Alpha-Actinin 1	A	B	1115	Training
2F37	Human Trpv2	A	B	1164	Training
2FGF	Human Basic Fibroblast Growth Factor	A	B	400	Training
2G3P	Infectivity Protein G3p	B	M	451	Training
2GAS	Isoflavone Reductase	A	B	1570	Training
2H44	Pde5a1	A	B	1043	Training
2IHL	Japanese Quail Egg White Lysozyme	A	E	400	Training
2IPI	Aclacinomycin Oxidoreductase	A	B	1330	Training
2J0P	Haem-Chaperone Protobacteria-Protein Hems	A	B	1017	Training
2J46	Ffh Ng Domain	B	C	1168	Training

2MBR	Uridine Diphospho-N-Acetylenolpyruvylglucosamine Reductase	A	B	766	Training
2Q7D	Human Inositol 1,3,4-Trisphosphate 5/6-Kinase	A	B	1784	Training
2RN2	Ribonuclease H	A	H	617	Training
2SCP	Sarcoplasmic Calcium-Binding Protein	A	B	1131	Training
2SHP	Shp-2	B	S	539	Training
2W20	Nana Sialidase	A	B	1103	Training
2WBF	Sera5e	X	Y	1236	Training
2WBM	Mthsbds	B	C	1443	Training
2WBQ	Vioc	A	B	1517	Training
2WSA	N-Myristoyl Transferase	A	B	1007	Training
2X26	Sulfonate Binding Protein Ssua	A	B	1377	Training
2YVW	Udp-N-Acetylglucosamine 1-Carboxyvinyl-Transferase	A	B	1520	Training
2YZ1	Murine Shps-1/Sirp	A	B	1378	Training
2Z6O	Ufm1 Cinhugating Enzyme 1	A	B	1158	Training
2ZYR	Lipase	A	B	1011	Training
3AAP	Lp1ntpdase	A	B	1387	Training
3B37	Aminopeptidase N	A	B	1149	Training
3C1D	Recx	A	B	1430	Training
3C8Y	Fe-Only Hydrogenase	A	B	1527	Training
3CJ1	Ntpdase 2	A	B	1126	Training
3CMS	Chymosin B	A	H	562	Training
3CU9	1,5-Alpha-L-Arabinanase	A	B	1086	Training
3ELS	Yeast Pml1p	A	B	1005	Training
3F0O	Merb	A	B	1043	Training
3GKJ	Npc1d(Ntd)	A	B	1014	Training
3GVO	Mouse Pumilio Puf-2 Domai?	A	B	1222	Training
3H30	Protein Kinase Ck2	A	B	1331	Training
3HZL	Nikd Aminooxidase	A	B	1402	Training
3IRB	Uncharacterized Protein From Duf35 Family	A	B	1060	Training

3KH7	Reduced Ccmg	A	B	1338	Training
3KK8	Camkii	A	B	1291	Training
3LVD	Gfp-Like Protein Acegfp-G222e	A	B	1005	Training
3M66	Mitochondrial Transcription Termination Factor3	A	B	1012	Training
3MHJ	Tankyrase-2	A	B	1102	Training
3MHT	Protein (Hhai Methyltransferase (E.C.2.1.1.73))	A	B	248	Training
3MHZ	2-Fluorohistidine Labeled Protective Antigen	A	B	1537	Training
3N5C	Arf6delta13	A	B	1458	Training
3PMG	Alpha-D-Glucose-1,6-Bisphosphate	A	B	540	Training
830C	Mmp-13	A	B	1236	Training
8PTI	Bovine Pancreatic Trypsin Inhibitor	A	R	700	Training
13PK	3-Phosphoglycerate Kinase	A	D	1393	Test
1AFK	Ribonuclease A	A	B	291	Test
1AQ0	1,3-1,4-Beta-Glucanase	A	B	722	Test
1B1J	Hydrolase Angiogenin	A	C	892	Test
1BEA	Bifunctional Amylase/Serine Protease Inhibitor	A	G	524	Test
1BF6	Phosphotriesterase Homology Protein	B	D	509	Test
1BS2	Protein (Arginyl-Trna Synthetase)	A	G	1210	Test
1BYO	Protein (Plastocyanin)	A	T	298	Test
1C02	Phosphotransferase Ypd1p	A	B	873	Test
1CK7	Protein (Gelatinase A)	A	B	953	Test
1DSU	Factor D	A	C	728	Test
1DZ4	Cytochrome P450-Cam	A	U	542	Test
1EPA	Epididymal Retinoic Acid-Binding Protein	A	B	642	Test
1FGK	Fgf Receptor 1	A	C	817	Test
1G2A	Polypeptide Deformylase	A	B	577	Test
1HF8	Clathrin Assembly Protein Short Form	A	O	538	Test
1IHB	Cyclin-Dependent Kinase 6 Inhibitor	A	B	329	Test
1INP	Inositol Polyphosphate 1-Phosphatase	A	C	798	Test

1LF2	Plasmepsin Ii	A	B	2168	Test
1LXK	Hyaluronate Lyase	A	B	1036	Test
1N45	Heme Oxygenase-1	A	B	1098	Test
1NUC	Staphylococcal Nuclease	A	S	606	Test
1PBG	6-Phospho-Beta-D-Galactosidase	A	R	271	Test
1PDA	Porphobilinogen Deaminase	A	B	885	Test
1PMI	Phosphomannose Isomerase	A	B	628	Test
1QJP	Outer Membrane Protein A	A	V	586	Test
1QME	Penicillin-Binding Protein 2x	A	B	647	Test
1RHS	Sulfur-Substituted Rhodanese	A	B	863	Test
1THE	Cathepsin B	A	B	729	Test
1UEB	Elongation Factor P	A	B	1150	Test
1URP	D-Ribose-Binding Protein	A	i	805	Test
1VLZ	Chey	A	B	435	Test
1XCA	Cellular Retinoic Acid Binding Protein Type Ii	A	B	617	Test
1YNQ	Aldo-Ketoreductase Akr11c21	A	B	1103	Test
2ATJ	Peroxidase C1a	A	G	932	Test
2BLS	Ampc Beta-Lactamase	B	E	610	Test
2CKI	Ulilysin	A	B	1185	Test
2END	Endonuclease V	A	D	610	Test
2EQA	Hypothetical Su5a Protein	A	B	1110	Test
2FGZ	Apo Pullulanase	A	B	1094	Test
2HLQ	Type Ii Bmp Receptor	A	B	1087	Test
2NAP	Protein (Periplasmic Nitrate Reductase)	b	p	747	Test
2TPS	Protein (Thiamin Phosphate Synthase)	A	B	809	Test
2UGI	Uracil-Dna Glycosylase Inhibitor	A	B	741	Test
2VT4	Beta1 Adrenergic Receptor	A	B	681	Test
2XOV	Rhomboid Protease Glpg	A	B	1200	Test
3FWK	Fmn Adenyltransferase	A	B	1048	Test

3GO5	Multidomain Protein With S1 Rna-Binding Domain	A	B	1141	Test
3ITA	Penicillin-Binding Protein Pbp6	A	B	1142	Test
3MG1	Orange Carotenoid Protein	A	B	1099	Test

Supplementary Table 2.2. Jeu de 200 interfaces biologiquement pertinentes (FDS data

PDB ID	Protein 1	Protein 2	Chain 1	Chain2	Interface, Å2	Selection	Set
1A14	Neuraminidase	Nc10 Fv	N	H	418	IPACdb	Training
1A2K	Ran Gtpase-Gdp	Nuclear Transport Factor 2	C	B	820	PPI affinity DB	Training
1AUT	Activated Protein C	Activated Protein C	C	L	882	IPACdb	Training
1AVZ	Hiv-1-Nef Protein	Fyn Kinase Sh3 Domain	B	C	654	PPI affinity DB	Training
1AY7	Rnase	Barstar	A	B	646	PPI affinity DB	Training
1BT6	Annexin A2	S100A10	B	C	505	Rognan et al., MedChemComm, 2014	Training
1BVN	Alpha-Amylase	Tendamistat	P	T	1133	PPI affinity DB	Training
1BXL	Bcl2	Bax	A	B	937	2P2I	Training
1C1Y	Rap-1a	Protein Kinase Raf-1	A	B	690	IPACdb	Training
1CJU	Ac Ii	Gs	B	C	820	IPACdb	Training
1DS2	Proteinase B	OMTKY3	E	I	582	IPACdb	Training
1E4K	Fc Fragment Of Human Igg 1	Fc Fragment Of Human Igg 1	A	B	1400	PPI affinity DB	Training
1EER	Erythropoietin	EPO Receptor	A	B	1032	PPI affinity DB	Training
1EM8	Dna Polymerase Iii Chi Subunit	Dna Polymerase Iii Psi Subunit	C	D	721	IPACdb	Training
1EMV	Colicin E9 Nuclease	Im9 Immunity Protein	A	B	800	PPI affinity DB	Training
1FQJ	Gt-Alpha	Rgs9	A	B	960	PPI affinity DB	Training
1GCQ	Grb2 C-Ter Sh3 Domain	Vav N-Ter Sh3 Domain	B	C	629	PPI affinity DB	Training
1GL1	Chymotrypsin	PMP-C (LCMI II)	A	I	857	PPI affinity DB	Training
1GLA	Glycerol Kinase	Glucose Specific Iiiglc	G	F	695	PPI affinity DB	Training
1GNG	GSK-3beta	GSK-3 Bp	A	X	761	Ivanov et al., TIPS, 2013	Training
1GPW	Hisf Protein	Amidotransferase Hish	A	B	1138	PPI affinity DB	Training
1GRN	Cdc42 Gtpase	Cdc42 Gap	A	B	1215	PPI affinity DB	Training
1H9D	Runx1 Domain Of Cbfalpha1	Dimerisation Domain Of Cbf-Beta	A	B	1095	PPI affinity DB	Training
1HWG	Ghbp	GH	B	C	485	LoopFinder	Training
1I5H	Ppxy	Nedd4	W	B	431	Rognan et al., MedChemComm, 2014	Training

1IXS	Ruva	Ruvb	A	B	699	IPACdb	Training
1J2J	Arf1 Gtpase.Gnp-Ranbd1	Gat Domain Of Gga1	A	B	643	PPI affinity DB	Training
1JDH	Beta Catenin	Htcf4	A	B	1777	Ivanov et al., TIPS, 2013	Training
1JIW	Alkaline Metallo-Proteinase	Proteinase Inhibitor	P	I	1057	PPI affinity DB	Training
1JQ5	Glycerol Dehydrogenase	Glycerol Dehydrogenase	A	B	1366	Dcbio dataset	Training
1JSU	Cyclin A	Cdk2	B	A	1798	Ivanov et al., TIPS, 2013	Training
1JU3	Cocaine Esterase	Cocaine Esterase	A	B	1338	Dcbio dataset	Training
1KAC	Adenovirus Fiber Knob Protein	Adenovirus Receptor	A	B	793	PPI affinity DB	Training
1KGY	Ephb2	Ephrin-B2	B	F	1233	LoopFinder	Training
1KKL	Hpr Kinase	Hpr	A	H	535	PPI affinity DB	Training
1KQ3	Glycerol Dehydrogenase	Glycerol Dehydrogenase	A	B	1549	Dcbio dataset	Training
1KTZ	Tgf-Beta	Tgf-Beta Receptor	A	B	522	PPI affinity DB	Training
1LW4	L-Allo-Threonine Aldolase	L-Allo-Threonine Aldolase	A	D	1736	Dcbio dataset	Training
1LZL	Bacterial Heroin Esterase	Bacterial Heroin Esterase	A	B	1249	Dcbio dataset	Training
1M10	Von Willebrand Factor Dom. A1	Glycoprotein Ib-Alpha	A	B	1142	PPI affinity DB	Training
1MAH	Acetylcholinesterase	Fasciculin	A	F	1104	PPI affinity DB	Training
1MQ8	Icam-1 Domain 1-2	Integrin Alpha-L I Domain	A	B	656	PPI affinity DB	Training
1MZW	U-Snrnp-Associated Cyclophilin	U4/U6 Snrnp 60kda Protein	A	B	466	IPACdb	Training
1N0W	Brca2	Rad51	B	A	1072	IPACdb	Training
1N8P	Cystathionine Gamma-Lyase	Cystathionine Gamma-Lyase	C	D	2089	Dcbio dataset	Training
1O17	Anthranilate Phosphoribosyltransferase	Anthranilate Phosphoribosyltransferase	A	D	963	Dcbio dataset	Training
1OFU	Cell Division Protein Ftsz	Sula Pa3008	A	X	812	IPACdb	Training
1P5F	Dj-1	Dj-1	A	B	1377	Dcbio dataset	Training
1P9M	Gp130	Il-6	A	B	728	Rognan et al., MedChemComm, 2014	Training
1PKH	Deaminase/Diphosphatase	Deaminase/Diphosphatase	A	B	1731	Dcbio dataset	Training
1PVH	Il6 Receptor Beta Chain D2-D3 Domains	Leukemia Inhibitory Factor	A	B	737	PPI affinity DB	Training
1QOP	Trp Synthase	Trp Synthase	B	C	1606	Dcbio dataset	Training
1R5Y	Queuine Trna-	Queuine Trna-Ribosyltransferase	A	B	1619	Dcbio dataset	Training

	Ribosyltransferase						
1R6Q	Clp Protease Subunit Clpa	Clp Protease Adaptor Protein Clps	A	C	1304	PPI affinity DB	Training
1S2Z	Rubrerythrin	Rubrerythrin	A	B	1677	Dcbio dataset	Training
1SML	Metallo Betalactamase L1	Metallo Betalactamase L1	A	B	1327	Dcbio dataset	Training
1SV0	Yan	Mae	A	C	504	LoopFinder	Training
1TGS	Trypsinogen	Psti	Z	I	906	IPACdb	Training
1TNR	Tnfr1	Tnfr1	A	R	673	2P2I	Training
1US7	Heat Shock Protein 82 N-Ter Domain	Hsp90 Co-Chaperone Cdc37 C-Ter Domain	A	B	576	PPI affinity DB	Training
1UUG	Uracyl-Dna Glycosylase	Glycosylase Inhibitor	A	B	1097	PPI affinity DB	Training
1UZ3	Emsy Protein	Emsy Protein	A	B	1254	Dcbio dataset	Training
1V2X	Trna (Gm18) Methyltransferase	Trna (Gm18) Methyltransferase	A	B	1751	Dcbio dataset	Training
1X7V	Pa3566 Protein	Pa3566 Protein	A	B	1260	Dcbio dataset	Training
1XD3	Uch-L3	Ubiquitin	A	B	1177	PPI affinity DB	Training
1XDT	Hbegf	Dipheria Toxin	R	T	1129	LoopFinder	Training
1XQS	Hspbp1	Hsp70 Atpase Domain	A	C	1240	PPI affinity DB	Training
1YCR	Hdm2	P53	A	B	751	2P2I	Training
1YFN	Sspb	Rsea	A	E	875	LoopFinder	Training
1ZE3	Chaperone Protein Fimc	Chaperone Protein Fimh	C	H	1521	Dcbio dataset	Training
1ZM4	Elongation Factor 2	Diphtheria Toxin A Catalytic Domain	A	B	847	PPI affinity DB	Training
2A5L	Trp Repressor Binding Protein Wrba	Trp Repressor Binding Protein Wrba	A	B	1305	Dcbio dataset	Training
2A9K	Ral-A.GDP	Mono-ADP-Ribosyltransferase C3	A	B	917	PPI affinity DB	Training
2AJF	Angiotensin-Converting Enzyme 2	Sars Spike Protein Receptor Binding Domain	A	E	913	PPI affinity DB	Training
2AQ3	Tcr Vbeta8.2	Sec3	A	B	599	PPI affinity DB	Training
2AQ6	Pyridoxine 5'-Phosphate Oxidase	Pyridoxine 5'-Phosphate Oxidase	A	B	1250	Dcbio dataset	Training
2B42	Xylanase	Xylanase Inhibitor	A	B	1331	PPI affinity DB	Training
2B4J	Integrase	LEDGF	A	B	1489	2P2I	Training

2B59	Chromosome Segregation Atpase	Cellulosomal Scaffolding Protein A	A	B	864	IPACdb	Training
2BHS	Cysteine Synthase B	Cysteine Synthase B	C	D	1570	Dcbio dataset	Training
2C0L	Trp Region Of Pex5	Sterol Carrier Protein 2	A	B	1043	PPI affinity DB	Training
2C4W	3-Dehydroquinate Dehydratase	3-Dehydroquinate Dehydratase	A	B	918	Dcbio dataset	Training
2CFT	Pyridoxal Phosphate Phosphatase	Pyridoxal Phosphate Phosphatase	A	B	976	Dcbio dataset	Training
2D9Q	Csf3 Cytokine	Colony-Stimulating Factor Receptor	A	B	734	IPACdb	Training
2E7J	Sep-Trna:Cys-Trna Synthase	Sep-Trna:Cys-Trna Synthase	A	B	1742	Dcbio dataset	Training
2ECQ	O-Acetylserine (Thiol)-Lyase	O-Acetylserine (Thiol)-Lyase	A	B	1684	Dcbio dataset	Training
2EXB	Penicillin-Binding Protein 4	Penicillin-Binding Protein 4	A	B	2012	Dcbio dataset	Training
2F9D	P14	Sf3b1	A	B	512	IPACdb	Training
2FWV	Hypothetical Protein Mtubf_01000852	Hypothetical Protein Mtubf_01000852	A	B	2171	Dcbio dataset	Training
2G2U	Beta-Lactamase Shv-1	Beta-Lactamase Inhibitory Protein	A	B	1279	Dcbio dataset	Training
2GOX	Complement C3d Fragment	Staphylococcus Aureus Efb-C	A	B	885	PPI affinity DB	Training
2HDI	Colicin 1 Receptor	Colicin -1a	A	B	1013	LoopFinder	Training
2HQS	Tolb	Pal	A	H	1219	PPI affinity DB	Training
2HRK	Glutamyl-T-Rna Synthetase	Gu-4 Nucleic Binding Protein	A	B	837	PPI affinity DB	Training
2HYM	Ifna	IFNAR2	A	B	860	Rognan et al., MedChemComm, 2014	Training
2J0T	Mmp1 Intersitial Collagenase	Metalloproteinase Inhibitor 1	A	D	795	PPI affinity DB	Training
2J3T	Trappc1	Trappc4	C	D	1885	LoopFinder	Training
2MTA	Methylamine Dehydrogenase	Amicyanin	H	L	1692	PPI affinity DB	Training
2NZL	Hydroxyacid Oxidase 1	Hydroxyacid Oxidase 1	A	B	1356	Dcbio dataset	Training
2O3B	Nuca Nuclease	Nuia Nuclease Inhibitor	A	B	820	PPI affinity DB	Training
2PCB	Cyt C Peroxidase	Cytochrome C	A	B	560	PPI affinity DB	Training
2PEH	Spf45	Sf3b1	B	D	477	LoopFinder	Training
2RKB	Serine Dehydratase-Like	Serine Dehydratase-Like	E	F	1165	Dcbio dataset	Training
2VDB	Serum Albumin	Peptostreptococcalalbumin-Binding Protein	A	B	946	PPI affinity DB	Training

2VEF	Dihydropterate Synthase	Dihydropterate Synthase	A	B	1599	Dcbio dataset	Training
2VR4	Beta Mannisodase	Beta Mannisodase	A	B	1352	Dcbio dataset	Training
2W6A	Arf Gtpase-Activating Protein Git1	Arf Gtpase-Activating Protein Git1	A	B	1445	Dcbio dataset	Training
2WG3	Hedgehog	HIP	A	C	979	LoopFinder	Training
2WTM	Est1E	Est1E	A	B	1336	Dcbio dataset	Training
2WXD	Myrosinase	Myrosinase	M	N	1002	Dcbio dataset	Training
2XAO	Bxlxl	Bak	B	A	1232	2P2I	Training
2Y0N	Msl3	Msl1	A	E	1147	LoopFinder	Training
2Y27	Phenylacetate-Coenzyme A Ligase	Phenylacetate-Coenzyme A Ligase	A	B	1460	Dcbio dataset	Training
2Y39	Protein CNRR	Protein CNRR	A	B	1343	Dcbio dataset	Training
2Z30	Tk-Subtilisin	Tk-Propeptide	A	B	1220	LoopFinder	Training
3A2Q	6-Aminohexanoate-Cyclic-Dimer Hydrolase	6-Aminohexanoate-Cyclic-Dimer Hydrolase	A	B	1802	Dcbio dataset	Training
3BDW	Cd94	Ngk2a	A	B	808	LoopFinder	Training
3BIM	Bcl6	Bcor	A	B	2042	IPACdb	Training
3BP8	Mlc Transcription Regulator	Pts Glucose-Specific Enzyme Eiicb	A	C	726	PPI affinity DB	Training
3BT1	Upar	Vitronectin	U	B	653	LoopFinder	Training
3BX7	Lipocalin 2	CTLA 4	A	C	1254	IPACdb	Training
3CKI	Tace	Timp-3	B	A	1108	LoopFinder	Training
3D36	Sporulation Kinase B	Dual Specificity Protein Phosphatase	A	B	1768	Dcbio dataset	Training
3DA8	Probable 5'-Phosphoribosylglycinamide Formyltransferase Purn	Probable 5'-Phosphoribosylglycinamide Formyltransferase Purn	A	B	964	Dcbio dataset	Training
3E1Z	Peptidase C1A	Papain	B	A	890	IPACdb	Training
3EPW	Iag-Nucleoside Hydrolase	Iag-Nucleoside Hydrolase	A	B	1360	Dcbio dataset	Training
3FAH	Aldehyde Oxidoreductase	Aldehyde Oxidoreductase	A	B	1203	Dcbio dataset	Training
3GUS	Glutathione S-Transferase P	Glutathione S-Transferase P	A	B	1243	Dcbio dataset	Training

3H6D	Deoxyuridine 5'-Triphosphate Nucleotidohydrolase	Deoxyuridine 5'-Triphosphate Nucleotidohydrolase	A	B	1724	Dcbio dataset	Training
3ITF	Periplasmic Adaptor Protein Cpxp	Periplasmic Adaptor Protein Cpxp	A	B	1426	Dcbio dataset	Training
3IUE	Pantothenate Synthetase	Pantothenate Synthetase	A	B	1184	Dcbio dataset	Training
3JYO	Quinate/Shikimate Dehydrogenase	Quinate/Shikimate Dehydrogenase	A	B	1256	Dcbio dataset	Training
3KD2	Cftr Inhibitory Factor (Cif)	Cftr Inhibitory Factor (Cif)	C	D	1395	Dcbio dataset	Training
3KNB	Titin	Obscurin-Like 1	A	B	701	IPACdb	Training
3KYS	Yap	Tef-1	B	A	1823	Ivanov et al., TIPS, 2013	Training
3LW6	Beta-4-Galactosyltransferase 7	Beta-4-Galactosyltransferase 7	A	B	1034	Dcbio dataset	Training
3O1N	3-Dehydroquinate Dehydratase	3-Dehydroquinate Dehydratase	A	B	1068	Dcbio dataset	Training
3OVP	Ribulose-Phosphate 3-Epimerase	Ribulose-Phosphate 3-Epimerase	A	B	1078	Dcbio dataset	Training
3QS7	Flt3	Fl	E	A	516	IPACdb	Training
3R0N	Poliovirus Receptor-Related Protein 2	Poliovirus Receptor-Related Protein 2	A	B	946	Dcbio dataset	Training
3V2A	Vegf	Vegfr2	A	R	705	IPACdb	Training
4AXG	Eif4e	Cup	B	D	502	Ivanov et al., TIPS, 2013	Training
4FJ3	03/03/14	Raf	A	P	510	Ivanov et al., TIPS, 2013	Training
4GQ6	Menin	MLL	A	B	835	2P2I	Training
4HQP	A7 Nicotinic Receptor	A-Bungarotoxin	A	I	777	LoopFinder	Training
4JOI	Snt1	Ten-1	A	D	1195	LoopFinder	Training
4MC1	Hiv-1 Pr	Hiv-A Pr	A	B	1845	IPACdb	Training
4NN6	Interleukin-7 Receptor Subunit Alpha	Cytokine Receptor-Like Factor 2	B	C	479	IPACdb	Training
4OJK	Rab11b	Cgmp-Dependent Protein Kinase 2	A	C	840	IPACdb	Training
4P78	Hic3b	Hic3a	A	D	1391	IPACdb	Training
1BS1	Dethiobiotin Synthetase	Dethiobiotin Synthetase	A	B	1548	Dcbio dataset	Test
1DFJ	Ribonuclease A	Rnase Inhibitor	E	I	1391	PPI affinity DB	Test

1EEJ	S-S Isomerase	S-S Isomerase	A	B	885	Dcbio dataset	Test
1EWY	Ferredoxin Reductase	Ferredoxin	A	C	822	PPI affinity DB	Test
1F2D	Carboxylate Deaminase	Carboxylate Deaminase	C	D	1957	Dcbio dataset	Test
1F34	Porcine Pepsin	Ascaris Inhibitor 3	A	B	1593	PPI affinity DB	Test
1F47	Zipa	Ftsz	A	B	601	2P2I	Test
1FFW	Chemotaxis Protein Chey	Chemotaxis Protein Chea	A	B	623	PPI affinity DB	Test
1FLT	Vegf	Flt1	W	X	627	Ivanov et al., TIPS, 2013	Test
1GXD	Prommp2 Type Iv Collagenase	Metalloproteinase Inhibitor 2	A	C	1306	PPI affinity DB	Test
1HCF	Neurotrophin-4	Neurotrophin-4	A	B	1811	PPI affinity DB	Test
1KO6	Nup98	Nup98	A	B	468	IPACdb	Test
1M9C	Cyclophilin A	Gag P24	A	D	511	IPACdb	Test
1NW9	Xiap	Casp-9	A	B	1111	2P2I	Test
1PXV	Saphopain	Sapostatin	A	C	1219	PPI affinity DB	Test
1QA9	Cd2	Cd58	A	B	758	PPI affinity DB	Test
1TUE	Hpv2	Hpv11	A	B	991	2P2I	Test
1UJ6	Ribose 5-Phosphate Isomerase	Ribose 5-Phosphate Isomerase	A	B	1338	Dcbio dataset	Test
1VK5	Rddm	Rddm	A	B	1194	Dcbio dataset	Test
1VKX	P50	P60	B	A	688	Ivanov et al., TIPS, 2013	Test
1X2R	Keap1	Nrf2	A	B	546	Rognan et al., MedChemComm, 2014	Test
1XX9	Fx1a	Ecotin	A	C	1135	IPACdb	Test
1Y7Q	Zinc Finger Protein 174	Zinc Finger Protein 174	A	B	1533	Dcbio dataset	Test
1Z3E	Spx	RNA Polyemerase	A	B	692	LoopFinder	Test
1Z92	Il-2	Il-2r	A	B	998	2P2I	Test
1ZHI	Bah Domain Of Orc1	Sir Orc-Interaction Domain	A	B	703	PPI affinity DB	Test
1ZLH	Carboxypeptidase A1	Carboxypeptidase Inhibitor	A	B	1082	Dcbio dataset	Test
2AB0	Yajl	Yajl	A	B	1339	Dcbio dataset	Test
2AFH	Nifd	Nifh1	A	F	481	LoopFinder	Test
2AST	Cks-1	Skp2	C	B	999	LoopFinder	Test
2BZ6	Blood Coagulation Factor Viia	Blood Coagulation Factor Viia	H	L	1012	Dcbio dataset	Test

2CAR	Inositol Triphosphate Pyrophosphatase	Inositol Triphosphate Pyrophosphatase	A	B	1140	Dcbio dataset	Test
2D0D	2-Hydroxy-6-Oxo-7-Methylocta-2,4-Dienoate Hydrolase	2-Hydroxy-6-Oxo-7-Methylocta-2,4-Dienoate Hydrolase	A	B	1040	Dcbio dataset	Test
2DVN	Hypothetical Protein Ph1917	Hypothetical Protein Ph1917	A	B	1096	Dcbio dataset	Test
2H7I	Enoyl-[Acyl-Carrier-Protein] Reductase [Nadh	Enoyl-[Acyl-Carrier-Protein] Reductase [Nadh	A	B	1541	Dcbio dataset	Test
2I25	Shark Single Domain Antigen Receptor	Hew Lysozyme	N	L	741	PPI affinity DB	Test
2I7D	5'(3')-Deoxyribonucleotidase	5'(3')-Deoxyribonucleotidase	A	B	1128	Dcbio dataset	Test
2QXV	Eed	Ezh2	A	B	1567	IPACdb	Test
2V52	Actin, Alpha Skeletal Muscle	Actin, Alpha Skeletal Muscle	M	B	1065	Dcbio dataset	Test
2VVT	Glutamate Racemase	Glutamate Racemase	A	B	1434	Dcbio dataset	Test
2Z1N	Ape0912	Ape0912	A	B	1625	Dcbio dataset	Test
3BZL	Escu	Escu	B	D	1108	Dcbio dataset	Test
3CM3	Dual Specificity Protein Phosphatase	Dual Specificity Protein Phosphatase	A	B	1030	Dcbio dataset	Test
3CPH	Ras-Related Protein Sec4	Rab Gdp-Dissociation Inhibitor	G	A	878	PPI affinity DB	Test
3F3E	Leut Transporter	Leut Transporter	A	B	1335	Dcbio dataset	Test
3F6Q	Integrin-Linked Protein Kinase	Lim And Senescent Cell Antigen-Like-Containing Domain Protein 1	A	B	993	Dcbio dataset	Test
3H0N	Uncharacterized Protein Duf1470	Uncharacterized Protein Duf1470	A	B	1071	Dcbio dataset	Test
3JRZ	Ccdb	Ccdb	A	B	961	Dcbio dataset	Test
4GEQ	Spc25	Spc24	A	B	1090	LoopFinder	Test
4GPO	Beta1 Adrenergic Receptor	Beta1 Adrenergic Receptor	A	B	903	IPACdb	Test

Supplementary Table 2.3. Descripteurs des PPI

Name	Description
nPTS	Total number of interaction points
Hydro	% of hydrophobic interaction points
Aro	% of aromatic interaction points
Hbond	% of hydrogen-bond interaction points
Ionic	% of of ionic bond interaction points
Hydro1	% of hydrophobic points (25 %<Burial <33.3%)
Hydro2	% of hydrophobic points (33.3 %<Burial <41.6%)
Hydro3	% of hydrophobic points (41.6 %<Burial <50%)
Hydro4	% of hydrophobic points (50 %<Burial <58.3%)
Hydro5	% of hydrophobic points (58.3 %<Burial <66.6%)
Hydro6	% of hydrophobic points (66.6 %<Burial <75%)
Hydro7	% of hydrophobic points 75 %<Burial <83.3%)
Hydro8	% of hydrophobic points (83.3 %<Burial <91.6%)
Hydro9	% of hydrophobic points (91.6%<Burial <100%)
Hydro10	% of hydrophobic points (Burial =100%)
Aro1	% of aromatic points (25 %<Burial <33.3%)
Aro2	% of aromatic points (33.3 %<Burial <41.6%)
Aro3	% of aromatic points (41.6 %<Burial <50%)
Aro4	% of aromatic points (50 %<Burial <58.3%)
Aro5	% of aromatic points (58.3 %<Burial <66.6%)
Aro6	% of aromatic points (66.6 %<Burial <75%)
Aro7	% of aromatic points 75 %<Burial <83.3%)
Aro8	% of aromatic points (83.3 %<Burial <91.6%)
Aro9	% of aromatic points (91.6%<Burial <100%)
Aro10	% of aromatic points (Burial =100%)
Hbond1	% of hydrogen bond points (25 %<Burial <33.3%)
Hbond2	% of hydrogen bond points (33.3 %<Burial <41.6%)
Hbond3	% of hydrogen bond points (41.6 %<Burial <50%)
Hbond4	% of hydrogen bond points (50 %<Burial <58.3%)
Hbond5	% of hydrogen bond points (58.3 %<Burial <66.6%)
Hbond6	% of hydrogen bond points (66.6 %<Burial <75%)
Hbond7	% of hydrogen bond points 75 %<Burial <83.3%)
Hbond8	% of hydrogen bond points (83.3 %<Burial <91.6%)
Hbond9	% of hydrogen bond points (91.6%<Burial <100%)
Hbond10	% of hydrogen bond points (Burial =100%)
Ionic1	% of ionic bond points (25 %<Burial <33.3%)
Ionic2	% of ionic bond points (33.3 %<Burial <41.6%)
Ionic3	% of ionic bond points (41.6 %<Burial <50%)
Ionic4	% of ionic bond points (50 %<Burial <58.3%)
Ionic5	% of ionic bond points (58.3 %<Burial <66.6%)
Ionic6	% of ionic bond points (66.6 %<Burial <75%)
Ionic7	% of ionic bond points 75 %<Burial <83.3%)
Ionic8	% of ionic bond points (83.3 %<Burial <91.6%)

Ionic9	% of ionic bond points (91.6%<Burial <100%)
Ionic10	% of ionic bond points (Burial = 100%)

Supplementary Table 2.4. Prédiction de 100 interface protéine-protéine (FDS jeu de test) par différetnes méthodes

PDB	Set	Chain1	Chain 2	Interface, Å^2	IChemPIC	NOXClass[a]	DiMoVo[b]	PISA	EPPIC
1bs1	Bio	A	B	1485	Bio	99.16	0.70	Bio	Bio
1dfj	Bio	E	I	1309	Xtal	100	0.04	Bio	Bio
1eej	Bio	A	B	858	Bio	98.84	0.51	Bio	na[c]
1ewy	Bio	A	C	762	Xtal	1.76	0.12	Xtal	Xtal
1f2d	Bio	C	D	1895	Bio	99.3	0.68	Bio	na
1f34	Bio	A	B	1522	Bio	na	0.56	Bio	Bio
1f47	Bio	A	B	569	Xtal	100	0.55	Bio	np[d]
1ffw	Bio	A	B	593	Xtal	99.3	0.43	Xtal	Xtal
1flt	Bio	W	X	588	Bio	99.24	-0.01	Bio	Xtal
1gxd	Bio	A	C	1247	Bio	98.49	0.47	Bio	Bio
1hcf	Bio	A	B	1721	Bio	100	0.79	Bio	Bio
1i5h	Bio	W	B	420	Bio	99.97	0.85	np	Xtal
1m9c	Bio	A	D	492	Bio	83.75	0.02	Xtal	Bio
1nw9	Bio	A	B	1066	Bio	99.99	0.62	Bio	Bio
1pxv	Bio	A	C	1174	Bio	99.78	0.40	Bio	Bio
1qa9	Bio	A	B	699	Bio	63	-0.04	Xtal	Xtal
1tue	Bio	A	B	938	Xtal	98.56	0.17	Xtal	Bio
1uj6	Bio	A	B	1281	Bio	86.84	0.65	Bio	Bio
1vk5	Bio	A	B	1146	Bio	99.77	0.20	Bio	Bio
1vkx	Bio	A	B	678	Bio	6.13	0.11	Bio	Bio
1x2r	Bio	A	B	537	Bio	100	0.25	Bio	Bio
1xx9	Bio	A	C	1068	Bio	94.22	0.68	Bio	Bio
1y7q	Bio	A	B	1209	Xtal	99.98	1.05	np	na
1z3e	Bio	A	B	679	Bio	90.94	0.10	Xtal	Bio
1z92	Bio	A	B	947	Bio	93.86	0.66	Xtal	Xtal
1zhi	Bio	A	B	668	Xtal	65.21	-0.07	Xtal	Xtal

1zlh	Bio	A	B	1033	Xtal	100	0.46	Bio	na
2ab0	Bio	A	B	1297	Bio	95.25	0.57	Bio	na
2afh	Bio	A	F	449	Xtal	3.02	-0.05	Xtal	Xtal
2ast	Bio	C	B	937	Bio	99.33	0.17	Bio	Xtal
2bz6	Bio	H	L	981	Bio	100	0.59	Bio	na
2car	Bio	A	B	1100	Bio	94.04	0.29	Bio	na
2d0d	Bio	A	B	1004	Xtal	1.91	0.56	Xtal	na
2dvn	Bio	A	B	635	Bio	86.85	0.19	Bio	na
2h7i	Bio	A	B	1496	Bio	82.54	0.60	Bio	na
2i25	Bio	N	L	714	Bio	98.74	0.51	Xtal	Bio
2i7d	Bio	A	B	1080	Bio	99.18	0.65	Bio	na
2qxv	Bio	A	B	1502	Bio	100	0.46	Bio	Bio
2v52	Bio	M	B	1030	Bio	100	0.46	Bio	na
2vvt	Bio	A	B	1339	Bio	98.99	0.18	Bio	na
2z1n	Bio	A	B	1587	Bio	99.96	0.92	Bio	Bio
3bzl	Bio	B	D	1083	Bio	100	0.73	Bio	na
3cm3	Bio	A	B	1012	Bio	100	0.49	Bio	na
3cph	Bio	G	A	848	Xtal	80.94	0.12	Bio	Bio
3f3e	Bio	A	B	1271	Xtal	34.27	0.62	Bio	na
3f6q	Bio	A	B	948	Bio	99.61	0.29	Bio	na
3h0n	Bio	A	B	1022	Bio	97.64	0.61	Bio	na
3jrz	Bio	A	B	932	Bio	96.48	0.74	Bio	na
4geq	Bio	A	B	1060	Bio	100	0.75	Bio	Bio
4gpo	Bio	A	B	863	Xtal	19.13	0.24	Bio	Xtal
13pk	Xtal	A	D	1393	Xtal	69.89	na	Bio	Xtal
1afk	Xtal	A	B	291	Xtal	0.35	na	na	Xtal
1aq0	Xtal	A	B	722	Xtal	9.26	na	Xtal	Xtal
1b1j	Xtal	A	C	892	Bio	92.48	na	Xtal	Xtal
1bea	Xtal	A	G	524	Xtal	39.16	na	na	Xtal

1bf6	Xtal	B	D	509	Xtal	0	na	Xtal	Xtal
1bs2	Xtal	A	G	1210	Bio	90.04	na	Xtal	Xtal
1byo	Xtal	A	T	298	Xtal	0	na	Xtal	Xtal
1c02	Xtal	A	B	873	Xtal	56.51	na	Bio	Bio
1ck7	Xtal	A	E	953	Bio	2.8	na	Bio	Xtal
1dsu	Xtal	A	C	728	Bio	5.76	na	Xtal	Xtal
1dz4	Xtal	A	U	542	Xtal	2.87	na	Bio	Xtal
1epa	Xtal	A	B	642	Xtal	20.79	na	Bio	Xtal
1fgk	Xtal	A	C	817	Xtal	24.27	na	Xtal	Xtal
1g2a	Xtal	A	B	577	Xtal	3.76	na	Xtal	Xtal
1hf8	Xtal	A	O	538	Xtal	2.1	na	Bio	Xtal
1ihb	Xtal	A	B	329	Xtal	0	na	Xtal	Xtal
1inp	Xtal	A	C	798	Xtal	1.97	na	Xtal	Xtal
1lf2	Xtal	A	B	2168	Bio	99.92	0.92	Bio	Xtal
1lxk	Xtal	A	B	1036	Xtal	1.04	0.00	Xtal	na
1n45	Xtal	A	B	1098	Bio	29.33	0.32	Xtal	na
1nuc	Xtal	A	S	606	Xtal	77.26	na	Xtal	Xtal
1pbg	Xtal	A	R	271	Xtal	0	na	Xtal	Xtal
1pda	Xtal	A	B	885	Xtal	2.9	na	na	Xtal
1pmi	Xtal	A	B	628	Xtal	0	na	na	Xtal
1qjp	Xtal	A	V	586	Xtal	99.98	na	Xtal	Bio
1qme	Xtal	A	B	647	Xtal	42.09	na	Xtal	Xtal
1rhs	Xtal	A	E	863	Xtal	0.68	na	na	Xtal
1the	Xtal	A	B	729	Xtal	51.75	na	Bio	Xtal
1ueb	Xtal	A	B	1150	Xtal	27.4	0.42	Bio	na
1urp	Xtal	A	i	805	Xtal	53.85	na	Xtal	Xtal
1vlz	Xtal	A	B	435	Xtal	8.01	na	Xtal	Xtal
1xca	Xtal	A	B	617	Xtal	84.55	na	Bio	Xtal
1ynq	Xtal	A	B	1103	Xtal	90.13	0.25	Bio	na

2atj	Xtal	A	G	932	Bio	11.13	na	na	Xtal
2bls	Xtal	B	E	610	Xtal	31.97	na	na	Xtal
2cki	Xtal	A	B	1185	Xtal	12.62	0.22	Bio	na
2end	Xtal	A	D	610	Xtal	87.71	na	Xtal	Xtal
2eqa	Xtal	A	B	1110	Xtal	63.85	0.51	Xtal	Xtal
2fgz	Xtal	A	B	1094	Xtal	0.75	0.12	Xtal	Xtal
2hlq	Xtal	A	B	1087	Bio	99.99	0.33	Xtal	na
2nap	Xtal	b	p	747	Bio	4.54	na	Xtal	Xtal
2tps	Xtal	A	B	809	Xtal	78.1	na	Xtal	Xtal
2ugi	Xtal	A	B	741	Xtal	93.7	na	Bio	Xtal
2vt4	Xtal	A	B	681	Xtal	95.21	0.32	Xtal	Xtal
2xov	Xtal	A	B	1200	Bio	99.85	0.37	Xtal	na
3fwk	Xtal	A	B	1048	Xtal	79.84	0.55	Xtal	na
3hzl	Xtal	A	B	1006	Bio	1.58	0.45	Xtal	na
3ita	Xtal	A	B	1142	Bio	47.59	0.52	Bio	Xtal
3mg1	Xtal	A	B	1099	Xtal	25.06	0.01	Xtal	na

[a] Xtal if score <50, Bio if score ≥ 50

[b] Xtal if score <0.5, Bio if score ≥ 0.5

[c] not applicable because the entry was in the training set of the prediction method

[d] no prediction

Supplementary Table 2.5. Prédiction de 142 PPI (Bahadur external set) par différentes méthodes

PDB	Chain1	Chain2	Interface, Å^2	Set	IChemPIC	NOXClass[a]	DiMoVo[b]	PISA	EPPIC
12as	A	B	1964	Biol	Bio	99.99	na[c]	Bio	Bio
1a3c	A	B	1019	Biol	Bio	62.78	na	na	Bio
1a4i	A	B	1396	Biol	Bio	na	na	Bio	Bio
1a4u	A	B	2604	Biol	Bio	100	na	Bio	Bio
1aa7	A	B	1123	Biol	Bio	94.98	na	Xtal	Bio
1ad3	A	B	4068	Biol	Bio	100	na	na	Bio
1ade	A	B	2925	Biol	Xtal	99.95	na	Bio	Bio
1af5	A	B	895	Biol	Bio	95.66	na	na	Bio
1afw	A	B	2491	Biol	Bio	na	na	na	Bio
1ajs	A	B	3530	Biol	Bio	na	na	na	Bio
1al0	B	F	1266	Biol	Bio	99.99	na	Bio	Bio
1amk	A	B	1507	Biol	Bio	99.93	na	na	Bio
1aor	A	B	1281	Biol	Bio	18.05	na	na	Bio
1aq6	A	B	2235	Biol	Bio	na	na	na	Bio
1auo	B	C	704	Biol	Xtal	0.33	na	na	Bio
1b3a	A	B	757	Biol	Bio	na	na	Bio	Bio
1b5e	A	B	2636	Biol	Bio	na	na	Bio	Bio
1b67	A	B	1650	Biol	Bio	na	na	Bio	Xtal
1b8a	A	B	4445	Biol	Bio	na	na	Bio	Bio
1b8j	A	B	3881	Biol	Bio	na	na	Bio	Bio
1bam	A	B	777	Biol	Bio	49.42	na	na	Xtal
1bbh	A	B	792	Biol	Bio	87.21	na	Bio	Bio
1bd0	A	B	3206	Biol	Bio	100	na	Bio	Bio
1bif	A	B	956	Biol	Xtal	57.59	na	na	Xtal
1biq	A	B	3086	Biol	Bio	100	na	Bio	Bio
1bis	A	B	1544	Biol	Bio	100	na	Bio	Bio

1bjw	A	B	2999	Biol	Bio	99.99	na	Bio	Bio
1bkp	A	B	2360	Biol	Bio	99.97	na	Bio	Bio
1bmd	A	B	1640	Biol	Bio	95.72	na	Bio	Bio
1brw	A	B	1094	Biol	Bio	75.9	na	Bio	Bio
1bsl	A	B	1926	Biol	Bio	93.23	na	Bio	Bio
1bsr	A	B	1986	Biol	Bio	100	na	na	Bio
1buo	A	B	2189	Biol	Xtal	100	na	na	Bio
1bxg	A	B	1084	Biol	Bio	72.03	na	Xtal	Xtal
1bxk	A	B	1334	Biol	Bio	97.64	na	Bio	Bio
1cdc	A	B	3990	Biol	Bio	100	na	Bio	Bio
1cg2	A	D	1351	Biol	Bio	97.82	na	na	Bio
1chm	A	B	3422	Biol	Bio	100	na	na	bio
1cmb	A	B	1864	Biol	Bio	na	na	na	Bio
1cnz	A	B	2535	Biol	Bio	na	na	Bio	Bio
1coz	A	B	1080	Biol	Bio	na	na	Bio	Bio
1csh	A	B	5176	Biol	Bio	100	na	na	Bio
1ctt	A	B	2043	Biol	Bio	99.93	na	na	Bio
1cvu	A	B	2523	Biol	Bio	99.59	na	Bio	Bio
1czj	A	B	855	Biol	Xtal	99.6	na	na	Bio
1daa	A	B	2369	Biol	Bio	100	na	na	Bio
1dor	A	B	2282	Biol	Bio	na	na	Bio	Bio
1dpg	A	B	2364	Biol	Bio	99.77	na	Bio	Bio
1dqs	A	B	1739	Biol	Bio	94.4	na	Bio	Bio
1dxg	A	B	759	Biol	Bio	99.99	na	Bio	Bio
1e98	A	B	791	Biol	Bio	98.7	na	Bio	Bio
1ebh	A	B	1824	Biol	Bio	98.41	na	Bio	Bio
1f13	A	B	2719	Biol	Xtal	99.56	na	Bio	Bio
1fip	A	B	1640	Biol	Bio	100	na	na	bio
1fro	A	B	3762	Biol	Bio	100	na	na	Bio

1gvp	A	B	929	Biol	Xtal	98.9	na	na	Bio
1hhp	A	B	1622	Biol	Bio	100	na	Xtal	Bio
1hjr	B	D	985	Biol	Bio	na	na	na	Bio
1hss	A	B	1128	Biol	Bio	na	na	na	Bio
1hxp	A	B	3492	Biol	Bio	100	na	Bio	Bio
1icw	A	B	965	Biol	Bio	97.34	na	na	Bio
1imb	A	B	1695	Biol	Bio	95.38	na	na	Xtal
1isa	A	B	951	Biol	Bio	na	na	na	Bio
1ivy	A	B	1666	Biol	Bio	99.67	na	Bio	Bio
1jhg	A	B	2294	Biol	Bio	100	na	na	Bio
1jsg	A	B	815	Biol	Bio	79.72	na	na	Xtal
1kba	A	B	517	Biol	Bio	19.3	na	na	Bio
1kpf	A	A	1912	Biol	Bio	100	na	na	Bio
1lyn	A	B	981	Biol	Bio	88.17	na	na	Xtal
1m6p	A	B	1086	Biol	Bio	97.3	na	Bio	Bio
1mkb	A	B	1648	Biol	Bio	99.99	na	Bio	Bio
1mor	A	B	2635	Biol	Bio	99.94	na	Bio	Bio
1nox	A	B	3161	Biol	Bio	100	na	na	Bio
1nse	A	B	2865	Biol	Bio	na	na	Bio	Bio
1nsy	A	B	2693	Biol	Bio	100	na	na	Bio
1oac	A	B	7381	Biol	Bio	100	na	na	Bio
1opy	A	B	1073	Biol	Bio	99.99	na	na	Xtal
1pgt	A	B	1249	Biol	Bio	94.68	na	na	Bio
1pre	A	B	2451	Biol	Xtal	99.92	na	na	Bio
1qfh	A	B	2376	Biol	Bio	na	na	Bio	Bio
1qhi	A	B	1749	Biol	Bio	99.99	na	Bio	Bio
1qr2	A	B	2011	Biol	Bio	99.99	na	Bio	Bio
1r2f	A	B	1814	Biol	Bio	99.97	na	Bio	Bio
1reg	X	Y	681	Biol	Xtal	28.31	na	Bio	Xtal

1rfb	A	B	2820	Biol	Bio	100	na	na	Bio
1rpo	A	B	1421	Biol	Bio	100	na	na	Bio
1ses	A	B	2281	Biol	Bio	99.95	na	na	Bio
1slt	A	B	555	Biol	Xtal	88.13	na	na	Xtal
1smn	A	B	913	Biol	Bio	87.13	na	na	Xtal
1smt	A	B	2020	Biol	Bio	na	na	na	Bio
1sox	A	B	1460	Biol	Bio	na	na	na	Bio
1tc1	A	B	1555	Biol	Bio	99.82	na	Bio	Bio
1tox	B	C	3906	Biol	Bio	100	na	na	Bio
1trk	A	B	4546	Biol	Bio	na	na	na	Bio
1uby	A	B	2223	Biol	Bio	99.9	na	na	Bio
1utg	A	B	1521	Biol	Bio	100	na	na	Bio
1vfr	A	B	3556	Biol	Bio	100	na	Bio	Bio
1vok	A	B	1666	Biol	Bio	na	na	Bio	Bio
1wtl	A	B	722	Biol	Bio	93.23	na	Bio	Xtal
1xso	A	B	692	Biol	Bio	na	na	na	Bio
2arc	A	B	831	Biol	Bio	97.84	na	Xtal	Bio
2ccy	A	B	841	Biol	Xtal	93.7	na	na	Bio
2hdh	A	B	1585	Biol	Bio	na	na	Bio	Bio
2ilk	A	B	4675	Biol	Bio	100	na	na	Bio
2lig	A	B	1685	Biol	Bio	99.94	na	Bio	Bio
2mcg	1	2	1746	Biol	Bio	99.17	na	Bio	Bio
2nac	A	B	3887	Biol	Bio	na	na	Bio	Bio
2ohx	A	B	1765	Biol	Bio	99.84	na	Bio	Xtal
2spc	A	B	2614	Biol	Bio	100	na	Bio	Bio
2sqc	A	B	843	Biol	Xtal	13.67	na	Xtal	Xtal
2tct	A	B	2744	Biol	Bio	99.99	na	na	Bio
2tgi	A	B	1315	Biol	Bio	100	na	na	Bio
3dap	A	B	2732	Biol	Bio	100	na	Bio	Bio

3grs	A	B	3390	Biol	Bio	100	na	na	Bio
3sdh	A	B	929	Biol	Bio	99.86	na	na	Bio
3ssi	A	B	894	Biol	Bio	69.3	na	na	Bio
4cha	B	C	2986	Biol	Bio	100	na	Bio	Bio
4kbp	B	C	1558	Biol	Bio	96.01	na	na	Bio
5csm	A	B	1903	Biol	Bio	99.85	na	na	Bio
5rub	A	B	2913	Biol	Bio	99.96	na	Bio	Bio
8prk	A	B	1014	Biol	Bio	88.78	na	Bio	Bio
9wga	A	B	2277	Biol	Bio	100	na	na	Bio
1a39	A	B	528	Xtal	Xtal	8.93	na	Xtal	Xtal
1ag9	A	B	444	Xtal	Xtal	49.32	na	Bio	Xtal
1bc2	B	C	657	Xtal	Bio	50	na	na	Bio
1caq	A	B	711	Xtal	Xtal	88.31	na	Xtal	Bio
1e0s	A	B	899	Xtal	Bio	90.33	na	Bio	Xtal
1feh	A	B	1626	Xtal	Bio	98.09	na	na	Xtal
1gjm	A	B	897	Xtal	Xtal	0.44	na	Xtal	Xtal
1hvf	A	B	541	Xtal	Xtal	0.49	na	Xtal	Xtal
1mbl	B	C	619	Xtal	Xtal	0.86	na	Xtal	Xtal
1ml1	K	I	743	Xtal	Xtal	77.03	na	Bio	Xtal
1mwc	A	B	453	Xtal	Xtal	1.24	na	Xtal	Xtal
1naw	A	B	1225	Xtal	Bio	95.58	na	Bio	Xtal
1qdm	A	B	822	Xtal	Xtal	96.17	na	Xtal	Xtal
1qpa	A	B	893	Xtal	Xtal	27.68	na	Bio	Xtal
1qs8	A	B	1186	Xtal	Xtal	69.68	na	Bio	Xtal
1qsn	A	B	815	Xtal	Xtal	100	na	Bio	Bio
1rne	A	B	1171	Xtal	Xtal	28.24	na	Bio	Xtal
1trn	A	B	761	Xtal	Xtal	72.15	na	Xtal	Xtal
3ng1	A	B	570	Xtal	Xtal	0	na	Xtal	Xtal
5tss	A	B	1452	Xtal	Xtal	95.1	na	Xtal	Xtal

[a] Xtal if score <50, Bio if score ≥ 50

[b] Xtal if score <0.5, Bio if score ≥ 0.5

[c] not applicable because the entry was in the training set of the prediction method

Supplementary Table 2.6. Prédiction ode143 i0nterfaces protéine-protéine (Ponstingl external set) par différentes méthodes

PDB	Chain1	Chain2	Interface, Å^2	Set	IChemPIC	NOXClass[a]	DiMoVo[b]	PISA	EPPIC
1a3c	A	B	991	Biol	Bio	62.78	na	na[c]	Bio
1ad3	A	B	3941	Biol	Bio	100	na	na	Bio
1af5	A	B	853	Biol	Bio	95.66	na	na	Bio
1afw	A	B	2398	Biol	Bio	na	na	na	Bio
1ajs	A	B	3443	Biol	Bio	na	na	na	Bio
1al0	B	F	1222	Biol	Bio	99.99	na	np[d]	Bio
1alk	A	B	3851	Biol	Bio	100	1.00	na	Bio
1amk	A	B	1472	Biol	Bio	99.99	na	na	Bio
1aom	A	B	1243	Biol	Bio	na	0.06	na	Bio
1aor	A	B	1234	Biol	Bio	18.05	na	na	Bio
1aq6	A	B	2217	Biol	Bio	na	na	na	Bio
1auo	B	C	667	Biol	Xtal	0.33	na	na	Bio
1bam	A	B	746	Biol	Bio	49.42	na	na	Xtal
1bif	A	B	895	Biol	Xtal	57.79	na	na	Xtal
1bsr	A	B	1920	Biol	Bio	100	na	na	Bio
1buo	A	B	1978	Biol	Xtal	100	na	na	Bio
1cg2	A	D	1300	Biol	Bio	97.82	na	na	Bio
1chm	A	B	3316	Biol	Bio	100	na	na	Bio
1cmb	A	B	1812	Biol	Bio	na	na	na	Bio
1cp2	A	B	953	Biol	Xtal	na	0.49	na	Bio
1csh	A	B	5087	Biol	Bio	100	na	na	Bio
1ctt	A	B	1989	Biol	Bio	99.93	na	na	Bio
1czj	A	B	798	Biol	Xtal	99.6	na	na	Bio
1daa	A	B	2300	Biol	Bio	100	na	na	Bio
1fip	A	B	1607	Biol	Bio	100	na	na	Bio

1fro	C	D	3628	Biol	Bio	100	na	na	Bio
1gvp	A	B	903	Biol	Xtal	98.8	na	na	Bio
1hjr	B	D	964	Biol	Bio	na	na	na	Bio
1hss	A	B	1099	Biol	Bio	na	na	na	Bio
1icw	A	B	987	Biol	Bio	97.34	na	na	Bio
1imb	A	B	1648	Biol	Bio	95.38	na	na	Xtal
1isa	A	B	920	Biol	Bio	na	na	na	Bio
1iso	A	B	3305	Biol	Bio	100	0.88	na	Bio
1jhg	A	B	2209	Biol	Bio	100	na	na	Bio
1jsg	A	B	793	Biol	Bio	79.72	na	na	Xtal
1kba	A	B	492	Biol	Bio	19.3	na	na	Bio
1kpf	A	B	1869	Biol	Bio	100	na	na	Bio
1lyn	A	B	945	Biol	Bio	88.17	na	na	Xtal
1mjl	A	B	1775	Biol	Bio	100	0.93	na	Bio
1mka	A	B	1618	Biol	Bio	99.99	0.97	na	Bio
1moq	A	B	2538	Biol	Bio	99.94	0.97	na	Bio
1nox	A	B	3034	Biol	Bio	100	na	na	Bio
1nsy	A	B	2611	Biol	Bio	100	na	na	Bio
1oac	A	B	7158	Biol	Bio	100	na	na	Bio
1opy	A	B	1046	Biol	Bio	99.99	na	na	Xtal
1otp	A	B	831	Biol	Bio	3.37	0.39	na	Bio
1pgt	A	B	1231	Biol	Bio	94.68	na	na	Bio
1pre	A	B	2291	Biol	Xtal	99.92	na	na	Bio
1puc	A	B	2169	Biol	Bio	100	0.58	na	Bio
1rfb	A	B	2645	Biol	Bio	100	na	na	Bio
1rpo	A	B	1401	Biol	Bio	100	na	na	Bio
1ses	A	B	2231	Biol	Bio	99.95	na	na	Bio
1slt	A	B	544	Biol	Xtal	88.13	na	na	Xtal
1smn	A	B	868	Biol	Bio	87.13	na	na	Xtal

1smt	A	B	1968	Biol	Bio	na	na	na	Bio
1sox	A	B	1412	Biol	Bio	na	na	na	Bio
1tox	B	C	3774	Biol	Bio	100	na	na	Bio
1trk	A	B	4487	Biol	Bio	na	na	na	Bio
1tys	A	B	2330	Biol	Bio	99.99	0.42	na	Bio
1uby	A	B	2169	Biol	Bio	99.9	na	na	Bio
1utg	A	B	1482	Biol	Bio	100	na	na	Bio
1wgj	A	B	961	Biol	Bio	na	0.58	na	Bio
1xso	A	B	655	Biol	Bio	na	na	na	Bio
2ccy	A	B	798	Biol	Xtal	93.7	na	na	Bio
2ilk	A	B	4566	Biol	Bio	100	na	na	Bio
2rsp	A	B	1493	Biol	Bio	99.53	0.79	na	Bio
2tct	A	B	2668	Biol	Bio	99.99	na	na	Bio
2tgi	A	B	1266	Biol	Bio	100	na	na	Bio
3grs	A	B	3283	Biol	Bio	100	na	na	Bio
3pgh	A	B	2376	Biol	Bio	99.59	0.96	na	Bio
3sdh	A	B	899	Biol	Bio	99.86	na	na	Bio
3ssi	A	B	866	Biol	Bio	69.3	na	na	Bio
4kbp	B	C	1472	Biol	Bio	96.01	na	na	Bio
5csm	A	B	1827	Biol	Bio	99.85	na	na	Bio
5tmp	A	B	1185	Biol	Bio	83.17	0.86	na	Bio
9wga	A	B	2296	Biol	Bio	100	na	na	Bio
16pk	A	B	461	Xtal	Xtal	2.88	-0.05	Xtal	Xtal
1a0k	A	B	445	Xtal	Xtal	1.2	0.09	Xtal	Xtal
1a19	A	B	535	Xtal	Xtal	42.01	0.35	na	Xtal
1a8o	A	B	517	Xtal	Bio	62.91	0.55	na	Xtal
1aay	A	B	443	Xtal	Bio	7.32	0.09	Bio	Xtal
1afk	B	C	889	Xtal	Xtal	93.9	na	na	Xtal
1ahq	A	B	506	Xtal	Xtal	2.37	-0.01	Xtal	Xtal

1am6	A	B	448	Xtal	Xtal	0	-0.27	Xtal	Xtal
1amj	A	B	623	Xtal	Xtal	1.18	0.16	na	Xtal
1aoh	A	B	683	Xtal	Xtal	27.31	0.03	na	Bio
1aua	A	B	763	Xtal	Xtal	0	0.36	na	Xtal
1aun	A	B	546	Xtal	Xtal	90.2	-0.07	na	Xtal
1avp	A	B	831	Xtal	Bio	100	0.50	Bio	Bio
1bea	A	B	560	Xtal	Xtal	40.29	na	na	Bio
1bmb	A	I	364	Xtal	Bio	99.99	0.32	Bio	Bio
1bn8	A	B	353	Xtal	Xtal	1.7	0.03	Xtal	Xtal
1bp1	A	B	870	Xtal	Bio	4.58	0.00	na	Xtal
1bry	Y	Z	487	Xtal	Xtal	1.04	0.37	na	Xtal
1bu1	A	B	491	Xtal	Xtal	61.35	0.07	Xtal	Xtal
1bwz	A	B	720	Xtal	Xtal	11.31	0.30	na	Xtal
1c3d	A	B	713	Xtal	Xtal	21.24	0.05	Xtal	Xtal
1ckm	A	B	1596	Xtal	Bio	50	0.28	na	Bio
1ctj	A	B	416	Xtal	Xtal	30.64	0.25	na	Xtal
1dff	A	B	499	Xtal	Xtal	0	0.37	na	Xtal
1djx	A	B	594	Xtal	Xtal	23.81	0.14	na	Xtal
1dmr	A	B	937	Xtal	Xtal	7.74	0.17	na	Xtal
1ema	A	B	432	Xtal	Xtal	0.65	0.28	Xtal	Xtal
1esf	A	B	553	Xtal	Xtal	0.42	-0.04	na	Xtal
1eso	A	B	419	Xtal	Xtal	9.98	0.20	Xtal	Xtal
1fdr	A	B	461	Xtal	Xtal	0	0.12	na	Xtal
1flp	A	B	317	Xtal	Xtal	0	0.05	Xtal	Xtal
1fsu	A	B	522	Xtal	na	59.58	-0.29	na	Xtal
1gci	A	B	420	Xtal	Xtal	4.44	0.07	Xtal	Xtal
1iae	A	B	408	Xtal	Xtal	2.24	0.11	na	Xtal
1inp	A	B	817	Xtal	Xtal	2.27	na	Xtal	Xtal
1ips	A	B	957	Xtal	Xtal	89.42	0.07	na	Xtal

1lrv	A	B	817	Xtal	Xtal	60.56	0.08	na	Xtal
1mb1	A	B	437	Xtal	Bio	37.67	0.04	Xtal	Bio
1mdt	A	B	610	Xtal	Xtal	1.36	0.23	na	Xtal
1mh1	A	B	477	Xtal	Xtal	0.97	-0.09	na	Xtal
1nuc	A	B	614	Xtal	Xtal	74.28	na	Xtal	Xtal
1ops	A	B	388	Xtal	Xtal	41.77	0.00	Xtal	Xtal
1pda	A	B	947	Xtal	Xtal	2.68	na	na	Bio
1pgs	A	B	531	Xtal	Xtal	4.35	0.07	na	Xtal
1pjr	A	B	886	Xtal	Xtal	27.22	-0.09	Xtal	Bio
1pmi	A	B	668	Xtal	Xtal	0	na	na	Xtal
1ps1	A	B	950	Xtal	Xtal	91.92	0.18	na	Bio
1rgp	A	B	531	Xtal	Bio	6.83	1.08	Xtal	Xtal
1rhs	A	B	641	Xtal	Xtal	0.37	na	na	Xtal
1uch	A	B	408	Xtal	Bio	15.41	0.01	Xtal	Xtal
1uro	A	B	1129	Xtal	Bio	98.78	0.54	Xtal	Bio
1yge	A	B	445	Xtal	Xtal	0	-0.09	Xtal	Xtal
232l	A	B	795	Xtal	Xtal	78.96	0.03	na	Xtal
2abx	A	B	585	Xtal	Xtal	95.51	0.01	na	Xtal
2atj	B	C	615	Xtal	Xtal	0.8	na	na	Xtal
2bls	A	B	756	Xtal	Xtal	2.34	na	na	Xtal
2cy3	A	B	129	Xtal	Xtal	1.59	0.28	Xtal	Xtal
2end	A	B	660	Xtal	Xtal	87.67	na	Xtal	Xtal
2fgf	A	B	396	Xtal	na	64.89	0.43	Xtal	Xtal
2gpr	A	B	815	Xtal	Bio	83.19	0.14	Xtal	Bio
2hex	B	C	607	Xtal	Xtal	13.88	0.22	na	Bio
2mhr	A	B	346	Xtal	Xtal	6.22	0.05	Xtal	Xtal
2pth	A	B	537	Xtal	Xtal	27.89	-0.27	Xtal	Bio
3dfr	A	B	399	Xtal	Xtal	0.59	0.15	Xtal	Xtal
3sil	A	B	647	Xtal	Xtal	45.48	0.12	Xtal	Xtal

5cp4	A	B	772	Xtal	Xtal	2.09	-0.17	Xtal	Xtal
8paz	A	B	519	Xtal	Xtal	2.39	0.02	Xtal	Xtal

[a] Xtal if score <50, Bio if score ≥ 50

[b] Xtal if score <0.5, Bio if score ≥ 0.5

[c] not applicable because the entry was in the training set of the prediction method

[d] no prediction

Supplementary Table 2.7. Prédiction de 66 interfaces protéine-proteine (IPAC validation set 3) par différents outils

PDB	Chain1	Chain2	Interface, Å^2	Kd, M	IchemPIC	NOXClass[a]	DiMoVo[b]	PISA	EPPIC
1r8s	A	E	1493	1.40E-04	Bio	100	0.53	Bio	Bio
1nw9	A	B	1057	7.40E-05	Bio	89.88	0.00	Xtal	Bio
2oob	A	B	417	6.00E-05	Xtal	99.95	0.65	Bio	Xtal
1i2m	C	D	1336	3.16E-05	Xtal	na[c]	0.55	Bio	Bio
1wq1	R	G	1458	1.70E-05	Bio	97.43	0.34	Bio	Xtal
1gcq	C	B	605	1.68E-05	na	99.89	0.73	Bio	Xtal
1ak4	A	D	516	1.60E-05	Xtal	na	0.62	Xtal	Bio
2pcc	C	D	530	1.00E-05	Xtal	97.87	0.68	Xtal	Xtal
1qa9	C	D	674	9.00E-06	Xtal	69.12	-0.07	Xtal	Xtal
1z0k	C	D	926	7.20E-06	Bio	100	0.54	Bio	Bio
1e96	A	B	592	6.00E-06	Xtal	35.24	0.02	Bio	Xtal
1sbb	A	C	1029	3.00E-06	Xtal	99.87	0.51	Bio	Xtal
1b6c	H	B	324	2.80E-06	Xtal	na	0.28	Bio	Xtal
1efn	C	D	604	2.50E-06	Bio	38.4	-0.18	Xtal	Xtal
1he8	A	B	651	2.50E-06	Xtal	99.69	0.73	Bio	Bio
2btf	A	P	1030	2.30E-06	Bio	99.4	0.41	Bio	Bio
2ot3	A	B	1152	1.80E-06	Bio	99.96	0.50	Bio	Bio
1gpw	C	D	1081	1.50E-06	na	97.92	0.15	Bio	Xtal
1j2j	A	B	604	1.10E-06	na	98.65	0.26	Bio	Bio
1e6e	A	B	1157	8.60E-07	Xtal	1.13	0.64	Xtal	Bio
1grn	A	B	1169	3.88E-07	na	99.11	0.21	Bio	Bio
2hrk	A	B	798	1.93E-07	na	97.49	0.26	Bio	Bio
2c0l	A	B	1004	1.09E-07	na	93.53	0.36	Bio	Bio
2hqs	B	C	1173	9.00E-08	na	99.69	0.56	Bio	Xtal
1buh	A	B	659	7.70E-08	Bio	na	0.01	Xtal	Xtal
1atn	A	D	887	4.50E-08	Bio	na	0.81	Xtal	Xtal

2cfh	B	D	1207	4.50E-08	Bio	92.49	0.42	Xtal	Bio
2hle	A	B	1060	4.00E-08	Bio	99.35	0.06	Xtal	Bio
1h1v	A	G	1036	2.30E-08	Bio	81.77	0.32	Xtal	Bio
2ajf	F	B	831	1.62E-08	na	100	-0.06	Bio	Xtal
1tmq	A	B	1199	1.10E-08	Bio	na	0.13	Bio	Bio
1pxv	C	A	1166	1.00E-08	Bio	95.16	0.12	Xtal	Xtal
1m10	A	B	1044	5.80E-09	na	89.88	0.80	Bio	Bio
2uuy	A	B	640	5.60E-09	Bio	99.91	0.05	Bio	Bio
1oph	A	B	679	5.00E-09	Bio	99.9	0.35	Bio	Bio
1ewy	C	A	774	3.57E-09	Xtal	99.4	0.40	Bio	Bio
1kxq	C	F	1030	3.50E-09	Bio	na	0.35	Bio	Xtal
1bkd	R	S	1581	3.30E-09	Bio	na	0.00	Xtal	Bio
1eaw	A	B	935	1.70E-09	Bio	99.97	-0.03	Xtal	Xtal
2b42	A	B	1262	1.07E-09	na	100	1.02	Bio	Bio
1ira	X	Y	1684	1.00E-09	Xtal	97.79	0.93	Bio	Bio
1kxp	A	D	1669	1.00E-09	Bio	65.13	0.04	Bio	Xtal
2i25	O	M	684	1.00E-09	Bio	99.43	0.20	Bio	Xtal
1t6b	X	Y	972	4.00E-10	Bio	100	0.37	Bio	Bio
1r0r	E	I	705	3.40E-10	Bio	99.93	0.83	Bio	Bio
1ibr	C	D	1696	3.00E-10	Bio	100	0.81	Bio	Bio
1acb	E	I	771	1.00E-10	Bio	99.99	0.59	Bio	Bio
1ppe	E	I	742	1.00E-10	Bio	88.46	0.21	Bio	Bio
1xqs	C	A	1175	4.00E-11	na	100	0.20	Xtal	Bio
1mah	A	F	1070	2.50E-11	na	99.78	0.40	Bio	Bio
1cgi	E	I	1025	1.60E-11	Bio	100	0.59	Bio	Bio
1bvn	P	T	1109	9.00E-12	na	na	-0.05	Xtal	Xtal
1avx	A	B	794	6.00E-14	Bio	93.53	0.31	Bio	Xtal
1ay7	A	B	617	6.00E-14	na	99.92	0.55	Bio	Bio
1dfj	E	I	1292	5.90E-14	Xtal	35.24	0.35	Xtal	Xtal

7cei	A	B	691	5.00E-15	Bio	99.35	0.52	Bio	Bio
1f34	A	B	1572	unknown	Xtal	na	0.86	Bio	Bio
1fq1	A	B	1016	unknown	Bio	99.55	0.28	Xtal	Xtal
1he1	C	A	1057	unknown	Bio	99.59	0.37	Bio	Bio
1udi	E	I	1010	unknown	Bio	98.61	0.40	Xtal	Bio
1xd3	C	D	1088	unknown	na	99.93	0.67	Bio	Xtal
1yvb	A	I	869	unknown	Bio	99.91	0.79	Bio	Bio
1z5y	D	E	672	unknown	Bio	68.14	0.27	Xtal	Bio
2h7v	D	B	702	unknown	Xtal	100	0.03	Bio	Bio
2nz8	A	B	1296	unknown	Bio	99.79	0.66	Bio	Bio
2sni	E	I	815	unknown	Bio	97.46	0.13	Xtal	Bio

[a] Xtal if score <50, Bio if score ≥ 50

[b] Xtal if score <0.5, Bio if score ≥ 0.5

[c] not applicable because the entry was in the training set of the prediction method

Chapitre 3

De la cavité au pharmacophore

3.1. Introduction

La conception de médicaments assisté par ordinateur[1] est devenue un outil standard pour assister les chimistes médicinaux dans l'identification et/ou l'optimisation de touches pour des cibles d'intérêt pharmaceutique. Les méthodes correspondantes sont classiquement divisées en deux catégories selon qu'elles sont basées sur la structure des ligands[2] ou des protéines cibles (site actif)[3]. Elles vont de pair avec l'évolution grandissante des connaissances sur les ligands biologiquement actifs ainsi que sur les sites de liaison protéine-ligand. Parmi les méthodes basées sur les ligands, la recherche par pharmacophore[4] est très populaire pour plusieurs raisons : (1) le concept de pharmacophore est très facile à comprendre et intuitif aussi bien pour les chémoinformaticiens que pour les chimistes médicinaux; (2) cette technique n'a besoin d'aucune connaissance préalable sur la structure tridimensionnelle de la protéine cible; (3) elle ne souffre pas des mêmes inconvénients[5] que les méthodes basées sur les structures de protéines (par exemple l'estimation de l'énergie libre de liaison) car leurs fonctions de score[6] sont purement topologiques; (4) aligner des ligands sur des pharmacophores aide naturellement à leur optimisation future de touches par ajout de propriétés manquantes ou délétion de propriétés non alignées.

Un criblage de chimiothèques par recherche pharmacophorique classique commence par l'alignement de toutes les molécules ayant le même effet sur une cible, puis par l'extraction des propriétés communes afin de créer le pharmacophore. La dernière étape consiste enfin à cribler la chimiothèque afin d'identifier les touches vérifiant les éléments du pharmacophore. Quand la structure du complexe protéine-ligand est disponible, un pharmacophore d'interaction peut être déterminé en alignant les éléments pharmacophoriques uniquement sur les atomes du ligand en interaction directe.[7–10] Il reste néanmoins énormément de protéines dont la structure tridimensionnelle (3D) est connue mais pour lesquelle les ligands sont toujours manquants (ex: les interfaces protéine-protéine). Dans ce cas de figure, il peut être intéressant de créer un pharmacophore à partir de la simple connaissance de la structure 3D du site de liaison ciblé. Plusieurs méthodes ont été proposées depuis une dizaine d'années pour combler le vide entre les méthodes basées sur la structure de protéine (docking) et la recherche par pharmacophore basée sur les ligands. Les pharmacophore fondés uniquement sur la structure de la cible sont généralement créés à l'aide de sondes (atomes, fragments) permettant de localiser les positions énergiquement favorables à une interaction protéine-ligand. Les méthodes basées sur des grilles (GRID[11], SuperStar[12], FTMap[13],

VolSite[14]) localisent les positions favorables dans une grille 3D englobant l'ensemble de la protéine ou sur une cavité définie comme étant un site de liaison potentiel. Les points de la grille correspondant aux valeurs d'interaction optimales sont ensuite sauvegardés pour chaque sonde [15–17] puis sont transformés en autant d'éléments du futur pharmacophore. Les méthodes à base de fragments essaient de prédire les points chaud d'interaction à partir de simulations par dynamique moléculaire des protéines hydratées (e.g. MCSS[18], SILCS[19], HSRP[20]) en presence de multiples copies de fragments représentant l'ensemble des propriétés possibles d'un pharmacophore (ex: accepteur et donneur de liaison hydrogène, hydrophobe). La dernière possibilité est de réaliser des prédiction topologiques de la position des éléments de pharmacophore en scannant une cavité ainsi que les acides aminés la composant, le but étant de créer des positions d'interaction idéale dans une espace 3D (sphère, cônes) où les atomes de ligands peuvent interagir favorablement avec la surface de la protéine. La première méthode à utiliser cette technique fût LUDI[21] qui a ensuite inspiré de nombreux algorithmes de génération de pharmacophores basés sur la structure des protéines (Virtual ligand[22], SBP[23], HS-Pharm[24], Snooker[25], Examplar[26]) .

Quelle que soit la méthode, le nombre d'éléments pharmacophoriques générés (quelques centaines) reste trop élevé et dépasse largement la complexité autorisée par la recherche de pharmacophore par des algorithmes d'alignement de sphères rigides. Le nombre d'éléments pharmacophoriques doit donc être considérablement réduit à des valeurs généralement inférieures à 10. Le nombre d'éléments peut être diminué en effectuant une présélection basée sur des critères énergétiques[18–20], des critères d'enfouissement[17], des critères de superposition avec des site d'hydratation[20] ou de localisation en se basant sur la connaissance des points chauds[24] d'interaction. Toutes ces méthodes se terminent par un algorithme d'agglomération hiérarchique des propriétés par type et par distance.

La recherche de pharmacophores créés à partir de la structure de la protéine cible a fait ses preuves et est au moins aussi efficace que de l'arrimage moléculaire quant à l'enrichissement en molécules actives lors de criblages virtuels[19,20,24,26]. Cette méthode souffre néanmoins d'un manque d'automatisation et nécessite de nombreuses interventions humaines, les étapes de construction citées précédemment étant fastidieuses et fortement dépendantes de nombreux facteurs (utilisateur, choix des sondes et des seuils d'énergie d'interaction, nombre de clusters). De plus la précision des pharmacophores basés sur les structures de protéines afin d'aligner un ligand dans une cavité d'intérêt a été peu étudiée[27] et rarement comparée aux résultats obtenus par arrimage moléculaire.

Pour répondre à ces problèmes, nous avons amélioré un outil de détection de cavité précédemment décrit au laboratoire (VolSite[14]) afin d'en automatiser les nombreuse étapes allant de la détection d'une cavité et la définition finale d'un pharmacophore. VolSite a été intégré à la suite logicielle IChem[28] afin de réaliser l'ensemble des étapes suivante : (1) détection de l'ensemble des cavités présentes à la surface d'une protéine cible, (2) prédiction de la droguabilité structurale de chaque cavité, (3) détection et création de pharmacophores à partir des structures 3D des cavités d'intérêt, (4) alignement de ligands sur les pharmacophores par un algorithme d'alignement de forme, (5) élimination des poses incorrectes, (6) évaluation des poses par quantification de la qualité de l'alignement.

Ce chapitre décrit notamment l'automatisation des étapes 1 à 5, l'évaluation quantitative des alignements ligand-pharmacophore (étape 6) restant encore à optimiser.

3.2. Méthodes

3.2.1. Jeux de données

Astex DiverseSet :

Les 85 entrées du jeu Astex DiverseSet[29] (**Annexe 3.1**) ont été téléchargées sur le site du CCDC (http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/astex_diverse/) et étudiées de la manière suivante. Pour chaque entrée, le complexe protéine ligand est récréé avec Sybyl-X-2.1.1 (Certera Inc, Princeton, U.S.A.), cela en ajoutant le ligand (fichier au format mol2) dans la protéine (fichier au format mol2). Les molécules d'eau liées sont récupérées dans le fichier de structure PDB original, l'ensemble des hydrogènes est supprimé, le complexe final hydraté (atomes lourds uniquement) est protoné à l'aide de Protoss[30]. Les ions et cofacteurs ne possédant pas d'atomes lourds dans une sphère de 4.5Å de rayon (centrée sur le centre de masse du ligand) ne sont pas conservés. Les molécules d'eau sont conservées à deux conditions : (1) l'atome d'oxygène est situé dans la sphère précédemment décrite; (2) la molécule d'eau réalise au moins deux liaisons hydrogènes (distance accepteur-donneur $\leq$ 3.5Å, l'angle donneur-hydrogène-accepteur $\geq$ 120°) avec la protéine. Le ligand et la protéine (avec les ions et cofacteurs) sont finalement enregistrés dans deux fichiers mol2 distincts.

ScPDB DiverseSet :

213 complexes protéine-ligand (**Annexe 3.2**) ont été sélectionés dans la base de données de complexes protéine-ligands sc-PDB[31], en tenant compte de la diversité des modes d'interactions mesurés par alignement de graphes d'interactions (Grim)[28]. Des groupes d'au moins 6 entrées sont créés à l'aide d'un simple algorithme d'agglomération et un score minimal de similarité (GrimScore $\geq$ 0.70) . Les entrées issues de la sc-PDB ont subi le même traitement que celles présentes dans le jeu Astex DiverseSet. Pour chaque entrée, la protéine et le ligand co-cristallisé sont enregistrés dans des fichiers mol2 séparés. Pour chaque ligand, au plus 250 conformères sont générés avec l'option '*FAST*' du générateur de conformères de CATALYST[32], puis sauvegardés au format sdf. La conformation cristallisée des ligands est retirée de l'ensemble conformationnel.

DUD-E :

Dix entrées (**Annexe 2.3**) représentant 5 familles de cibles importantes (récepteurs couplés aux protéines G, récepteurs nucléaires, protéine kinases, protéases, autres enzymes) sont extraites du jeu DUD-E[33] et traités de la même manière que les autres jeux de données.

3.2.2. Arrimage moléculaire (Docking)

L'arrimage moléculaire a été réalisé avec Surflex-Dock[34] (version 2.745). Un protomol[34] est généré depuis la liste des acides aminés, ions, cofacteurs et molécules d'eau dont au moins un atome lourd est distant de moins de 4.5Å du centre de masse du ligand. Le protomol est ensuite utilisé pour arrimer des conformations aléatoires du ligand à l'aide de paramètres standards de Surflex-Dock en utilisant l'argument "*–pgeom*" qui raffine la structure du ligand dans son environnement protéique après arrimage. Nous conservons uniquement la meilleure pose selon le score de docking calculé (*pKd*).

3.2.3. Recherche de pharmacophores protéine-ligand (RL-Pharm)

Le protocole RL-Pharm de génération de pharmacophore récepteur-ligand[7] de Discovery Studio v4.5 a été utilisé afin de générer les pharmacophores. Les éléments pharmacophoriques (accepteur, donneur, positif ionisable, négatif ionisable, aromatique et hydrophobe) sont alignés sur les atomes lourds du ligand et conservés uniquement si l'atome de ligand est en interaction directe avec la protéine cible (incluant les molécules d'eau liées) selon un ensemble de règles topologiques[7]. Un pharmacophore unique regroupant l'ensemble des éléménts pharmacophoriques retenus est sauvegardé au format chm.

Les structures 3D des ligands sont converties au format sdf depuis les fichiers mol2 en utilisant l'outil "*Convert*" (Molecular Networks, Erlangen, Germany) et utilisées comme entrées pour la génération de pharmacophores 3D en utilisant le protocole "*Generate Conformations*" de Discovery Studio. La méthode est configurée en mode *FAST*, elle génère un maximum de 100 conformères par ligand avec un seuil d'énergie de 20 kcal/mol par rapport au minimum global. La position initiale du ligand est supprimée afin de conserver un fichier au format sdf contenant un ensemble de conformères pour chaque ligand. Le protocole "*Screen Library*" de Discovery Studio est utilisé à des fins de criblage. Un maximum de 100 pharmacophores (toutes les combinaisons de 3 à 7 propriétés) est utilisé pour aligner le ligand en mode rigide. Seul le meilleur alignement (valeur de fit la plus élevée) est conservé.

3.2.4. Détermination de pharmacophores basés sur la cavité (IChem)

L'algorithme Volsite préalablement développé au laboratoire[14], a été implémenté au sein de la suite logicielle IChem[28] avec quelques améliorations. Premièrement, les atomes d'hydrogène sont ajoutés à l'aide de Protoss[30] afin d'optimiser le réseau de liaisons hydrogènes intra et intermoléculaires pour toutes les molécules présentes dans les fichiers de structure. Les propriétés pharmacophoriques (hydrophobe, aromatique, donneur, accepteur, positif ionisable, négatif ionisable, métallique) sont détectées à la volée en se référant au type atomique tout en considérant les ions, cofacteurs, molécules d'eau et groupes prosthétiques comme faisant partie de la protéine. Deuxièmement, les interactions hydrophobes sont redéfinies en utilisant des règles plus strictes que dans la précédente version. Les atomes pouvant réaliser une interaction hydrophobe sont restreints aux atomes de carbone et de soufre non liés à un hétéroatome ainsi qu'aux halogènes. La création de pharmacophores à partir de la cavité suit quatre étapes (**Figure 3.1**) :

1 – Détection de cavité à basse résolution : En utilisant le fichier de coordonnées de la protéine cible, nous déterminons les coordonnées extrêmes de la grille auxquelles sont ajoutées 8 Å dans chacune des 6 directions possibles. La grille ayant pour dimension la distance précédemment calculée est positionnée sur le centre géométrique de la protéine avec une résolution de 1.5Å, ce qui crée des cubes (voxels) de 3.375 $Å^3$ de volume. A chaque voxel est associé un point en son centre ainsi qu'une propriété. Si un atome lourd de protéine est situé à moins de 2Å du centre du voxel, celui-ci est considéré comme inaccessible et appartenant à la protéine (propriété "*IN*"). Pour le reste des voxels, nous vérifions l'enfouissement en générant, depuis le centre de ces coordonnées, un ensemble de 120 segments de 8 Å de longueur. Si le nombre de segments interceptant un voxel '*IN*' (*Nri*) est inférieur à 40, le voxel est considéré comme en dehors de toute cavité et se voit assigner la propriété "*OUT*".

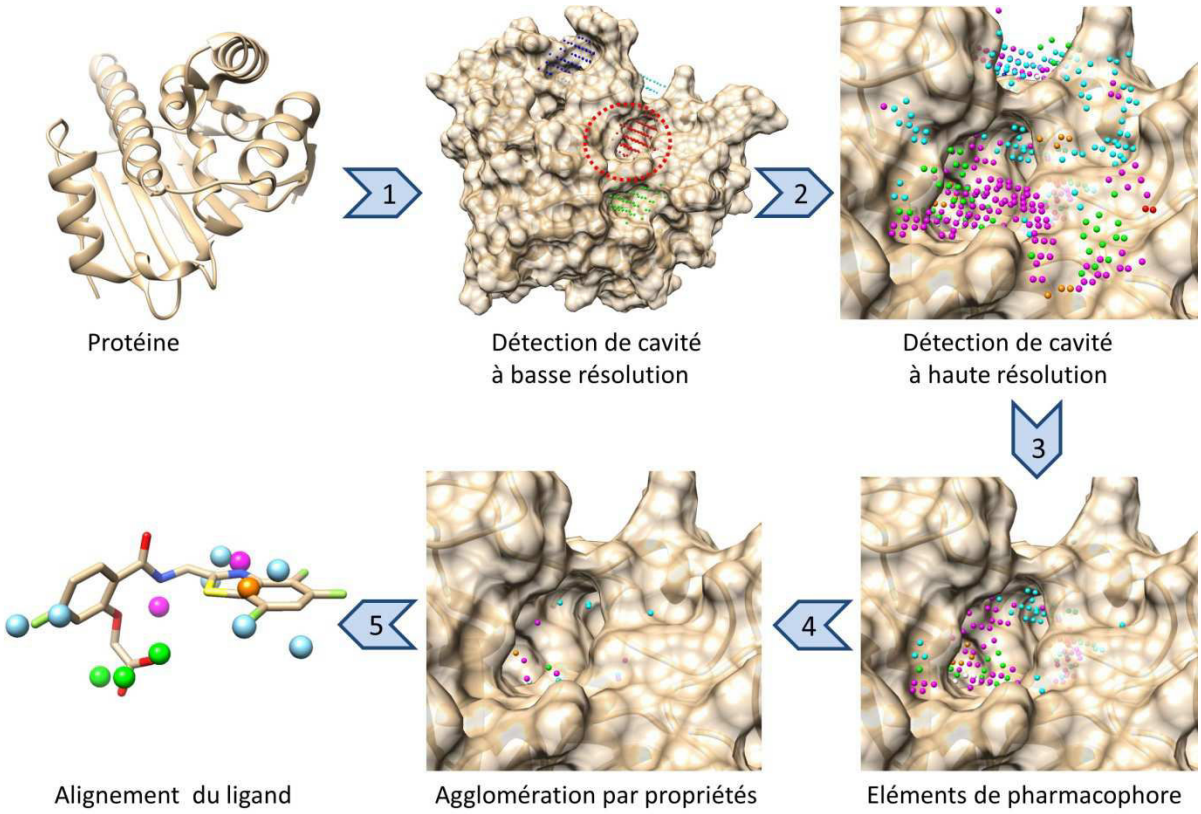


Figure 3.1. *Vue d'ensemble de la méthode.* ***1)*** *A partir d'une structure 3D de protéine sur laquelle les hydrogènes ont été rajoutés, les cavités sont automatiquement générées par VolSite et décrites comme un ensemble de points pharmacophoriques (points cyans, rouges et verts) ;* ***2)*** *Les cavités prédites droguables (cercle rouge) sont recalculées dans une grille plus fine (1 Å de résolution). Les caractéristiques pharmacophoriques (hydrophobe, cyan; aromatique, orange ; accepteur et négatif ionisable, vert; donneur et positif ionisable, magenta) sont assignées en fonction des atomes de protéines les plus proches ;* ***3)*** *Les éléments pharmacophoriques sont filtrés à l'aide de règles topologiques (enfouissement > 80/120, distance au centre de la grille < 8 Å) ;* ***4)*** *Agglomération des éléments pharmacophoriques;* ***5)*** *Alignement ligand-pharmacophore par recouvrement de formes et propriétés pharmacophoriques selon la méthode Shaper.*

Les voxels isolés (moins de 3 voxels adjacents utilisables) sont supprimés. La cavité est ainsi définie par les voxels restants. Ceux situés à moins de 4 Å d'un atome de protéine se voient attribuer une des 8 propriétés pharmacophoriques possible par complémentarité à l'atome de protéine le plus proche en suivant les règles d'interaction définies précédemment. Les voxels ne correspondant pas à ces règles topologiques se voient attribuer la propriété "*NULLE*" et ne seront plus utilisés ici. Les résidus de protéine servant à typer les voxels de cavité sont conservés dans un fichier site (format mol2). Nous attribuons enfin un score de droguabilité à toute cavité identifiée selon une machine d'apprentissage à vecteurs supports

définie précédemment[14]. Seules les cavités prédites droguables (score positif) sont utilisées pour la génération de pharmacophores.

2 - Représentation de la cavité à haute définition: Les cavités droguables sont recalculées avec deux modifications dans la procédure, (i) le centre de de la nouvelle grille de 20 Å de côté est placé sur le centre de gravité de la cavité basse résolution. (ii) La résolution de la grille passe à 1 Å pour une meilleure définition des points de cavité (6 000 voxels par cavité en moyenne). Chaque voxel se voit assigner une propriété pharmacophorique comme précédemment.

3 - Sélection et raffinement des éléments pharmacophoriques : Un pharmacophore idéal a été calculé pour chacun des 213 complexes protéine-ligand du jeu sc-PDBDiverse. Dans les pharmacophores idéaux, une propriété pharmacophorique n'est positionnée que sur l'atome de ligand en interaction directe (selon les règles d'IChem) avec la protéine cible. L'étude de la disposition spatiale de ces éléments de pharmacophores idéaux nous a permis de déterminer des valeurs seuils pour deux propriétés (enfouissement, distance au centre de la cavité) afin de réduire le nombre de points de pharmacophores finaux sans perte d'information cruciale. Les points pharmacophoriques au centre de voxels ayant un enfouissement inférieur à 80 ou étant à plus de 8 Å du centre de gravité de la cavité sont ainsi éliminés.

Les propriétés des points de pharmacophores restants sont ensuite raffinées de manière spécifique. Un point se verra assigné ainsi comme accepteur seulement s'il remplit la condition suivante : l'atome de protéine le plus proche est un donneur de liaison hydrogène avec un angle donneur-hydrogène-point compris entre 120 et 180 degrés. Les points pharmacophoriques ne vérifiant pas cette règle se voit assigner la propriété miroir du deuxième atome de protéine le plus proche et ainsi de suite (3^{e} plus proche, 4^{e} plus proche, etc…) jusqu'à ce l'ensemble des règles soit vérifié. Si aucune règle n'est validée (tous les atomes de protéines suivants sont à plus de 4 Å du point pharmacophorique par exemple), le voxel et le point pharmacophorique en son centre sont supprimés. Les règles d'assignation des points de pharmacophore aromatiques ont aussi été légèrement modifiées en tenant compte des plans des cycles des acides aminés aromatiques environnants. En plus de la distance au centre du cycle aromatique (<4Å entre le centre et le point), nous ajoutons une seconde distance seuil de 1.5Å entre le point pharmacophorique et un point situé à 4 Å du centre du cycle aromatique voisin sur une normale au plan de ce dernier et dans les deux directions. Cette définition nous permet ainsi de détecter des interactions aromatiques ligand-récepteur

face-face ou les deux normales aux plans aromatiques en interactions sont décalées d'au moins 1.5 Å. La dernière propriété modifiée concerne les points pharmacophoriques hydrophobes. La propriété hydrophobe est ainsi restreinte aux points de cavités dont l'environnement à moins de 4 Å contient au moins 50% de résidus hydrophobes (alanine, valine, leucine, isoleucine, proline, méthionine, phénylalanine, tyrosine, tryptophane) et pour lequel deux atomes de protéines hydrophobes sont distant de moins de 4 Å du point de cavité à assigner.

4 – Définition du pharmacophore final: Les points de pharmacophores restants sont regroupés par propriété en utilisant un algorithme d'agglomération hiérarchique et une distance seuil de 3.1 Å. Des sphères d'exclusions correspondant à des régions de l'espace occupés par la protéine et interdites au ligand sont enfin définies selon la méthode suivante. Une sphère unique est placée pour chaque acide aminé tapissant la cavité aux coordonnées atomiques correspondant au barycentre des atomes lourds les plus proches (< 4 Å) des points de pharmacophores. Leur diamètre est proportionnel au nombre d'atomes lourds utilisés pour les définir (1.15 Å pour 1 atome proche, 1.25 Å pour 2, 1.35 Å pour 3, 1.45 Å pour 4, 1.55 Å pour 5, 1.6 Å pour 6 et 1.7 Å de diamètre pour 7 ou plus d'atomes proches).

Les pharmacophores finaux (avec ou sans sphère d'exclusion) d'une taille moyenne de 35 éléments sont conservés au format chm de CATALYST[35] ainsi que sous la forme d'un fichier mol2. Le pharmacophore est décrit par les items suivants:

- la propriété : hydrophobe, aromatique, accepteur, donneur, négatif ionisable, ionisable positif ionisable, métallique.

- les coordonnées atomiques de la propriété (tête).

- un vecteur de 3Å de longueur dans la direction de la queue (accepteur, donneur, aromatique) dirigé vers l'atome de protéine complémentaire.

- les attributs spéciaux pour la propriété aromatique (centre, plan)

- sphères localisées sur les têtes et queues des vecteurs de rayon 1.6 et 2.2 Å, respectivement.

Les éléments à double propriété donneur et accepteur sont représentés par deux éléments séparés (donneur, accepteur) partageant les mêmes coordonnées atomiques en leur tête.

3.2.5. Alignement des ligands sur les pharmacophores (Shaper)

L'algorithme Shaper préalablement développé au laboratoire[14] a été utilisé pour aligner les atomes de ligands sur les points de pharmacophores. Shaper est un outil utilisant les bibliothèques "*OEChem*" et "*OEShape*" (OpenEye Sciencific Software, Santa Fe, U.S.A.) pour décrire les formes moléculaires par des gaussiennes et pour aligner deux objets moléculaires (atomes de ligands sur points de cavité) par maximisation de l'intersection des deux volumes correspondants[36]. Pendant l'alignement, les points de cavité restent fixes alors que les conformères du ligand subissent des translations/ rotations. Les meilleurs alignements sont ensuite triés par un score de couleur (les couleurs étant les propriétés pharmacophoriques) au moyen, d'un champ de force spécifique. Le champ de force customisé (**Annexe 2.4**) est composé de motif SMARTS pour décrire 10 propriétés pharmacophoriques de ligands (hydrophobe, cycle aromatique, cycle aliphatique, accepteur, donneur, accepteur et donneur, anion, cation et exclusion), 7 propriétés de points de cavité (hydrophobe, cycle, donneur, accepteur, accepteur et donneur, cation, anions) et 33 règles d'alignement afin de calculer le score de superposition des propriétés par similarité selon la métrique FitTverskyCombo:

$$FitTverskyCombo = \frac{OS_{C,L}}{0.15\ IS_C + 0.95\ IS_L + OS_{C,L}} + \frac{OC_{C,L}}{0.15\ IC_C + 0.95\ IC_L + OC_{C,L}}$$

$OS_{C,L}$ représente le volume commun aux formes du pharmacophore et du ligand, IS_C et IS_L les volumes non alignés, $OC_{C,L}$ est le volume commun aux couleurs du pharmacophore et du ligand, IC_C et IC_L les volumes des couleurs non alignées. Contrairement au score de Tanimoto qui met un poids égal sur les deux éléments alignés, le score de FitTversky donne un poids plus important (0.95) à l'objet mobile (le ligand). La métrique est asymétrique et varie entre 0 et 2.

L'alignement des ligands a montré quelques limites qui nous ont fait changer la méthode d'alignement (**Figure 3.2**).

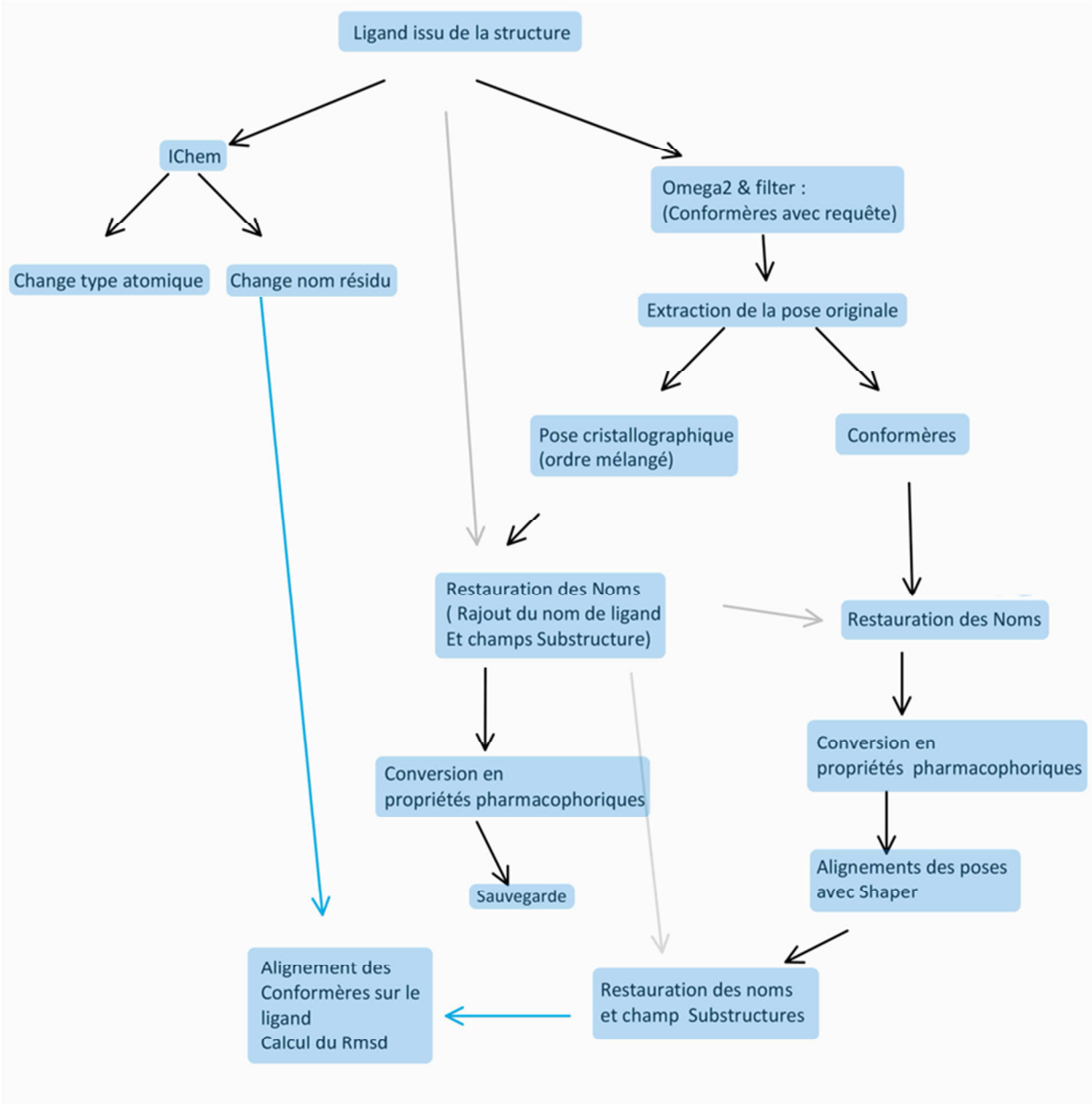


Figure 3.7. *Diagramme représentant les modifications apportées au ligand avant le processus d'alignement (flèche grise: lecture, flèche noire: modification, flèche cyan: alignement des coordonnées 3D). Chaque boite correspond à la création d'un nouveau fichier.*

Le fait d'utiliser un champ de force personnalisé a montré que l'interprétation des ligands et du pharmacophore par *OEChem* est erroné. Afin de pallier à ce problème, nous n'alignons pas les atomes de ligands mais des pseudoatomes décrivant les propriétés pharmacophoriques de celui-ci. Il en résulte un plus petit nombre de points à aligner. Ce procédé simple d'aspect demande plusieurs tâches d'harmonisation des noms, types atomiques et ordre de lecture. Les ligands sont d'abord traités par IChem afin d'avoir un fichier mol2 de référence avec un bon type atomique. Le conformères sont ensuite générés par le logiciel *Omega2* (OpenEye) La première pose est donc extraite et conservée en tant que référence. L'ensemble est converti en propriétés pharmacophoriques à l'aide d'IChem. On peut enfin aligner les conformères sur les pharmacophores et comparer les poses obtenues à la pose co-cristallisée si celle-ci est connue.

Les poses ainsi générées par IChem sont filtrées par le nombre de contacts répulsifs avec la protéine (clash1, distance seuil < 1.7Å; Clash2, distance seuil < 2.3Å) puis triées par score FitTverskyCombo.

3.3 Résultats et discussion

Le concept de pharmacophore, plus que centenaire[37], reste fréquemment utilisé à des fins de criblage virtuel afin d'identifier des molécules bioactives. La plupart du temps, le pharmacophore est défini à partir de ligands partageant un mécanisme d'action et une cible protéique. Plus récemment, le concept a été étendu à des pharmacophores déduits directement de structures cristallographiques protéine-ligand. Quand seule la structure 3D de la protéine est connue sans qu'aucun ligand n'ait pu être préalablement identifié, définir un pharmacophore simple et utilisable est plus difficile car cela impose une perception des éléments pharmacophoriques à partir de la seule connaissance d'un site de liaison. Ce procédé est très complexe: il requiert la détection de cavités droguables à la surface de la protéine cible, la détermination des régions de l'espace ou des atomes de ligands interagiront de manière optimale avec les points d'ancrages supposés les plus importants de la cavité.

Souvent le nombre de points pharmacophoriques générés *in situ* dans la cavité excède de loin le seuil toléré par des algorithmes d'alignement ligand-pharmacophore. Par conséquent, les points pharmacophoriques initiaux doivent être élagués de manière rationnelle, généralement à partir de cartes énergétiques, afin de conduire à un pharmacophore utilisable (< 10 points) à des fins de criblage virtuel.

Parmi les diverses méthodes de génération de pharmacophores à partir de structures de cavités, nombre d'entre elles reposent sur des calculs longs et complexes de dynamique moléculaire interdisant leur utilisation même à faible débit. Même s'il existe un effort récent pour simplifier les étapes de construction précédemment décrites, il y a toujours un besoin en un logiciel unique, rapide, fiable, automatisant le procédé entier, depuis la détection de la cavité jusqu'à la définition du pharmacophore final.

3.3.1. Détermination de pharmacophores à partir de cavités

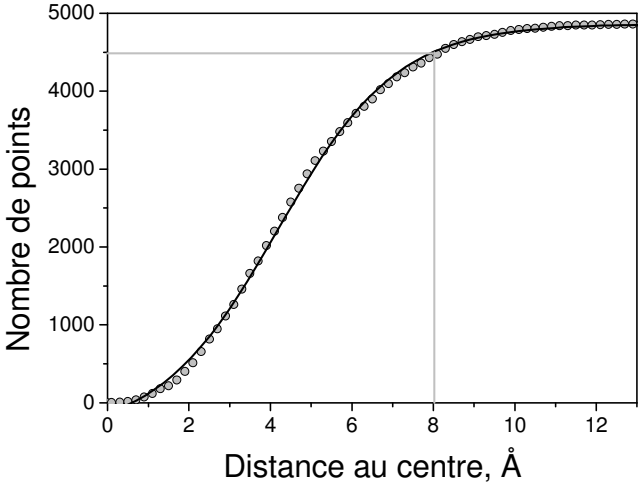
Les pharmacophores basés sur la cavité sont générés par un procédé en 4 étapes (**Figure 3.1**). Premièrement, les cavités potentiellement droguables sont déterminées à la volée depuis la structure de protéine en utilisant les paramètres standard de notre outil VolSite. La méthode centre la protéine dans une grille de résolution 1.5 Å et assigne les propriétés pharmacophoriques (hydrophobe, aromatique, donneur, accepteur, positif ionisable, négatif ionisable, chélateur de métal) aux différents voxels accessibles, en fonction

des atomes de protéines les plus proches. La droguabilité potentielle de chaque cavité détectée est déterminée par un modèle d'apprentissage à vecteurs supports (SVM) montrant une très bonne précision de classification (89%) sur des jeux de tests, par rapport aux méthodes concurrentes[14]. Pour chaque cavité prédite droguable, une nouvelle détection avec une résolution plus fine (1Å) est réalisée de manière à ce que la cavité soit centrée dans une grille de 20 Å de côté. Dans une troisième étape, les points de cavité obtenus sont élagués afin de diminuer considérablement leur nombre et de les transformer en points de pharmacophore.
L'algorithme VolSite précédemment publié a été modifié pour prendre en compte les positions explicites des hydrogènes. Le principal avantage d'utiliser les positions explicites des hydrogènes sur la protéine est l'optimisation de la définition des accepteurs de liaison hydrogène en suivant le vecteur correspondant (donneur-hydrogène-voxel). Nous avons aussi redéfini la définition des points de pharmacophore aromatiques afin de prendre en compte des interactions face-face décalées. La propriété hydrophobe a aussi été raffinée afin d'améliorer la qualité des pharmacophore générés. Au moins deux atomes hydrophobes de protéine doivent être distants de moins de 4Å d'un voxel hydrophobe, l'environnement immédiat doit être composé de plus de 50% de résidus hydrophobes).

La première conséquence de ces changements est que l'assignation des propriétés ne se fait plus forcément en une étape. Par exemple, un atome de protéine hydrophobe (carbone CB d'une alanine) ne peut réaliser une interaction hydrophobe au voxel le plus proche s'il ne vérifie pas les conditions précédemment énoncées; même s'l s'agit de l'atome de protéine le plus proche du voxel. Dans ce cas, la vérification des atomes suivants les plus proches est réalisée jusqu'à ce qu'un atome de protéine vérifie l'ensemble des règles. Si aucun atome ne convient, aucune propriété pharmacophorique n'est assignée au voxel correspondant

Une fois les points de cavités assignés, un élagage est réalisé par vérification de deux propriétés: l'enfouissement et la distance au centre. Des valeurs seuils pour ces deux propriétés ont étés définis par observation de points pharmacophoriques idéaux directement générés à partir de 213 complexes protéines-ligands de structure connue (**Figure 3.3**). L'examen de la distribution de ces deux propriétés pour 4871 points pharmacophoriques, montre que la distance au centre de la cavité est généralement inférieure à 8 Å alors que l'enfouissement du point pharmacophorique est très majoritairement supérieur à 80/120.

A



B

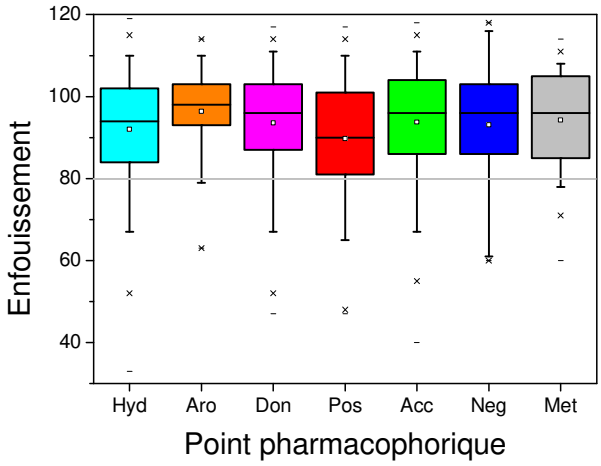


Figure 3.3 : *Propriétés des 4871 éléments pharmacophoriques générés à partir de 231 complexes protéine-ligand divers (sc-PDBDiverseSet); A) Distance de l'élément (en Å) au centre de la cavité, exprimée en nombre cumulé de points pharmacophoriques. La distribution cumulative suit une fonction sigmoïde de Boltzmann (R^2 = 0,999); B) Boites à moustaches de la distribution de l'enfouissement des caractéristiques pharmacophoriques (Hyd, hydrophobe, Aro, aromatique, Don, liaison hydrogène donneur; Pos, ionisable positif, Acc, liaison hydrogène accepteur, Neg: ionisable négatif; Met : Métal) exprimé par le nombre de projections de 8 Å (sur un total de 120) provenant du centre du voxel et intersectants des atomes de protéines. La boîte délimite les 25e et 75e percentiles, les moustaches délimitent les 5e et 95e percentiles. Les valeurs médianes et moyennes sont indiquées par une ligne horizontale et une case vide dans la boîte. Les croix délimitent les 1% et 99e percentiles, respectivement. Les valeurs minimales et maximales sont indiquées par un tiret*

L'application de ces filtres permet de réduire considérablement le nombre de points pharmacophoriques potentiels de 800 lors de la définition initiale à 300 si l'enfouissement est supérieur à 80, puis à 150 si distance au centre est inférieure à 8 Å (**Figure 3.4**)

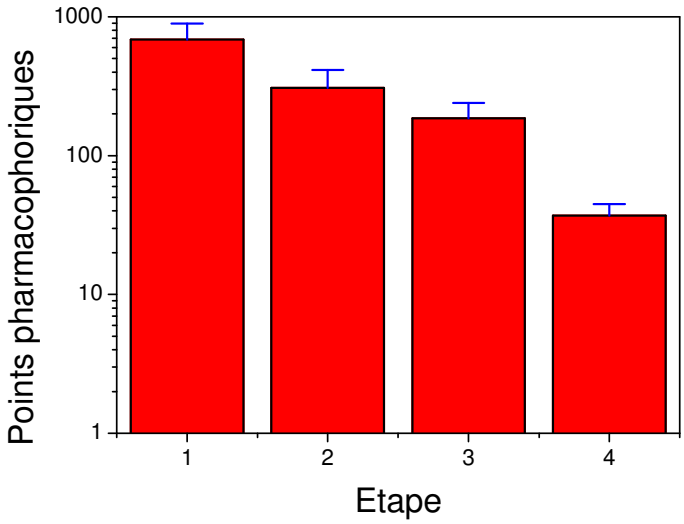


Figure 3.4: *Décroissance du nombre de points pharmacophoriques en fonction des étapes de notre protocole*

Au fur et à mesure des étapes de notre protocole automatisé, nous avons vérifié que des éléments pharmacophoriques importants n'étaient pas perdus en cours de route, en comparant les pharmacophores obtenus à chaque étape avec les pharmacophores idéaux obtenus directement à partir des 85 entrées du jeu de contrôle AstexDiverseSet (85 complexes). Nous avons considéré comme un succès le fait de placer un point de cavité à moins de 2 Å d'un vrai élément pharmacophorique de même type (**Figure 3.5**). Nous observons ainsi que le taux de récupération de vrais éléments pharmacophoriques décroit logiquement au fur et à mesure que le nombre de points de cavité diminue. Dans la définition initiale (800 points en moyenne), 95 à 100% des vrais éléments pharmacophoriques se superposent aux points de cavité, si ce n'est les points hydrophobes où on observe déjà une perte de 20% de vrais éléments pharmacophoriques observés sur les ligands du jeu AstexDiverseSet (Etape 1, **Figure 3.5**). La seconde étape de raffinement de la cavité n'influe que très peu les statistiques observées à l'étape 1. L'élagage puis l'agglomération des points de cavités en points pharmacophoriques finaux réduit les pourcentage de couverture aux alentours de 60% pour les points hydrophobes et 70-80% pour les autres propriétés pharmacophoriques (**Figure 3.5**).

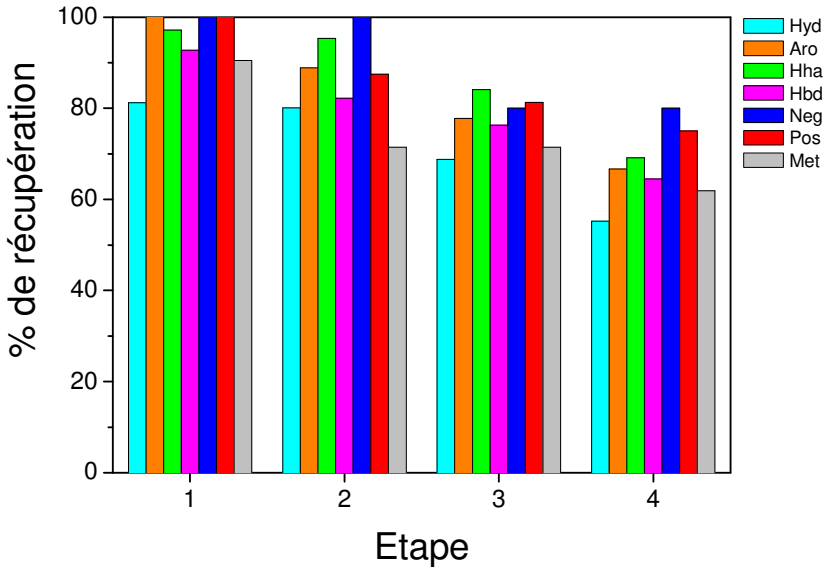


***Figure 3.5** : Pourcentage de récupération de propriétés pharmacophoriques idéales tout au long du protocole. Une caractéristique est définie récupérée s'il existe un voxel généré par IChem avec une propriété de même type que le vrai point pharmacophorique (set AstexDiverseSet) et distante de moins de 2 Å.*

Les points pharmacophoriques hydrophobes sont notoirement les plus difficiles à reproduire du fait du manque de directionalité rencontrée dans ce type d'interaction. L'examen des propriétés hydrophobes prédites par rapport aux vrais éléments pharmacophoriques observés sur les ligands du jeu de test, laisse apparaitre que le vrai point pharmacophorique n'est que rarement au centre d'un cluster de voxels hydrophobes. Il n'en demeure pas moins que les performances observées restent excellentes au vu du faible nombre de points pharmacophoriques (environ 30) obtenus en fin de protocole.

3.3.2. Alignement ligand-pharmacophore

L'alignement de ligands sur des pharmacophores composés en moyenne de 30 points s'est vite avéré impossible au moyen de méthodes classiques employant des sphères rigides, telle qu'implémentées dans la plupart des méthodes dont CATALYST. La combinatoire de superpositions possibles d'un ligand possédant en moyenne 20-30 atomes à un pharmacophore de 30 points est tellement grande qu'il est impossible de la satisfaire dans un temps raisonnable. Nous avons donc opté pour des méthodes de superposition plus floues et plus rapides en considérant non plus les atomes comme des sphères rigides mais comme des

Gaussiennes décrivant la forme des espèces moléculaires (ligands, points de pharmacophore) à aligner. Cette méthode a été développée dans le logiciel ROCS d'OpenEye[38] afin de comparer des ligands selon leur forme, avec énormément de succès.[39] Nous avons donc adapté le logiciel Shaper, préalablement développé au laboratoire pour aligner une paire de points de cavité, à l'alignement de ligand à des points de cavités[14]. Notre générateur de pharmacophore ainsi que la méthode d'alignement a été comparé, sur le jeu de test AstexDiverseSet de 85 complexes protéine-ligand, à deux méthodes concurrentes: l'arrimage moléculaire, et le criblage de pharmacophore protéine-ligand RL-Pharm (**Tableau 3.1**). L'arrimage moléculaire a été retenu comme une méthode standard de positionnement de ligands à partir de la simple structure d'une protéine cible. Ce sera notre référence. La superposition RL-Pharm à des pharmacophores protéine-ligand déduits de ces mêmes complexes, définit quant à elle la limite supérieure de précision possible pour toute recherche pharmacophorique.

Tableau 3.1. *Précision de positionnement de 85 ligands (Astex Diverse Set) à leur site de liaison par arrimage moléculaire (Surflex-Dock), alignement à un pharmacophore récepteur-ligand (RL-Pharm) et alignement à un pharmacophore de cavité (IChem Volsite)*

Méthode	**rmsd, Å**[a]	**% entrées <1 Å**	**% entrées <2 Å**	**% entrées < 3 Å**
Surflex-Dock	2.54	25	66	78
RL-Pharm	1.70	40	75	86
IChem-VolSite[b]	3.69	2	27	39
IChem-VolSite[c]	2.30	2	39	84

[a] écart quadratique moyen des coordonnées entre la pose prédite et la pose cristallographique
[b] pose correspondant au meilleur score FitTverskyCombo
[c] meilleure solution possible (plus petit rmsd à la pose cristallographique)

Le jeu de données AstexDiverseSet représente des complexes protéine-ligand sélectionnés manuellement et particulièrement difficiles à reproduire comme en témoigne la qualité des poses prédites par Surflex-Dock, un outil d'arrimage jugé parmi les meilleurs. Considérant uniquement la meilleure pose, l'écart quadratique moyen des coordonnées (rmsd) est de 2.54 Å avec 66% des ligands dont la pose est acceptable (rmsd < 2 Å, **Tableau 3.1**).

La méthode de recherche de pharmacophorique protéine-ligand RL-Pharm, que nous avons développé en partenariat avec Accelrys[7], produit logiquement les meilleurs résultats (rmsd moyen de 1.70 Å) dans la mesure ou les pharmacophores ont été précisément déduits des complexes protéine-ligand dont la structure est à reproduire. La qualité des résultats est donc essentiellement dépendante de la précision du générateur de conformères (la pose cristallographique ayant été omise dans le jeu de conformères) et de la qualité de la routine ajustant ces conformères au pharmacophore.

Les résultats obtenus par notre méthode (IChem VolSite) sont encourageants (rmsd moyen de 3.6 Å) mais encore inférieurs à ceux obtenus par arrimage moléculaire, notre référence pour la présente étude Si l'on considère l'ensemble des poses générées par notre méthode d'alignement (500 en moyenne), au moins l'une d'entre elles est proche de la solution cristallographique (plus petit rmsd de 2.3 Å en moyenne, **Tableau 3.1**). Il ne nous a malheureusement pas encore été possible de détecter ces poses correctes au moyen ni d'un score d'alignement ni d'une énergie d'interaction calculée par deux fonctions de score empirique (PLP, Chemscore).

3.4. Conclusion

La perception automatique de pharmacophores à partir de la simple structure de la cavité protéique d'intérêt reste un problème non résolu par des méthodes cheminformatiques automatisées et rapides. Alors que l'arrimage moléculaire permet de bien poser des ligands mais pas de les scorer, la recherche de pharmacophore basés sur les cavités pourrait permettre de répondre à cet inconvénient, étant donné que les recherches pharmacophoriques sont rapides et indépendantes de tout calcul énergétique. En étudiant les propriétés de vrais points de pharmacophore issus de complexes protéine-ligand co-cristallisés, nous avons pu mettre en évidence des filtres simples d'enfouissement et de distance au centre de cavité, nécessaires à la réduction du nombre de points pharmacophoriques générés automatiquement à partir de la structure de la cavité. Ces filtres, ajoutés à une agglomération finale des éléments pharmacophoriques nous permettent la définition automatique de pharmacophores simples (30 points en moyenne) sans perte importante d'information. Nous avons estimé à environ 20-30 le pourcentage de points pharmacophoriques effectivement perdus par les 4 étapes de notre protocole automatisé avec une exception notable pour les points hydrophobes où environ 40% des points sont perdus en cours de route. Cette déperdition explique en partie les résultats certes encourageants mais encore non satisfaisants obtenus lors de l'alignement final ligand-pharmacophore. Nous avons clairement à faire face à un problème d'évaluation quantitative de l'alignement car un jeu de bonnes solutions existe presque toujours, sauf pour les ligands très flexibles (> 10 liaisons de rotation), mais les diverses métriques d'évaluation de l'alignement que nous avons retenues ont été incapables de les identifier. Un modèle d'apprentissage (forêts aléatoires) basé sur les propriétés du ligand, de la cavité et de l'alignement ne permet également pas d'améliorer la qualité de la première pose retenue. Diverses pistes restent à étudier afin de distinguer les bonnes des mauvaises poses dans les solutions fournies. Tout d'abord, nous allons essayer d'utiliser un filtre supplémentaire à l'étape importante d'élagage en calculant au moyen d'une fonction empirique (PLP, Chemscore) l'énergie d'interaction du point pharmacophorique avec son environnement. L'observation de vrais points pharmacophoriques déduits de complexes protéine-ligand nous donnera la distribution et les valeurs seuils énergétiques à utiliser pour chaque propriété pharmacophorique. Nous espérons que la diminution du nombre de points pharmacophoriques à aligner permettra de réduire la proportion de faux alignements néanmoins bien scorés. Il est également possible d'assigner des poids différents dans le champ de force d'alignement aux points pharmacophoriques

polaires (accepteurs, donneurs, négatif et positif ionisables) afin d'éviter l'alignement simple de points hydrophobes sans conservation des couleurs. Il est enfin possible d'utiliser d'autres techniques d'alignement de Gaussiennes. La méthode que nous avons utilisée utilise une grille pour aligner les volumes des objets à comparer. Elle est recommandée par OpenEye pour sa vitesse et des objets de taille importante (> 20 points pour l'objet mobile) mais elle ne discrimine pas les couleurs pharmacophoriques (tous les points ont un rayon unique assimilés à celui d'un atome de carbone). Dans la mesure où nous pensons réduire le nombre de points de pharmacophores et que l'alignement du ligand ne se fait plus sur la totalité de ses atomes mais sur ses points pharmacophoriques déduits (en bien plus petit nombre), il nous sera possible d'utiliser une méthode exacte de recouvrement de points en testant systématiquement toutes les combinaisons possibles. Nous en attendons une meilleure qualité globale de l'alignement et un meilleur respect des couleurs pharmacophoriques.

Nous pensons être en mesure d'obtenir une qualité de pose voisine de celle obtenue par arrimage moléculaire. La qualité d'alignement et la fonction de score associée devront dans un second temps être vérifiées par criblage virtuel de la base de données DUD-E afin de calculer des taux d'enrichissement en vrais actifs lorsque ceux-ci sont mélangés à un grand nombre d'inactifs chimiquement similaires, ceci pour une dizaine de cibles d'intérêt pharmaceutique.

3.5. Bibliographie

1. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev*. **66,** 334–395 (2014).
2. Ripphausen, P., Nisius, B. & Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discov. Today* **16,** 372–376 (2011).
3. Spyrakis, F. & Cavasotto, C. N. Open challenFges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys*. **583,** 105–119 (2015).
4. Leach, A. R., Gillet, V. J., Lewis, R. A. & Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **53,** 539–558 (2010).
5. Plewczynski, D., Łaźniewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **32,** 742–755 (2011).
6. Wolber, G., Seidel, T., Bendix, F. & Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **13,** 23–29 (2008).
7. Meslamani, J. et al. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **52,** 943–955 (2012).
8. Salam, N. K., Nuti, R. & Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **49,** 2356–2368 (2009).
9. Koes, D. R. & Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* **40,** W409–414 (2012).
10. Wolber, G. & Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **45,** 160–169 (2005).
11. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28,** 849–857 (1985).
12. Verdonk, M. L., Cole, J. C. & Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **289,** 1093–1108 (1999).

13. Brenke, R., Kozakov, D., Chuang, G.-Y., Beglov, D., Hall, D., Landon, M.L., Mattos, C. & Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25,** 621–627 (2009).
14. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).
15. Ahlström, M. M., Ridderström, M., Luthman, K. & Zamora, I. Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J. Chem. Inf. Model.* **45,** 1313–1323 (2005).
16. Ortuso, F., Langer, T. & Alcaro, S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics* **22,** 1449–1455 (2006).
17. Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R. & Blundell, T. L. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **59,** 4314–4325 (2016).
18. Miranker, A. & Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **11,** 29–34 (1991).
19. Yu, W., Lakkaraju, S. K., Raman, E. P., Fang, L. & MacKerell, A. D. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *J. Chem. Inf. Model.* **55,** 407–420 (2015).
20. Hu, B. & Lill, M. A. Protein pharmacophore selection using hydration-site analysis. *J. Chem. Inf. Model.* **52,** 1046–1060 (2012).
21. Böhm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **6,** 61–78 (1992).
22. Schüller, A. et al. A pseudo-ligand approach to virtual screening. *Comb. Chem. High Throughput Screen.* **9,** 359–364 (2006).
23. Kirchhoff, P. D., Brown, R., Kahn, S., Waldman, M. & Venkatachalam, C. M. Application of structure-based focusing to the estrogen receptor. *J. Comput. Chem.* **22,** 993–1003 (2001).
24. Barillari, C., Marcou, G. & Rognan, D. Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J. Chem. Inf. Model.* **48,** 1396–1410 (2008).

25. Roland, W. S. U. et al. Snooker structure-based pharmacophore model explains differences in agonist and blocker binding to bitter receptor hTAS2R39. *PloS One* **10,** e0118200 (2015).
26. Johnson, D. K. & Karanicolas, J. Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein-Protein Interactions. *J. Chem. Inf. Model.* **56,** 399–411 (2016).
27. Hu, B. & Lill, M. A. Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking. *J. Chem. Inf. Model*. **53,** 1179–1190 (2013).
28. Desaphy, J., Raimbaud, E., Ducrot, P. & Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **53,** 623–637 (2013).
29. Hartshorn, M. J. et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **50,** 726–741 (2007).
30. Bietz, S., Urbaczek, S., Schulz, B. & Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminformatics* **6,** 12 (2014).
31. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites-10 years on. *Nucleic Acids Res*. **43,** D399–404 (2015).
32. Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment. (Dassault Systèmes, 2015).
33. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* ***55,*** 6582–6594 (2012).
34. Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des*. **21,** 281–306 (2007).
35. Kurogi, Y. & Güner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **8,** 1035–1055 (2001).
36. Grant, J. A., Gallardo, M. A. & Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **17,** 1653–1666 (1996).
37. Güner, O. F. & Bowen, J. P. Setting the record straight: the origin of the pharmacophore concept. *J. Chem. Inf. Model.* **54,** 1269–1283 (2014).

38. Rush, T.S., Grant, J.A., Mosyak, L., Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.*, **48**, 1489-1495 (2005)

39. Hawkins, P.C.D., Skillman, A.G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, **50**, 74-82 (2007)

3.6. Annexes

Annexe 3.1. *Jeu de données AstexDiverseSet de 85 complexes protein-ligand*

PDB ID	Ligand ID	Protéine
1g9v	RQ3	Hemoglobin alpha chain
1gkc	NFH	92 kda type iv collagenase
1gm8	SOX	Penicillin g acylase beta subunit
1gpk	HUP	Acetylcholinesterase
1hnn	SKF	Phenylethanolamine n-methyltransferase
1hp0	AD3	Inosine-adenosine-guanosine-preferring nucleoside hydrolase
1hq2	PH2	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase
1hvy	D16	Thymidylate synthase
1hwi	115	Hmg-coa reductase
1hww	SWA	Alpha-mannosidase ii
1ia1	TQ3	Dihydrofolate reductase
1ig3	VIB	Thiamin pyrophosphokinase
1j3j	CP6	Bifunctional dihydrofolate reductase-thymidylate synthase
1jd0	AZM	Carbonic anhydrase xii
1jje	BYS	Imp-1 metallo beta-lactamase
1jla	TNK	Hiv-1 rt a-chain
1k3u	IAD	Tryptophan synthase alpha chain
1ke5	LS1	Cell division protein kinase 2
1kzk	JE2	Protease
1l2s	STC	Beta-lactamase
1l7f	BCZ	Neuraminidase
1lpz	CMB	Blood coagulation factor xa
1lrh	NLA	Auxin-binding protein 1
1m2z	DEX	Glucocorticoid receptor
1meh	MOA	Inosine-5'-monophosphate dehydrogenase
1mmv	3AR	Nitric-oxide synthase, brain
1mzc	BNE	Protein farnesyltransferase beta subunit
1n1m	A3M	Dipeptidyl peptidase iv soluble form
1n2j	PAF	Pantothenate synthetase
1n2v	BDI	Queuine trna-ribosyltransferase
1n46	PFA	Thyroid hormone receptor beta-1
1nav	IH5	Hormone receptor alpha 1, thra1
1of1	SCT	Thymidine kinase
1of6	DTY	Phospho-2-dehydro-3-deoxyheptonate aldolase, tyrosine-inhibited
1opk	P16	Proto-oncogene tyrosine-protein kinase abl1
1oq5	CEL	Carbonic anhydrase ii
1owe	675	Urokinase-type plasminogen activator
1oyt	FSN	Thrombin heavy chain
1p2y	NCT	Cytochrome p450-cam
1p62	GEO	Deoxycytidine kinase

1pmn	984	Mitogen-activated protein kinase 10
1q1g	MTI	Uridine phosphorylase putative
1q41	IXM	Glycogen synthase kinase-3 beta
1q4g	BFL	Prostaglandin g/h synthase 1
1r1h	BIR	Neprilysin
1r55	097	Adam 33
1r58	AO5	Methionine aminopeptidase 2
1r9o	FLP	Cytochrome p450 2c9
1s19	MC9	Vitamin d3 receptor
1s3v	TQD	Dihydrofolate reductase
1sg0	STL	Nrh dehydrogenase [quinone] 2
1sj0	E4D	Estrogen receptor
1sq5	PAU	Pantothenate kinase
1sqn	NDR	Progesterone receptor
1t40	ID5	Aldose reductase
1t46	STI	v-kit hardy-zuckerman 4 feline sarcoma viral oncogene homolog
1t9b	1CS	Acetolactate synthase, mitochondrial
1tow	CRZ	Fatty acid-binding protein, adipocyte
1tt1	KAI	Glutamate receptor, ionotropic kainate 2
1tz8	DES	Transthyretin
1u1c	BAU	Uridine phosphorylase
1u4d	DBQ	Activated cdc42 kinase 1
1uml	FR4	Adenosine deaminase
1unl	RRC	Cyclin-dependent kinase 5
1uou	CMU	Thymidine phosphorylase
1v0p	PVB	Cell division control protein 2 homolog
1v48	HA1	Purine nucleoside phosphorylase
1v4s	MRK	Glucokinase isoform 2
1vcj	IBA	Neuraminidase
1w1p	GIO	Chitinase b
1w2g	THM	Thymidylate kinase tmk
1x8x	SO4	Tyrosyl-trna synthetase
1xm6	5RM	Camp-specific 3',5'-cyclic phosphodiesterase 4b
1xoq	ROF	Camp-specific 3',5'-cyclic phosphodiesterase 4d
1xoz	CIA	Cgmp-specific 3',5'-cyclic phosphodiesterase
1y6b	AAX	Vascular endothelial growth factor receptor 2
1ygc	905	Coagulation factor vii
1yqy	915	Lethal factor
1yv3	BIT	Myosin ii heavy chain
1yvf	PH7	Hcv ns5b polymerase
1ywr	LI9	Mitogen-activated protein kinase 14
1z95	198	Androgen receptor
2bm2	PM2	Human beta2 tryptase
2br1	PFP	Serine/threonine-protein kinase chk1
2bsm	BSM	Heat shock protein hsp90-alpha

Annexe 3.2. *Jeu de données sc-PDBDiverseSet de213 complexes protein-ligand*

Cluster	PDB	Ligand	Protéine	Entrées
0	10gs	VWW	Glutathione S-transferase P	18
1	1kjx	IMP	Adenylosuccinate synthetase	9
2	2r3a	SAM	Histone-lysine N-methyltransferase SUV39H2	25
3	2r3f	SC8	Cyclin-dependent kinase 2	6
4	3e5h	GNP	Ras-related protein Rab-28	171
5	13pk	ADP	Phosphoglycerate kinase, glycosomal	6
6	2fde	385	Protease	81
7	1v3s	ATP	Signaling protein	31
8	2fdp	FRP	Beta-secretase 1	66
10	1v45	3DG	Purine nucleoside phosphorylase	11
12	3orf	NAD	Dihydropteridine reductase	58
14	3orn	3OR	Dual specificity mitogen-activated protein kinase kinase 1	11
15	3oro	AGS	Serine/threonine protein kinase	65
16	2r4b	GW7	Receptor tyrosine-protein kinase erbB-4	9
18	3e65	XXZ	Nitric oxide synthase, inducible	8
20	3orz	BI4	3-phosphoinositide-dependent protein kinase 1	10
21	1klk	PMD	Dihydrofolate reductase	9
22	1a28	STR	Progesterone receptor	19
24	2r4f	RIE	3-hydroxy-3-methylglutaryl-coenzyme A reductase	17
28	1a2n	TET	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	9
32	1v79	FR7	Adenosine deaminase	6
34	2r4t	ADP	Glycogen synthase	13
36	2feq	34P	Prothrombin	76
37	2r59	PH0	Leukotriene A-4 hydrolase	7
38	1a42	BZU	Carbonic anhydrase 2	9
39	2r5c	C6P	Kynurenine aminotransferase	23
42	3otf	CMP	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4	18

45	1knr	FAD	L-aspartate oxidase	125
46	1knu	YPA	Peroxisome proliferator-activated receptor gamma	7
52	3ou2	SAH	SAM-dependent methyltransferase	174
53	1kol	NAD	Glutathione-independent formaldehyde dehydrogenase	75
54	2r6h	FAD	NADH:ubiquinone oxidoreductase, Na translocating, F subunit	32
55	2r6j	NDP	Eugenol synthase 1	6
58	1kor	ANP	Argininosuccinate synthase	21
59	2r6w	LLB	Estrogen receptor	17
68	2r7m	AMP	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5'-monophosphate synthetase	28
70	1a4z	NAD	Aldehyde dehydrogenase, mitochondrial	31
73	1v9n	NDP	Malate dehydrogenase	10
79	3ow3	SMY	cAMP-dependent protein kinase catalytic subunit alpha	8
80	3e7x	AMP	D-alanine--poly(phosphoribitol) ligase subunit 1	14
83	3e87	G95	RAC-beta serine/threonine-protein kinase	8
85	1vbm	YSA	Tyrosine--tRNA ligase	9
87	1vc2	NAD	Glyceraldehyde 3-phosphate dehydrogenase	52
88	3owa	FAD	Acyl-CoA dehydrogenase	22
89	3owb	BSM	Heat shock protein HSP 90-alpha	21
90	1vcf	FMN	Isopentenyl-diphosphate delta-isomerase	8
92	2r8o	T5X	Transketolase 1	32
96	1kp8	ATP	60 kDa chaperonin	21
100	3e8x	NAP	BH1520 protein	17
102	1kpg	SAH	Cyclopropane mycolic acid synthase 1	33
104	1a80	NDP	2,5-diketo-D-gluconic acid reductase A	30
105	3e92	G6A	Mitogen-activated protein kinase 14	21
107	1vdc	FAD	Thioredoxin reductase 1	28
111	3e9h	KAA	Lysine--tRNA ligase	10
116	4c4f	7CE	Dual specificity protein kinase TTK	8
118	3ox4	NAD	Alcohol dehydrogenase 2	7

119	1kqb	FMN	Oxygen-insensitive NAD(P)H nitroreductase	18
120	2r97	FMN	NAD(P)H dehydrogenase (quinone)	12
122	4c58	824	Cyclin-G-associated kinase	7
125	4c5o	FAD	Putative monooxygenase	30
126	4c61	LMM	Tyrosine-protein kinase JAK2	7
128	3oy1	589	Mitogen-activated protein kinase 10	7
130	3oy3	XY3	Tyrosine-protein kinase ABL1	28
133	1kqm	ANP	Myosin heavy chain, striated muscle	10
135	2r9r	NAP	Voltage-gated potassium channel subunit beta-2	10
136	1kqn	NAD	Nicotinamide mononucleotide adenylyltransferase 1	6
151	4c8g	C5P	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	9
156	2fky	N2T	Kinesin-like protein KIF11	17
162	4ca6	3EF	Angiotensin-converting enzyme	21
173	3p0n	BPU	Tankyrase-2	12
180	1vhn	FMN	tRNA-dihydrouridine synthase	28
181	2rd2	QSI	Glutamine--tRNA ligase	6
183	3ebh	BES	M1 family aminopeptidase	12
184	1vhw	ADN	Purine nucleoside phosphorylase DeoD-type 1	18
185	3p19	NAP	Putative blue fluorescent protein	91
189	4ccb	OFG	ALK tyrosine kinase receptor	8
192	3p23	ADP	Serine/threonine-protein kinase	39
195	2foi	JPA	Enoyl-acyl carrier reductase	22
199	1adc	PAD	Alcohol dehydrogenase E chain	17
200	4cdg	ADP	Bloom syndrome protein	16
202	4cdq	7VR	Polyprotein	6
203	3p3c	3P3	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	8
212	2fpt	ILB	Dihydroorotate dehydrogenase (quinone), mitochondrial	14
223	3eei	MTM	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	8
225	3eej	53R	Strain CBS138 chromosome J complete sequence	13

228	3p5s	AVU	CD38 molecule	7
232	3efq	714	Farnesyl pyrophosphate synthase	6
234	2rh1	CAU	Beta-2 adrenergic receptor	6
238	3p7n	FMN	Sensor histidine kinase	8
241	1aj2	2PH	Dihydropteroate synthase	6
242	2fsn	ADP	Archaeal actin homolog	15
246	2fsv	NAP	NAD(P) transhydrogenase subunit beta	9
247	1kyi	ATP	ATP-dependent protease ATPase subunit HslU	6
249	3p88	P88	Bile acid receptor	8
252	2fto	TMP	Thymidylate synthase	10
255	3p8x	ZYD	Vitamin D3 receptor	66
256	1akw	FMN	Flavodoxin	9
258	3ehg	ATP	Sensor histidine kinase DesK	26
259	3p8z	36A	RNA-directed RNA polymerase NS5	12
261	1kyx	CRM	6,7-dimethyl-8-ribityllumazine synthase	20
263	1am1	ADP	ATP-dependent molecular chaperone HSP82	14
264	3ehx	BDL	Macrophage metalloelastase	13
267	2fv9	002	Disintegrin and metalloproteinase domain-containing protein 17	6
269	2fvc	888	Genome polyprotein	12
274	4d86	ADP	Poly [ADP-ribose] polymerase 14	6
276	2rkg	AB1	Pol protein	11
278	3p9j	P9J	Aurora kinase A	11
279	2rku	R78	Serine/threonine-protein kinase PLK1	13
281	2rl5	2RL	Vascular endothelial growth factor receptor 2	13
295	4d9t	0JG	Ribosomal protein S6 kinase alpha-3	7
296	4d9w	X32	Thermolysin	13
303	1l2t	ATP	Uncharacterized ABC transporter ATP-binding protein MJ0796	12
307	1l4e	RBZ	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase	8
314	2g1n	1IG	Renin	8

319	2udp	UPP	UDP-glucose 4-epimerase	11
322	1aqb	RTL	Retinol-binding protein 4	6
323	3elj	GS7	Mitogen-activated protein kinase 8	10
325	3elm	24F	Collagenase 3	6
333	4dc3	2FA	Putative adenosine kinase	14
336	3en4	KS1	Proto-oncogene tyrosine-protein kinase Src	22
351	2uuo	LK3	UDP-N-acetylmuramoylalanine--D-glutamate ligase	8
354	1aux	AGS	Synapsin-1	7
355	3eos	PK2	Queuine tRNA-ribosyltransferase	12
356	3epp	SFG	mRNA cap guanine-N7 methyltransferase	14
360	3ept	FDA	Putative FAD-monooxygenase	10
364	4dfp	0L7	DNA polymerase I, thermostable	9
369	4dgm	AGI	Casein kinase II subunit alpha	8
378	1b0h	LYS_LYS_ALN	Periplasmic oligopeptide-binding protein	10
379	3pd3	A3T	Threonine--tRNA ligase	6
383	3eqp	T95	Activated CDC42 kinase 1	7
388	1b0p	TPP	Pyruvate-flavodoxin oxidoreductase	7
389	3erk	SB4	Mitogen-activated protein kinase 1	7
400	1vrw	NAD	Enoyl-ACP reductase	15
401	4dk5	0KO	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	7
407	1vso	AT1	Glutamate receptor ionotropic, kainate 1	9
411	4dko	0LM	Envelope glycoprotein gp160	6
412	1b3d	S27	Stromelysin-1	16
419	3peh	IBD	Endoplasmin homolog, putative	7
427	2ga2	A19	Methionine aminopeptidase 2	6
431	2uyy	NA7	Putative oxidoreductase GLYR1	13
432	4dlk	ATP	Phosphoribosylaminoimidazole carboxylase, ATPase subunit	15
434	1vtk	TMP	Thymidine kinase	9
444	1lhn	AON	Sex hormone-binding globulin	8

453	1lik	ADN	Adenosine kinase	6
454	2v0i	UD1	Bifunctional protein GlmU	10
459	1b9i	PXG	Putative UDP-kanosamine synthase aminotransferase subunit	7
475	3pjg	UGA	UDP-glucose 6-dehydrogenase	6
478	2v1u	ADP	ORC1-type DNA replication protein 1	11
488	3ewr	APR	Non-structural protein 3	6
501	4dqw	ATP	Inosine-5'-monophosphate dehydrogenase	13
502	3plq	RP2	cAMP-dependent protein kinase type I-alpha regulatory subunit	6
503	4dr9	BB2	Peptide deformylase	16
511	1w05	W05	Isopenicillin N synthase	20
512	4drx	GTP	Tubulin alpha chain	10
521	3exh	TPP	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial	19
532	1bif	AGS	6-phosphofructo-2-kinase	13
535	1bjy	CTC	Tetracycline repressor protein class D	7
538	1lvg	5GP	Guanylate kinase	7
540	2glx	NDP	1,5-anhydro-D-fructose reductase	7
551	3eyg	MI1	Tyrosine-protein kinase JAK1	16
567	1boo	SAH	Modification methylase PvuII	17
586	2gqt	FAD	UDP-N-acetylenolpyruvoylglucosamine reductase	14
598	2v6g	NAP	3-oxo-Delta(4,5)-steroid 5-beta-reductase	6
609	3ptq	NFG	OSIGBa0135C13.7 protein	8
616	3f3y	4OA	Bile salt sulfotransferase	6
621	2gtb	AZP	Orf1ab polyprotein	6
627	2v95	HCY	Corticosteroid-binding globulin	6
632	4dya	0MF	Nucleocapsid protein	6
640	1w7k	ADP	Dihydrofolate synthase	10
649	4e0i	FAD	Mitochondrial FAD-linked sulfhydryl oxidase ERV1	10
650	3f82	353	Hepatocyte growth factor receptor	8
665	1c1c	612	Reverse transcriptase/ribonuclease H	6

677	3pzb	NAP	Aspartate-semialdehyde dehydrogenase	9
685	1c30	ADP	Carbamoyl-phosphate synthase large chain	6
689	3fbu	COA	Acetyltransferase, GNAT family	7
704	3q0u	LL3	HTH-type transcriptional regulator EthR	6
759	1cbf	SAH	Cobalt-precorrin-4 C(11)-methyltransferase	6
765	2ha8	SAH	Probable methyltransferase TARBP1	11
768	2vfz	UPF	N-acetyllactosaminide alpha-1,3-galactosyltransferase	9
773	4e7z	ADP	Unconventional myosin-VI	10
785	1wkg	POI	Acetylornithine/acetyl-lysine aminotransferase	11
801	4eaw	0NQ	RNA-directed RNA polymerase	9
802	4eb3	0O3	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	7
826	3flk	NAD	D-malate dehydrogenase [decarboxylating]	8
885	2vna	NAP	Prostaglandin reductase 2	6
889	3qcf	NXY	Receptor-type tyrosine-protein phosphatase gamma	7
903	1mp3	TTP	Glucose-1-phosphate thymidylyltransferase	6
969	3fy4	FAD	(6-4)DNA photolyase	11
978	2hsd	NAD	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase	6
984	3qgz	ADN	Histidine triad nucleotide-binding protein 1	7
1031	2vw w	7X2	Ephrin type-B receptor 4	7
1061	3g5e	Q74	Aldose reductase	7
1073	3qov	ADP	Phenylacetate-coenzyme A ligase	8
1099	2w0j	ZAT	Serine/threonine-protein kinase Chk2	7
1163	1xoi	288	Glycogen phosphorylase, liver form	6
1202	3r04	UNQ	Serine/threonine-protein kinase pim-1	7
1260	3gjq	TRP_GLU_HIS_ASP_AC E	Caspase-3	7
1265	2we3	DUT	Deoxyuridine 5'-triphosphate nucleotidohydrolase	6
1271	1xwk	GDN	Glutathione S-transferase Mu 1	7
1310	4fhh	0U3	Vitamin D3 receptor A	9

1418	2wqo	VGK	Serine/threonine-protein kinase Nek2	6
1440	3rll	RLL	Androgen receptor	10
1453	4fsm	HK1	Serine/threonine-protein kinase Chk1	12
1505	1o6h	W37	Squalene--hopene cyclase	7
1717	4gfd	0YB	Thymidylate kinase	6
1719	4gfn	SUY	DNA gyrase subunit B	10
1801	4gpj	0Q1	Bromodomain-containing protein 4	6
1845	4gv2	5ME	Poly [ADP-ribose] polymerase 3	19
2170	3iub	FG2	Pantothenate synthetase	6
2615	4jd4	JDM	Dihydroorotate dehydrogenase (fumarate)	10
2716	1sqb	AZO	Cytochrome b	6
2898	4kfn	1QR	Nicotinamide phosphoribosyltransferase	8
3197	3zcm	PX3	Integrase	13

Annexe 3.3. *Jeu de données DUD-E (10 entrées)*

Protéine	**PDB**	**Ligand**	**DUD-E**	
			Actifs	Decoys
G Protein-Coupled receptors				
Adenosine A2A receptor (AA2AR)	3eml	ZMA	482	31498
Beta2 adrenergic receptor (ADRB2)	3ny8	JRZ	231	14993
Récepteurs nucléaires				
Androgen receptor (ANDR)	2am9	TES	269	14343
Glucocortocoid receptor (GCR)	3bqd	DAY	258	14987
Autres enzymes				
Adenosine deaminase (ADA)	2e1w	FR6	93	5449
Prostaglandin G/H synthase 2 (PGH2)	3ln1	CEL	435	23135
Proteases				
Angiotensin-converting enzyme (ACE)	3bkl	KAW	282	16860
Renin (RENI)	3g6z	A7T	104	6955
Protein kinases				
Fibroblast growth factor receptor 1 (FGFR1)	3c4f	C4F	139	8691
RAC-alpha protein kinase (AKT1)	3cqw	CQW	293	16426

Annexe 3.4. *Champ de force customisé pour l'alignement ligand-pharmacophore (Shaper)*

```
########################################
#        Pharmacophoric properties                    #
########################################
TYPE donor
TYPE acceptor
TYPE DonnAcc
TYPE cation
TYPE anion
TYPE rings
TYPE hydrophobe
TYPE ringAliph
TYPE exclusion

########################################
#        Pattern  Types                               #
########################################
PATTERN DonnAcc [15O;X0]
PATTERN rings [15C]
PATTERN hydrophobe [13C]
PATTERN acceptor [15O,14O]
PATTERN donor [15O,14N]
PATTERN cation [15N]
PATTERN anion [17O]

########################################
#      Interaction Definitions                        #
########################################

INTERACTION donor donor attractive gaussian weight=1.0 radius=1.0
INTERACTION acceptor acceptor attractive gaussian weight=1.0 radius=1.0
INTERACTION cation cation attractive gaussian weight=1.0 radius=1.0
INTERACTION anion anion attractive gaussian weight=1.0 radius=1.0
INTERACTION DonnAcc DonnAcc attractive gaussian weight=1.0 radius=1.0
INTERACTION DonnAcc donor attractive gaussian weight=1.0 radius=1.0
INTERACTION DonnAcc acceptor attractive gaussan weight=1.0 radius=1.0
INTERACTION donor cation attractive gaussian weight=1.0 radius=1.0
INTERACTION acceptor anion attractive gaussian weight=1.0 radius=1.0
INTERACTION cation donor  attractive gaussian weight=1.0 radius=1.0
INTERACTION anion acceptor  attractive gaussian weight=1.0 radius=1.0
INTERACTION anion cation  repulsive gaussian weight=1.0 radius=1.0
INTERACTION cation anion   repulsive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe donor   repulsive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe acceptor   repulsive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe DonnAcc  repulsive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe cation  repulsive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe anion   repulsive gaussian weight=1.0 radius=1.0
INTERACTION donor hydrophobe    repulsive gaussian weight=1.0 radius=1.0
INTERACTION acceptor hydrophobe  repulsive gaussian weight=1.0 radius=1.0
```

```
INTERACTION DonnAcc hydrophobe   repulsive gaussian weight=1.0 radius=1.0
INTERACTION cation hydrophobe  repulsive gaussian weight=1.0 radius=1.0
INTERACTION anion hydrophobe   repulsive gaussian weight=1.0 radius=1.0
INTERACTION rings donor   repulsive gaussian weight=1.0 radius=1.0
INTERACTION rings acceptor   repulsive gaussian weight=1.0 radius=1.0
INTERACTION rings DonnAcc  repulsive gaussian weight=1.0 radius=1.0
INTERACTION rings cation  repulsive gaussian weight=1.0 radius=1.0
INTERACTION rings anion   repulsive gaussian weight=1.0 radius=1.0
INTERACTION donor rings    repulsive gaussian weight=1.0 radius=1.0
INTERACTION acceptor rings  repulsive gaussian weight=1.0 radius=1.0
INTERACTION DonnAcc rings   repulsive gaussian weight=1.0 radius=1.0
INTERACTION cation rings  repulsive gaussian weight=1.0 radius=1.0
INTERACTION anion rings   repulsive gaussian weight=1.0 radius=1.0
```

Chapitre 4

Caractérisation des interfaces protéine-protéine de structure cristallographique connue

4.1. Introduction

Les interfaces protéine-protéine (PPI) sont au cœur du fonctionnement de la cellule et attirent de plus en plus l'industrie pharmaceutique afin d'identifier des petites molécules (le plus souvent des inhibiteurs) capable de les moduler sélectivement[1]. Etant donné la très grande complémentarité des protéines en interface, les modulateurs de PPIs constituent une nouvelle génération de candidat-médicaments très prometteurs de par leur très grande sélectivité pour une cible unique[2]. De très nombreuses bases de données ont été décrites afin de répertorier l'ensemble des PPIs au niveau génomique, protéomique et structural[1]. Ces bases de données[3] permettent notamment d'interroger l'interactome protéine-proteine (variant de 130 000 à 600 000 selon les estimations[4]) selon une pathologie et permettent une visualisation de réseaux d'interactions afin d'identifier les nœuds les plus importants. Au niveau structural, essentiel pour concevoir de manière rationnelle des modulateurs de PPI, peu de données sont disponibles. Trois bases de données de ligands[5–7] recensent de manière non-exhaustive les quelque milliers modulateurs de PPIs connus, la plupart ciblant un faible nombre de PPIs multi-investiguées (ex: p53-MDM2, BclXL-Bak). Encore plus tenues sont les données structurales sur les PPI cibles. La base de données 2P2Idb[5] recense 27 PPIs pour lesquelles une structure cristallographique est connue à la fois pour le complexe, les monomères à l'état apo (libre) ou holo (co-cristallisé avec un modulateur). Timbal[7] affiche 1695 structures cristallographiques de complexes protéine-protéine, protéine-inhibiteur et protéine-acide nucléiques. Il est plus que probable que des dizaines de milliers de PPI biologiquement importantes et de structure connue n'aient encore jamais été répertoriées de par la simple absence de ligands capables de les moduler.

L'objectif de ce travail est donc de réaliser pour la première fois une cartographie exhaustive de l'ensemble des PPIs de structure cristallographique connue au moyen de logiciels spécifiques capables de parcourir la Protein DataBank et de réaliser les opérations suivantes de manière complètement automatisée:

- Détection des interfaces;
- Prédiction de leur pertinence biologique;

- Recherche de cavités droguables à l'interface et au voisinage de celle-ci;
- Recherche de ligands occupant ces cavités droguables.

La Protein Data Bank (PDB) contient à ce jour près de 120 000 structures de macromolécules à haute résolution, dont la majorité est composée de protéines. Cela en fait la première source d'informations pour l'étude des interfaces entre protéines. Une structure tridimensionnelle est composé d'un ou plusieurs éléments structuraux, principalement des chaines peptidiques que l'on assimile à des protéines. Dès lors qu'il y a plusieurs chaines peptidiques dans une structure, il y a potentiellement une interface. L'amélioration de la qualité des structures (notamment leur résolution) rend l'étude de la centaine de milliers de contact inter-protéines possible, sachant qu'un ensemble non négligeable d'entre eux est biologiquement pertinent. La discrimination des contacts biologiquement viables est réalisable à l'aide de différents outils tels que PISA[8] ou EPPIC[9] mais leur domaine d'applicabilité ne couvre pas l'ensemble des interfaces présentes dans la PDB; la détection et caractérisation d' interfaces de petite taille restant problématique[10]. La caractérisation de l'interface comme biologiquement pertinente n'est pas la seule information à sauvegarder, il faut aussi différentier les propriétés phyisicochimiques clés des interfaces viables de celles des artefacts de cristallisation.

La modulation de PPI est possible de deux manières: (1) de manière orthostérique au moyen d'une petite molécule rentrant en compétition avec l'un des deux partenaires, (ii) de manière allostérique par une molécule capable de favoriser ou d'empêcher de manière indirecte la liaison des deux chaines protéiques. La première voie est de très loin la plus utilisée en chimie médicinale[1] mais se heurte au fait que les PPIs sont souvent plates et délocalisées sur une très grande surface, ce qui résulte dans la conception de molécules de très haut poids moléculaire avec un indice de développabilité clinique faible. La seconde voie a été très peu investiguée jusqu'à présent et a surtout été la conséquence de criblages aléatoires ou d'observations fortuites[2]. Nous ne privilégieront ici aucune des deux voies. De ce fait, nous allons répertorier l'ensemble des cavités droguables et leurs ligands où qu'ils se trouvent, soit à la surface d'un dimère, soit à la surface d'un des deux protomères en interaction.

Ce chapitre détaille la procédure d'analyse de la Protein Data Bank, il présente les statistiques obtenues puis développe une classification topologique des cavités présentes aux interfaces.

4.2. Matériel et Méthodes

Le traitement de la Protein Data Bank a été réalisé à l'aide de scripts écrit en python 2.7.11 et exécuté avec anaconda 4.0[11]. Les données finales seront stockées dans une base de données MariaDB[12]. L'ensemble des calculs utilisés ici a été réalisé grâce au centre de calcul de l'IN2P3 du CNRS.[13] Pour la suite de ce chapitre la Protein Data Bank sera nommée "PDB " et le fichier au format PDB sera cité comme la structure protéique.

4.2.1. Lecture des entrées PDB

Un total de 115 041 structures de protéines (contenu de la PDB au 1er juin 2016) a été utilisé pour détecter toutes les interfaces protéine-protéine de structures connues et qui pourraient être modulées par des molécules de faibles poids moléculaires, selon un protocole illustré en **Figure 4.1**. Une interface protéine-protéine est définie par un minimum de 20 interactions non covalentes entre deux chaînes peptidiques réparties sur au moins 10 acides aminés. Chaque molécule présente dans un fichier PDB (protéine, peptide, acide nucléique, ligand, solvant, ion, ajout de cristallisation) est caractérisée au moyen de critères simples se basant sur l'existence de noms de résidus. La protéine est composée d'aminoacides naturels reliés par une liaison peptidique puis de cofacteurs et d'ions importants pour le maintien de son intégrité structurale. Les brins d'acides nucléiques (ARN, ADN) n'ont pas été pris en compte dans cette étude. Une molécule est considérée comme un ligand lorsqu'elle remplit les conditions suivantes:

- poids moléculaire est inférieur à 500;
- structure de type organique, peptidique ou nucléotidique
- interactions moléculaires non covalentes avec son environnement
- surface exposée au solvant, à l'état lié, inférieure à 50%.

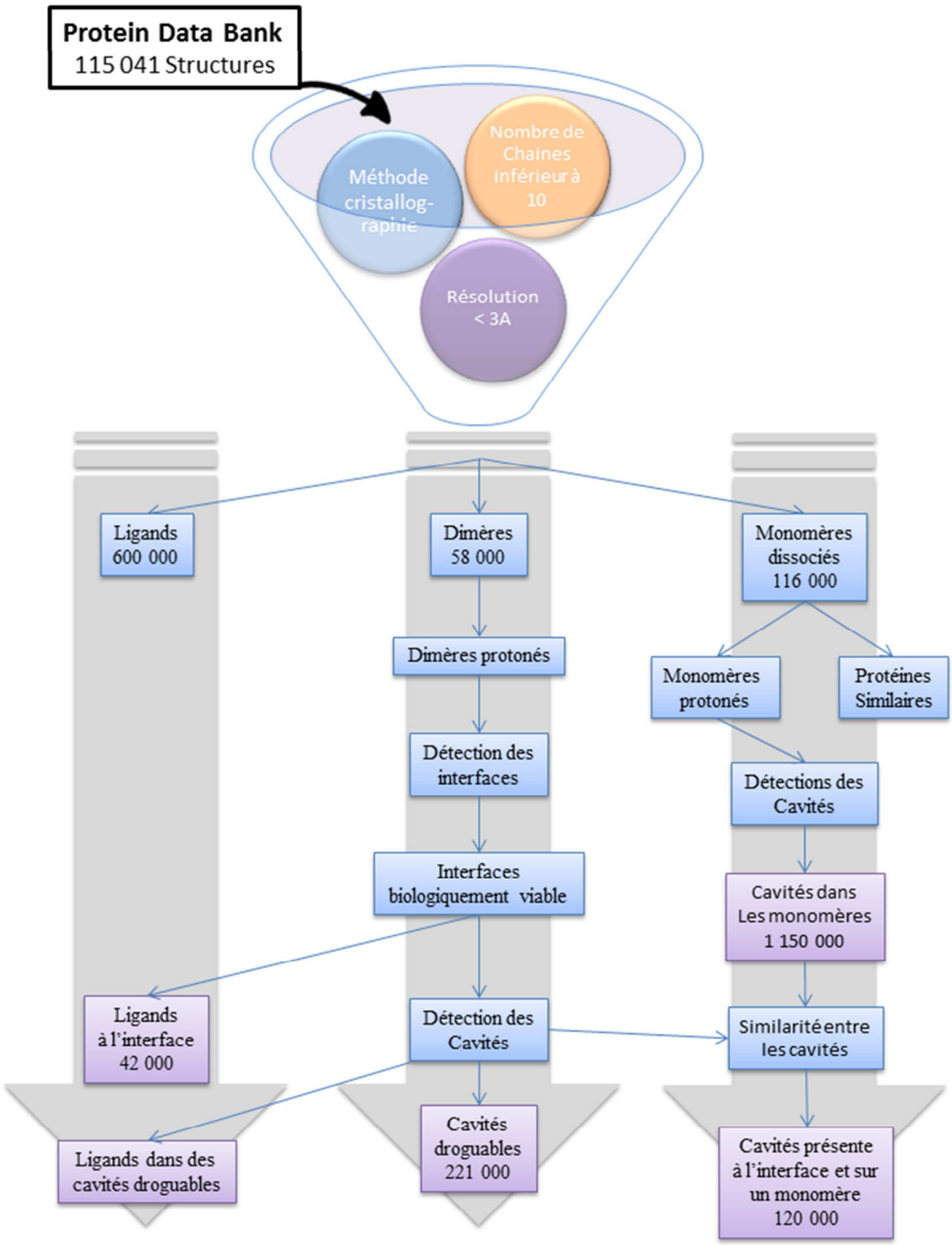


***Figure 4.8**: Diagramme présentant le processus de filtre et d'analyse de la PDB. Les structures parcourent l'ensemble du diagramme de haut en bas. Les différents scripts sont représentés en bleu et les productions finales sont en violet. Le nombre d'objets (ligands, cavités, monomères, dimères) est indiqué en bas de chaque catégorie.*

L'étude a été réalisée sur l' unité asymétrique fournie pour chaque entrée de la PDB, en parcourant les fichiers au moyen de BioPython 1.66.[13] Seules les structures composées de

2-9 chaînes protéiques ont été analysées. Les entrées plus complexes d'au moins 9 chaines (ex: PDB id 4U3M, ribosome 80S de levure) seront analysées plus tard afin de ne pas impacter le temps de calcul d'opérations déjà complexes sur de simples monomères (ex: détection de cavités). Les informations relatives à ces entrées seront donc jointes plus tard aux résultats décrits dans ce chapitre.

Les informations principales sont issues de l'entête du fichier de structure: le code PDB de référence de la structure, nom des protéines composant la structure, nom des chaines et données expérimentales. Lorsqu'il est présent, le numéro d'accession UniProt est extrait du fichier afin de récupérer des informations sur les chaînes de la structure. À ce stade sont éliminées les structures non obtenues par diffraction des rayons de X et ayant une résolution supérieure à 3Å.

4.2.2. Détection des interfaces biologiquement pertinentes

La détection d'interfaces se réalise sur des complexes dimériques (paires de chaînes), reconstitués de manière systématique à partir de chaque chaîne sauvegardée en fichiers distincts. Ainsi, une structure trimérique composée de 3 chaines peptidiques (A, B, C) donne naissance à trois structures de dimères possibles (AB, BC, AC). Pour chaque fichier de structure de dimère, les atomes d'hydrogènes ont été ajoutés à l'aide de Protoss[14], un outil récemment décrit qui optimise le réseau de liaison hydrogène intra- et inter-moleculaire au moyen d'une fonction de score empirique. De manière intéressante, Protoss considère la combinatoire des états d'ionisation et tautomériques possibles pour la totalité des atomes présents quelle que soit leur origine (protéine, solvant, cofacteur, ion, ligand). Le logiciel modifiant les en-têtes des fichiers PDB qui sont utiles pour la suite du processus, un script python récupère les en-têtes des fichiers originaux et les replace dans les sorties de Protoss. Chaque fichier est ensuite converti au format mol2 (TRIPOS Intl., St.Louis, U.S.A.) afin de contrôler le type de chaque atome présent. Lors de cette conversion certains acides aminés portant un atome de sélénium (sélénocystéine et sélénométhionine) sont standardisés et transformés en acides aminés naturels correspondants. Les coordonnées multiples des atomes ne sont pas tolérées et nous ne conserverons que les coordonnées dont le taux d'occupation est le plus élevé.

Les structures au format mol2 sont ensuite analysées afin d'identifier les interfaces entre deux chaines avec un outil (detectPPI) développé au laboratoire. Les interfaces sont détectées en deux temps: (1) les zones d'interactions sont d'abord déterminées en comptant le nombre d'atomes portés par des chaînes différentes, distants de moins de 5 Å, et délocalisés sur au moins 10 acides aminés; (2) si ce nombre est supérieur à 20, les interactions moléculaires non covalentes sont déterminées de manière explicite dans la zone de contact avec le logiciel IChem[15] également développé au laboratoire. Rappelons que les interactions entre deux chaines sont formalisées par des pseudoatomes (IPAs) placés à mi-distance des atomes en interaction et annotés par type d'interaction (hydrophobe, aromatique, liaison hydrogène, liaison ionique; **Figure 4.2**)

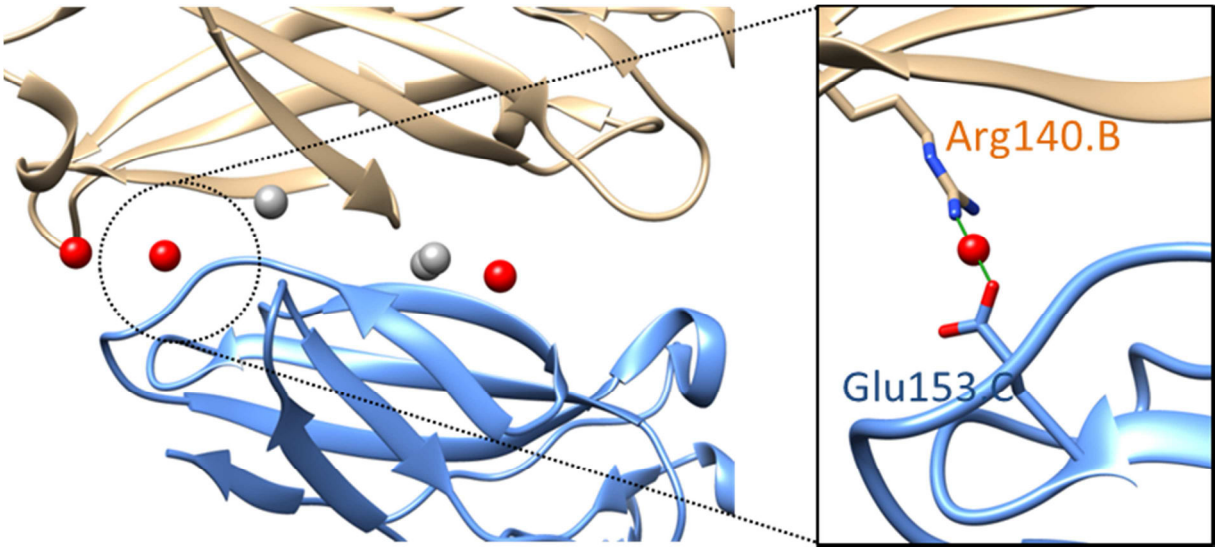


***Figure 4.2**: Exemple d'interface (PDB ID: 4NNY) entre la sous unité alpha du récepteur à l'interleukine-7 (rubans beige, chaîne B) et le cytokine receptor-like factor 2 (rubans bleus, chaîne C). Six pseudoatomes (spheres) sont places à mi-distance de chaque paire d'atomes en interaction et héritent d'une propriété correspondant à l'interaction correspondante. L'insert à droite est un zoom sur une liaison ionique (trait vert) montrant les chaines latérales en interaction (Arg140.B, Glu153.C) et le pseudoatome d'interaction (sphère rouge)*

La pertinence biologique de chaque interface est ensuite prédite avec le logiciel IChemPIC[10] préalablement décrit dans le chapitre 2, afin de ne conserver que les interfaces prédites comme biologiquement pertinentes.

4.2.3. Détection des cavités droguables

Détection

Une partie de ce travail de thèse a été consacrée à l'intégration puis la modification du programme Volsite[16] dans la suite IChem. Pour rappel, VolSite est un outil programmé en langage C++ permettant la détection de cavités à la surface d'une protéine d'intérêt, et au voisinage d'un ligand défini par l'utilisateur. La protéine est placée au sein d'une grille 3D de taille 20 Å, de résolution 1.5 Å, et centrée sur le ligand prédéfini (**Figure 4.3**).

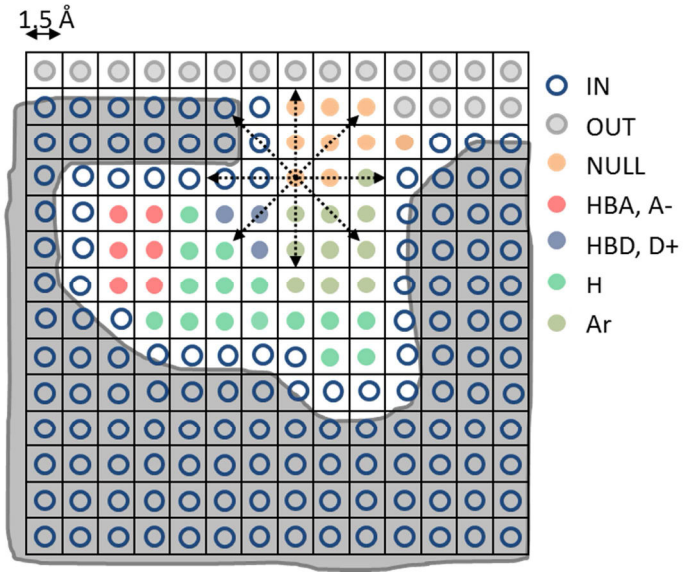


***Figure 4.3**: Grille tridimensionnelle (projetée ici en deux dimensions par souci de clarté) englobant la protéine cible (surface grise). Chaque voxel de 1.5 Å de côté est assigné une propriété (IN: protéine, OUT: extérieur). Pour ceux dont l'enfouissement est supérieur à une valeur seuil (50/120 par défaut), la propriété pharmacophorique complémentaire de celle de l'atome de protéine le plus proche leur est assignée (HBA: donneur de liaison hydrogène, A-: négatif ionisable, HBD: donneur de liaison hydrogène, D+: positif ionisable, H: hydrophobe, Ar: aromatique). Les voxels en dehors de la cavité ont une propriété nulle (NULL).*

La nature de chaque cube (protéine, cavité, extérieur) que l'on nommera ici voxel, est définie en fonction de sa localisation et de son enfouissement. Ce dernier est calculé en projetant, à partir du centre de chaque voxel, 120 rayons de 8 Å de longueur et en comptabilisant ceux interceptant un voxel contenant un atome de protéine. Une propriété pharmacophorique est ensuite attribuée à chaque voxel de cavité, propriété complémentaire de celle de l'atome de protéine le plus proche. Ainsi un voxel aura par example la propriété "donneur de liaison hydrogène" si l'atome de protéine voisin est un accepteur de liaison hydrogène. Un total de 8 propriétés pharmacophoriques (hydrophobe, aromatique, donneur, accepteur, accepteur et donneur, positif, négatif, nulle; **Tableau 4.1**) est ainsi répertorié pour chaque voxel, le centre de chacun d'entre eux délimitant une image inverse de la cavité accessible au ligand. Cette image inverse est sauvegardée au format mol2 permettant ainsi sa visualisation par n'importe quel logiciel de modélisation.

***Tableau 4.1**: Liste des propriétés de voxels délimitant une cavité*

Propriété	Nom	Atome de protéine voisin
hydrophobe	CA	hydrophobes*
aromatique	CZ	aromatique
accepteur	O	donneur
donneur	N	accepteur
accepteur/donneur	OG	accepteur/donneur
positif ionisable	NZ	négatif ionisable
négatif ionisable	OD1	positif ionisable
nulle	DU	aucun à moins de 4 Å

* au moins 2 atomes hydrophobes voisin (vor règles ci-dessous)

Mon travail a consisté à modifier spécifiquement le programme VolSite de manière à ce que l'ensemble des cavités présentes à la surface d'une structure protéique (monomérique ou multimérique) soient détectées à la volée, indépendamment de toute présence d'un ligand potentiel. Afin d'éviter la formation d'une multitude de petites cavités reliées entre elles par de très courts canaux, les cavités identifiées ont été assemblées avec un algorithme agglomératif en partant des voxels les plus enfouis jusqu'à ceux dont l'enfouissement reste supérieur à une valeur seuil de 50/120. Chaque voxel n'appartenant pas à la protéine est parcouru en commençant par le plus enfoui. Le premier voxel représente le premier cluster. Les voxels suivants sont ensuite assemblés de manière agglomérative à un des groupes préexistant au moyen de règles incluant le nombre de voxels, leur enfouissement et la densité du cluster.

La formalisation finale de la cavité se fait par un algorithme d'agrégation des clusters. La fusion entre deux clusters a lieu en prenant en compte la taille des clusters impliqués, la distance minimale entre eux ainsi que le gain de la fusion (**Tableau 4.2**).

***Tableau 4**.2: Règles d'agrégation des clusters de voxels*

Taille du cluster 1	Taille du cluster 2	Distance	Gain
1	1	Diagonale d'un voxel	/
1	5	Coté d'un voxel	0.1
5	15	Coté d'un voxel	1
15	15	Coté d'un voxel	1.5
15	30	Coté d'un voxel	2
30	30	Coté d'un voxel	2.5
70	70	Coté d'un voxel	4
/	/	Coté d'un voxel	8

Dans une grille standard, un voxel a 26 voisins. Le gain est le nombre de voxels voisins appartenant au second cluster divisé par le nombre maximal de voisins. Cette fonction a été paramétrée sur un ensemble de 80 protéines avec leurs ligands, le but étant d'optimiser le nombre, la taille et la forme des cavités. La modification de la méthode d'agglomération n'a pas engendré de modification majeure de la génération des cavités au niveau de voxels interceptant les atomes de ligands. Une fois les cavités créées, il a fallu corriger quelques erreurs dues à la méthode de calcul d'enfouissement des voxels qui tend à créer des cavités

creusées au niveau de leurs bordures extérieures. Un protocole de "lissage" de la cavité (**Figure 4.4**) est alors entrepris en parcourant à nouveau les voxels les moins enfouis mais avec de nouvelles règles. Un voxel est ainsi rajouté à la cavité, si celui-ci est relié de manière ininterrompue à au moins 11 voxels de cette cavité et en partageant au moins une face complète avec un des voxels voisins **(Figure 4.4)**.

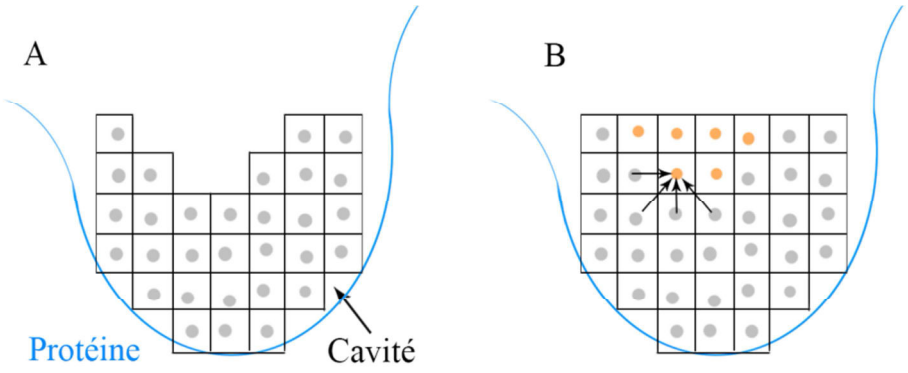


Figure 4.4: Procédé de lissage de la cavité représenté en 2D, A) Représentation de la cavité générée par l'ancienne version de VolSite; B) Cavité dont l'extrémité la moins enfouie a été lissée, en 2D. Les voxels orange sont rajoutés lors du lissage couche par couche.

Finalement le dernier critère important est la taille des agglomérats de voxels finaux. Un agglomérat sera définit comme cavité uniquement si celui-ci fait plus de vingt voxels.

Annotation pharmacophorique

L'atome de protéine le plus proche de chaque voxel de cavité n'est pas toujours le plus représentatif afin d'en assigner la propriété pharmacophorique. Par exemple, un voxel dont les 3 atomes de protéine les plus proches sont deux hydrophobes (distance de 2.9 et 3.1 Å) et un accepteur de liaison hydrogène (distance de 3.0 Å), se verra assigner la propriété "hydrophobe" bien que la propriété "accepteur" puisse être plus adaptée. La méthode d'assignation originale de VolSite[16] a donc été raffinée à l'aide de règles plus strictes. Les interactions polaires (liaisons ioniques et liaison hydrogènes) sont notamment priorisées même si un atome hydrophobe de protéine est le plus proche du voxel. Ainsi, si les conditions de distance de d'angles sont vérifiées, un voxel se verra attribué la propriété donneur ou accepteur de liaison hydrogène même si un atome de protéine plus proche est incapable d'engager une liaison hydrogène. La définition de l'assignation "hydrophobe" a aussi été renforcée. Une propriété hydrophobe de voxel n'est plus liée à un seul atome hydrophobe de

protéine voisin mais à un environnement dit hydrophobe. Un voxel ne se verra ainsi assigner la propriété hydrophobe qu'à trois conditions: (1) au moins la moitié des acides aminés environnants (< 5 Å) doivent être hydrophobes (définition selon une liste préétablie dans IChem), (2) au moins deux atomes hydrophobes de protéine se situent à moins de 5 Å de distance du centre du voxel, (iii) il n'y a pas d'atomes polaires de protéine présents à une distance comprise entre l'atome hydrophobe le plus proche et l'atome hydrophobe suivant, dans ce cas présent la propriété polaire est prioritaire.

Estimation de la droguabilité

La présence d'une cavité à la surface d'une protéine ne signifie pas qu'un ligand peut venir s'y fixer. La droguabilité ou ligandabilité[17] est une valeur arbitraire définissant si une cavité peut accueillir un candidat médicament avec une affinité suffisante. Dans VolSite, la droguabilité est estimée par une machine d'apprentissage à vecteurs supports (SVM) entrainée sur un jeu (NRDLD) de 113 structures de cavités (71 droguables et 42 non-droguables) représentées par une empreinte de 73 descripteurs[16]. La valeur de droguabilité est présentée sous la forme d'un réel variant autour de zéro, mais devant être interprétée de manière binaire. Une valeur positive indique que la cavité est prédite comme droguable. Par rapport à des algorithmes concurrents (Fpocket, SiteMap, DrugPred), VolSite présente les meilleures prédictions (précision de 89%) sur le jeu de cavités test issues de l'ensemble NRDLD[16].

Dans la mesure où nous avons légèrement modifié les règles de détection et d'annotation pharmacophoriques des voxels de cavité, nous avons refait un modèle d'apprentissage sur le même jeu de données (NRDLD) en validation croisé 5-fois au moyen du logiciel svm_light 6.02. Une grille de deux paramètres a été mise en place, elle optimise la valeur γ du noyau rbf (g) et la marge C du modèle. Ceux-ci ont été définis de la manière la plus exhaustive possible en variant γ de 0 à 1 par pas de 1.10^{-6} et C de 0 à 100 par pas de 1. L'optimisation a été réalisée sur le cluster de l'IN2P3 à l'aide de 200 nœuds. Le meilleur jeu de paramètres (γ =1e$^{10-6}$, C=100) conduisant à la meilleure valeur de F-mesure (F-mesure = 0.86) a été conservé pour prédire la droguabilité de nouvelles cavités. Il est à noter que cette nouvelle implémentation n'altère pas significativement la précision de la méthode (précision de 89% et 88 % pour l'ancienne et la nouvelle implémentation, respectivement). La valeur de droguabilité prédite à la volée par VolSite lors de l'examen de la PDB, sert de tri afin de ne conserver que les cavités droguables dont le score est positif.

Calcul de similarité

La similarité de deux cavités a été estimée par Shaper[16], un outil permettant l'alignement de forme des cavités de manière similaire au programme d'alignement de forme des ligands ROCS[18]. Chaque centre de voxel de cavité est ici considéré comme un atome typé pharmacophoriquement dont le volume est approximé par une Gaussienne. L'alignement de deux cavités se fait en optimisant le recouvrement des Gaussiennes, donc des volumes des deux cavités à comparer. L'alignement est dirigé vers la meilleure superposition possible des formes. Cet alignement est ensuite scoré pour le recouvrement des propriétés pharmacophoriques prédéfinies au moyen d'un champ de force spécifique (**Annexe 4.1**) définissant les propriétés pharmacophoriques, les règles d'alignement et les poids assignés à chaque alignement de propriétés. La qualité de l'alignement est estimée par la métrique TanimotoCombo:

$$TanimotoCombo = \frac{OS_{A,B}}{IS_A + IS_B + OS_{A,B}} + \frac{OC_{A,B}}{IC_A + IC_B + OC_{A,B}}$$

$OS_{A,B}$: volume commun aux cavités A et B alignées
IS_A, IS_B: volume des cavités individuelles A et B non alignées
$OC_{A,B}$: volume commun aux propriétés pharmacophoriques des cavités A et B alignées
IC_A, IC_B : volume des propriétés pharmacophoriques des cavités individuelles A et B non alignées
Le score de similarité TanimotoCombo varie de 0 (cavités entièrement dissimilaires) à 2 (cavités identiques)

4.2.4. Annotation de l'interface

Pour chaque PPI biologiquement pertinente, un certain nombre de propriétés attenantes à l'entrée PDB, les chaines en contact, la nature de l'interface et la localisation de cavités et de ligands sont répertoriées (**Tableau 4.**3)

Tableau 4.3: *Annotation de chaque interface*

Protéine	Chaînes	Interface	Cavité	Ligand
Code PDB	Code	Pseudoatomes d'interaction	Taille, $Å^3$	Code HET
Structure	Nom	Taille, Å2	73 descripteurs	masse
	N° UniProt	45 descripteurs	droguabilité	Distance à l'interface, Å
		Pertinence biologique	Enfouissement	Structure (mol2)
			Distance à l'interface, Å	
			Classification	
			Structure	

Chaque interface est décrite sous la forme d'un fichier de structure mol2 contenant les pseudo-atomes d'interactions (IPAs), un descripteur de 45 valeurs décrivant les interactions présentes à l'interface et leur valeur d'enfouissement. L'ensemble est référencé pour deux protéines données et un complexe unique. Les cavités détectées sont obligatoirement associées à une interface précise, elles sont aussi décrites sous formes de structure mol2 et d'un descripteur de 73 valeurs décrivant l'enfouissement des différentes interactions des différents types physico-chimique. Les différentes cavités et ligands sont classés en fonction de leur distance à l'interface, la distance étant la plus petite distance entre point de cavité/ atome de ligand et pseudoatome d'interface. Cavités et ligands ne sont conservés que si la distance à l'interface est inférieure à 8 Å.

4.3. Résultats

4.3.1. Composition en oligomères de la PDB

L'analyse exhaustive de la PDB réalisée au cours de ce travail répond à trois objectifs majeurs:

- analyser chaque interaction moléculaire à l'échelle atomique entre deux chaînes d'une même structure en se focalisant sur les interfaces prédites biologiquement pertinentes;
- identifier la totalité des cavités droguables à l'interface et au voisinage de celle-ci;
- caractériser les ligands occupant ces cavités.

La PDB est une base de données en constante évolution. A la date 1er juin 2016, nous avons pu analyser 115 041 entrées différentes. Nous avons uniquement sélectionné les structures obtenues par diffraction des rayons X car leur qualité est facile à analyser au moyen de descripteurs simplex (ex: résolution). Les entrées obtenues par RMN décrites sous forme d'ensembles ont été écartées afin de facilite l'analyse, de même que les entrées obtenues par d'autres méthodes (ex: microscopie électronique, diffraction de neutrons) souvent obtenues à basse résolution. La seconde étape de filtrage a consisté à ne retenir que les structures RX de résolution suffisante (< 3 Å) et oligomériques (2-9 chaînes).

Ces filtres ont rejeté 19 109 entrées obtenues par une méthode autre que la diffraction des rayons X, 36 407 structures monomériques, et enfin 2 182 structures possèdant au moins dix chaines protéiques (**Figure 4.5**). Les 57 339 entrées oligomériques restantes décrivent un total de 320 389 chaines soit une moyenne de 3 chaines par structure.

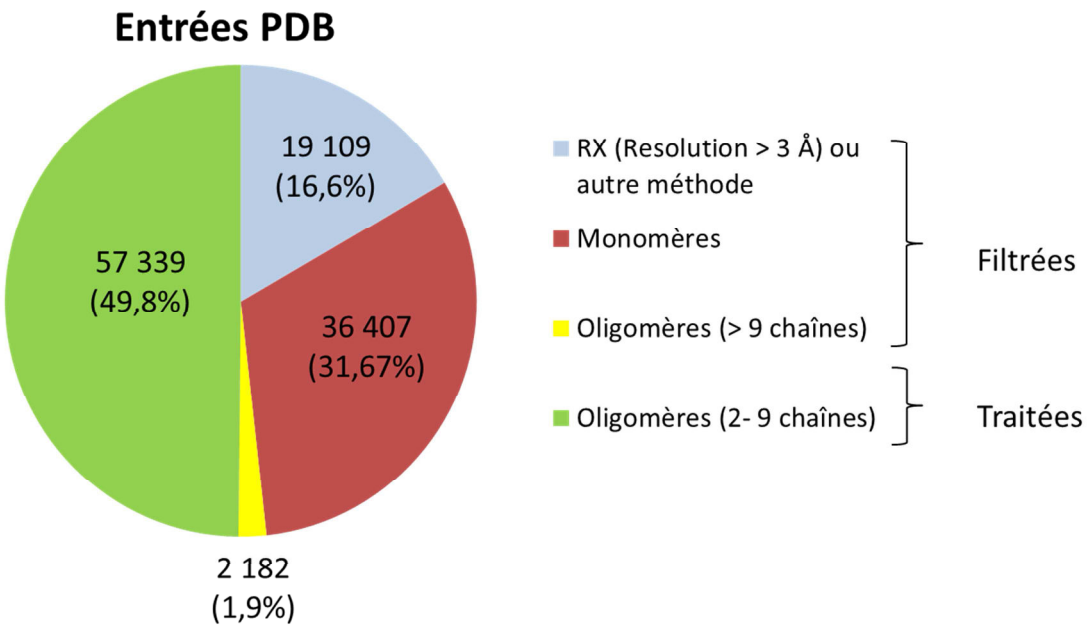


***Figure 4.5**: Filtrage des entrées par source expérimentale, résolution et niveau d'oligomérisation*

4.3.2. Détection des interfaces biologiquement pertinentes

L'intégralité des interfaces entre deux chaines a été détectée, l'ensemble des interactions non covalentes étant explicitement définie (contacts hydrophobes, interactions aromatiques, liaisons hydrogènes et liaisons ioniques). Pour chaque interaction, nous générons un pseudoatome (IPA) situé à mi-distance des atomes en interaction (**Figure 4.2**). Les interfaces complexes sont ainsi représentées d'une manière simplifiée à l'aide d'un ensemble d'IPAs (80 à 110 en moyenne) décrivant la nature et l'enfouissement des interactions correspondantes. Nous avons tout particulièrement été attentifs aux placements des liaisons hydrogènes grâce à l'ajout explicite d'atomes d'hydrogène sur l'ensemble des structures permettant à la fois une optimisation des états de protonation et des formes tautomériques possibles sur la totalité des molécules en présence (protéine, co-facteur, ligand, solvant, ion)

Sur les 295 128 interfaces détectées, 23 % d'entre elles ont été prédites biologiquement pertinentes par IChemPic (**Figure 4.6**).

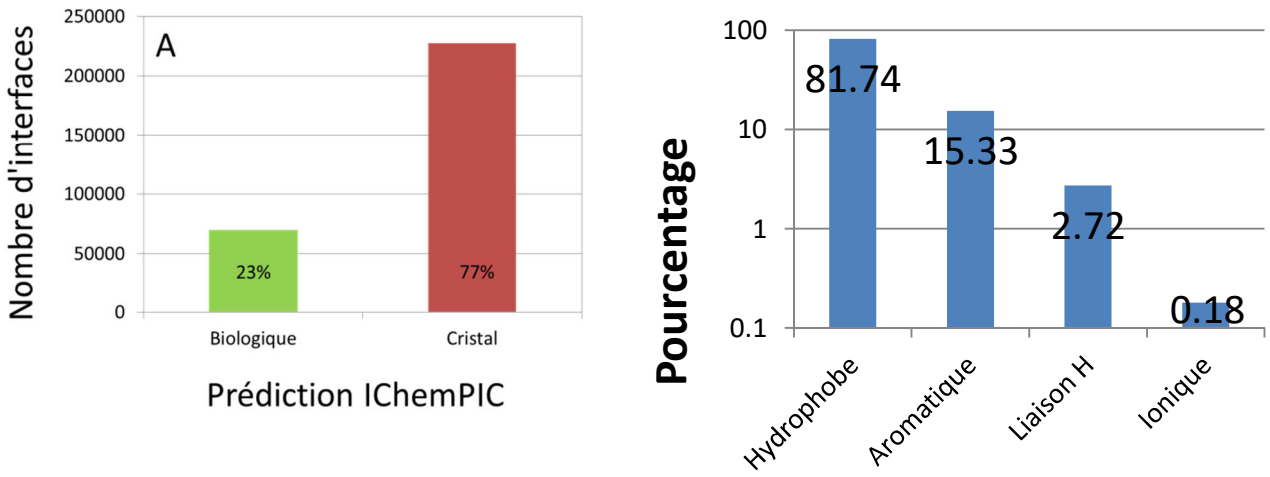


***Figure 4.6**. Interfaces prédites biologiquement pertinentes par IChemPIC. A) Représentation du nombre d'interfaces prédites biologiquement pertinente (biologique) et non pertinente (cristal) sur l'ensemble des contacts protéine-protéine présentes aux seins des structures multimériques de la Protein Data Bank. B) Distribution des interactions moléculaires (IPAs) observées aux interfaces biologiquement pertinentes.*

La très grande proportion d'interfaces cristallographiques sans pertinence biologique s'explique par la méthode combinatoire de définition des interfaces ou toutes les combinaisons possibles sont évaluées indépendamment de leur disposition spatiale relative. Il n'en demeure pas moins que 67 880 interfaces ont été prédites comme biologiquement relevante par notre modèle de forêts aléatoire IChemPIC (**Figure 4.6**). La distribution des interactions moléculaires observées à ces interfaces d'intérêt laisse apparaitre sans surprise une forte proportion d'interactions hydrophobes (80%, **Figure 4.6**).

4.3.3. Cavités droguables

Détection

Chaque interface prédite biologiquement pertinente est un site d'étude potentiel. Afin de déterminer celles que l'on veut étudier plus particulièrement, nous avons identifié les cavités présentes à la surface de tout dimère prédit biologiquement pertinent ainsi qu'en dissociant artificiellement le dimère en considérant chacun des deux monomères correspondants pris isolément. Cette approche nous permet de détecter des cavités à la surface du dimère mais également à la surface d'un des deux monomères avant association et en considérant un modèle de liaison entièrement rigide. Bien entendu, ce mode rigide d'association n'est pas pertinent pour tous les complexes. C'est pourquoi nous effectuerons plus tard une recherche de cavités sur des protéines uniquement monomériques (par exemple co-cristallisée avec un inhibiteur de PPI) mais préalablement identifiées comme faisant partie d'une interface biologiquement pertinente.

Sur l'ensemble des entrées présentes dans la PDB, qu'elle soient sous forme monomérique (dimère dissocié) ou dimérique, VolSite a détecté 2 180 354 cavités soit une moyenne de 40 cavités par entrée et 7 cavités par interface. Sur ces 2 180 354 cavités, 73% d'entre elles (1 595 113) ont un volume inférieur à 230Å^3 soit 70 voxels (**Figure 4.7**). Ces cavités de petite taille ont été éliminées de notre analyse car nous considérons qu'elles sont inadaptées à accueillir un candidat médicament potentiel. Cette valeur seuil de 230 Å^3 a été définie en considérant la distribution du volume des cavités droguables dans la base de données sc-PDB de sites protéine-ligand droguables[19] développée au laboratoire.

La droguabilité des 585 241 cavités restantes (27%) a été prédite par notre machine d'apprentisssage SVM incluse dans VolSite et laisse apparaitre une très forte proportion (380 746 soit 65%) de cavités réellement droguables (**Figure 4.7**). Il nous faut ici rappeler que ces cavités peuvent exister à la fois à la surface du complexe dimérique et de chacun des deux monomères, sous des formes identiques ou non.

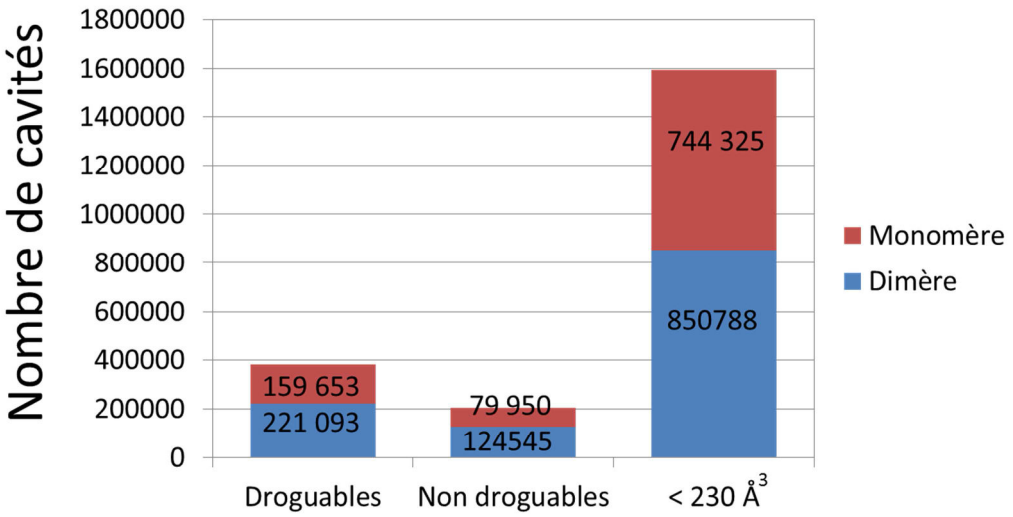


***Figure 4.7** : Histogramme empilé des différentes cavités identifiées à la surface des entrées traitées de la PDB possédant une interface biologiquement pertinente. Les cavités de petite taille (< 230 Å^3) ne sont pas soumises à une prédiction de droguabilité par VolSite. La quantité de chaque type de cavité est indiquée, soit à la surface d'un dimère soit du dimère dissocié (monomère)*

Classification

Nous avons classé les cavités droguables en 4 catégories en fonction de leur enfouissement dans les formes dimèriques (E_d) et monomériques (E_m), la contribution des monomères respectifs à leur constitution, et leur distance D_i à l'interface (**Figure 4.8**):

- Les cavités interfaciales sont présentes uniquement à l'interface du dimère et sont entièrement enfouies;
- Les cavités de bordure: formées par l'association de 2 monomères, elles sont accessibles chez le dimère mais s'ouvrent chez les deux monomères le constituant. Elles se retrouvent en périphérie immédiate de la zone d'interface proprement dite;
- Les cavités orthostériques sont uniquement accessibles lorsque les deux monomères sont dissociés, sont localisées à l'interface du dimère correspondant et accueillent la majorité des inhibiteurs de PPI co-cristallisés à ce jour;
- Les cavités allostériques: présentes indifféremment à la surface d'un des deux monomères et du dimère correspondant, elles se situent à distance de l'interface et ne changnte pas ou très peu de conformation lors de la formation du complexe dimérique.

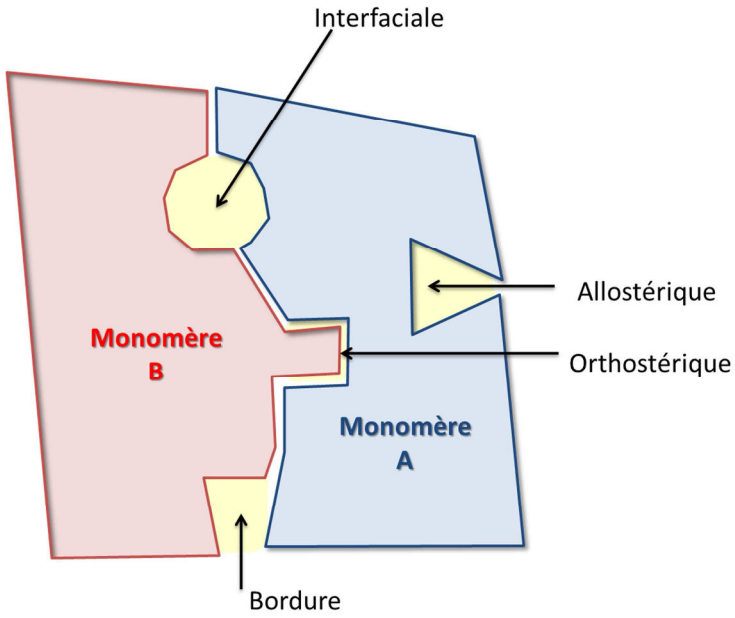


***Figure 4.8**. 4 types de cavités possibles à l'interface et au voisinage de celle-ci*

Afin de comptabiliser et de caractériser de manière plus approfondie chacune de ces quatre catégories, nous avons calculé les propriétés suivantes pour l'ensemble des 380 746 cavités droguables observée à la surface des dimères et de chacun des monomères les constituant:

- volume (Å^3) approximé par le calcul du nombre de voxels de cavités;
- enfouissement moyen à la surface du dimère (E_d) et des monomères séparés (E_m);
- la distance à l'interface (D_i) , plus petite distance entre un pseudo-atome d'interaction (IPA) et un centre de voxel de cavité.
- Similarité entre la cavité à la surface du dimère et celle à la surface d'un des deux monomères correspondant (TanimotoCombo).

La vérification de l'exactitude de ces catégories a été réalisée en plusieurs phases. Dans un premier temps nous avons analysé 10 cavités interfaciales connues[20]. Nous avons par la suite analysé 10 cavités choisies aléatoirement parmi toutes celles possédant des propriétés similaires (D_i < 3Å et E_d > 90). Sur ces 10 cavités de dimères choisies, une analyse visuelle nous confirme que 9 d'entre elles sont bien des cavités interfaciales, une seule étant une cavité de bordure (**Tableau 4.4**). Nous avons ensuite réalisé le même test sur 10 cavités choisies

aléatoirement mais présentant un enfouissement beaucoup plus faible (E_d < 70). 9 de ces 10 cavités sont des cavités de bordure, une seule étant interfaciale.

Pour compléter ces tests nous avons observé 10 cavités présumées allostériques (présentes sur un seul monomère, distance à l'interface > 3 Å, enfouissement < 90). Ces 10 cavités se sont avérées être réellement allostériques à l'interface cible bien que 3 d'entre elles soient également proches d'une autre interface biologiquement viable pour cette entrée PDB (**Tableau 4.4**). Enfin, 10 cavités dites orthostériques (présentes sur un seul monomère, distance à l'interface < 3 Å) ont été observées; ce sont toutes des cavités orthostériques dans lesquelles au moins un acide aminé du protomère manquant est entièrement enfoui.

Tableau 4.4 : *Propriétés de cavités choisies aléatoirement correspondants à la définition d'une cavité interfaciale (D_i < 3Å et E_d > 90).*

PDB	Proposition de Classification	Distance à l'interface, D_i	Enfouissement à la surface du dimère, E_d	Vérification visuelle
1h6d	Interfaciale	1,70	100	Interfaciale
1o94	Interfaciale	2,41	98	Interfaciale
1wd6	Interfaciale	2,35	100	Interfaciale
1x29	Interfaciale	2,6	103	Interfaciale
2c3c	Interfaciale	2,00	108	Interfaciale
2ev1	Interfaciale	1,90	95	Interfaciale
2p7o	Interfaciale	2,01	95	Interfaciale
2v7t	Interfaciale	1,63	103	Interfaciale
3mma	Interfaciale	2,06	96	Interfaciale
3mph	Interfaciale	1,90	98	Bordure
1k5m	Bordure	2,12	66	Bordure
4xdj	Bordure	2,31	67	Bordure
4fql	Bordure	2,40	69	Bordure
4cew	Bordure	2,95	68	Bordure
3pq6	Bordure	2,05	65	Bordure
1s78	Bordure	2,29	69	Bordure
1tju	Bordure	2,69	66	Bordure
3qhf	Bordure	2,56	68	Bordure

5dd6	Bordure	2,46	69	Bordure
3are	Bordure	2,90	70	Interfaciale
3vu7	Allostérique	12	75	Allostérique
1njj	Allostérique	10.3	77	Allostérique
3vbe	Allostérique	10.67	73	Allostérique
4in7	Allostérique	8.75	81	Allostérique*
4u5c	Allostérique	17.56	72	Allostérique*
3noc	Allostérique	20.3	71	Allostérique
3noc	Allostérique	21.07	87	Allostérique*
4kue	Allostérique	12.16	74	Allostérique
3sxn	Allostérique	19.41	77	Allostérique
3nog	Allostérique	21.4	71	Allostérique
1iy8	Orthostérique	0.74	81	Orthostérique
4drs	Orthostérique	0.86	72	Orthostérique
2vxx	Orthostérique	0.57	74	Orthostérique
3faz	Orthostérique	0.9	68	Orthostérique
3fa4	Orthostérique	0.56	71	Orthostérique
3ics	Orthostérique	1.69	76	Orthostérique
2p6t	Orthostérique	0.47	73	Orthostérique
4h80	Orthostérique	2.17	99	Orthostérique
1nw4	Orthostérique	1.56	79	Orthostérique
5fiw	Orthostérique	0.87	82	Orthostérique

** Allostérique pour l'interface cible mais impliquée dans une autre interface à la surface d'un autre dimère de cette entrée*

A la suite de ces observations nous avons déterminé un arbre de décision permettant un classement des cavités droguables (**Figure 4.9**) retrouvées à la surface des structures contenant des interfaces pertinentes. Le premier critère observé est la distance à l'interface: si une cavité est situé à une distance (la plus courte) supérieure à 3Å de l'interface, elle est classé comme allostérique.

Le deuxième critère vérifié est l'enfouissement moyen de la cavité: si l'enfouissement à la surface du dimère est inférieur à 90 (90 projections enfouies sur les 120 existantes), la cavité sera considérée comme une cavité de bordure accessible au solvent

Les cavités restantes sont donc proches de l'interface et très enfouies, elles sont classées en fonction de la participation d'un ou des deux protomères à la cavité. Si une chaine participe à la définition de la cavité, c'est une cavité orthostérique. Sinon, nous avons affaire à une cavité interfaciale (**Figure 4.9**).

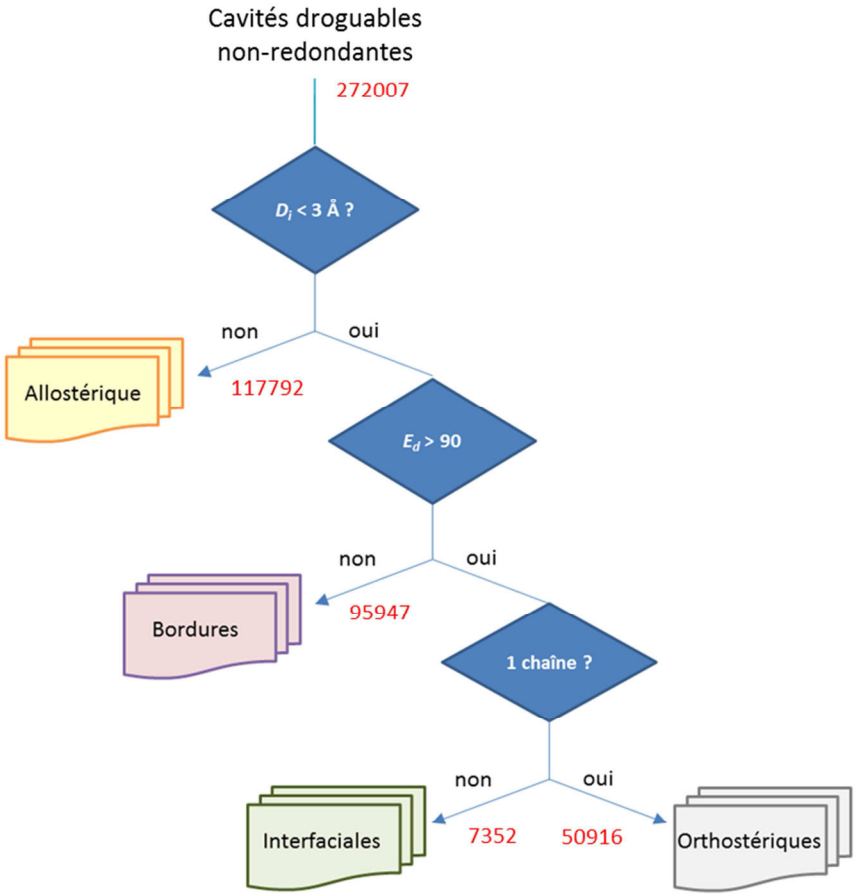


***Figure 4.9**: Classification des cavités droguables selon un arbre de décision*

Afin d'avoir un compte aussi exact que possible, nous avons éliminé les cavités redondantes observées à la fois à la surface du dimère et des monomères correspondants (TanimotoCombo (Shaper) >1.4) et avons ainsi classé 272 007 cavités en:

- 117 792 cavités allostériques;
- 95 947 cavités de bordure;
- 50 916 cavités orthostériques;
- 7 352 cavités interfaciales.

Les cavités dites allostériques ne sont certainement pas toutes exploitables pour une modulation de la PPI correspondante. Néanmoins, leur très grand nombre illustre leur potentiel complètement inexploité jusqu'ici. Le nombre de cavités droguables dites de bordure est également très important et conforte une étude préliminaire réalisée sur un petit ensemble d'entrées non redondantes de la PDB[22]. Ces cavités présentent également un potential interessant de par leur accessibilité à la surface de l'oligomère. Bien que nous n'ayons pas encore considéré les entrées PDB monomériques et leur cavités droguables (ex: co-cristallisées avec des inhibiteurs de PPI[21]), nous avons pu identifier plus de 50 000 cavités droguables orthostériques présentes à la surface d'un monomère et en interaction avec au moins un résidu du monomère partenaire. Ces cavités sont bien évidemment les plus interessantes à étudier pour l'identification d'inhibiteurs orthostériques. Enfin, nous avons pu identifier plus de 7000 cavités interfaciales complètement enfouies à la surface du dimère et représentant des cibles idéales à des stabilisateurs d'interface[20].

Volume

Les cavités droguables ont pour la plupart un volume inférieur à 350Å^3 (**Figure 4.10**) et restent donc très petites par rapport aux cavités droguables de la sc-PDB qui ont un volume moyen de 800 Å3 [16]. Ces cavités de faible taille, malgré qu'elles soient prédites droguables par notre modèle d'apprentissage sont probablement inexploitables à des fins de recherche de modulateurs d'interface. Un très grand nombre de cavités présentent néanmoins un volume idéal (entre 600 et 1000 Å^3) pour accueillir des modulateurs de haute affinité.

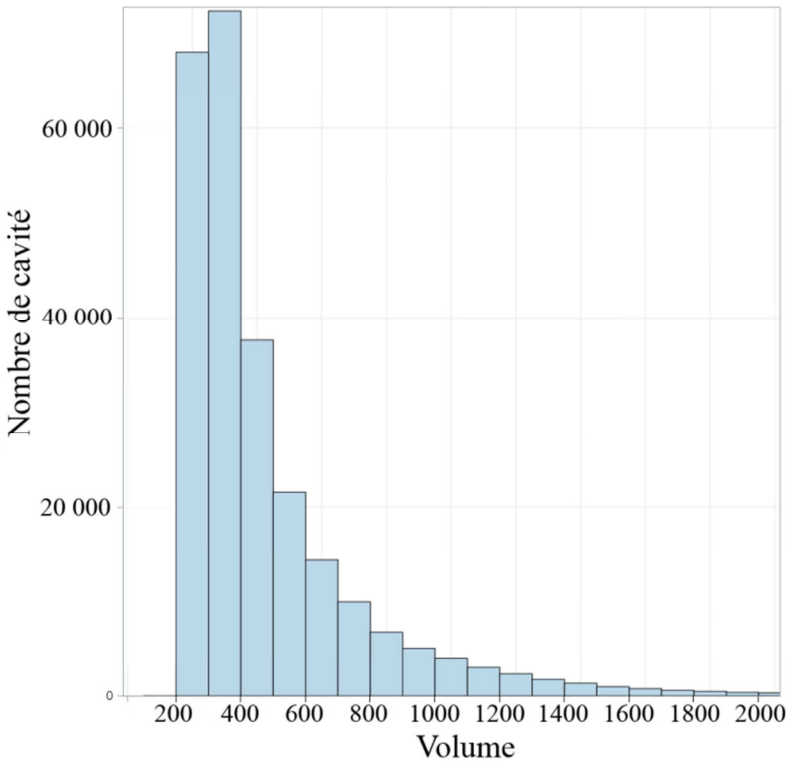


Figure 4.10: *Distribution des volumes de cavités (monomères, dimères)*

Distance à l'interface protéine-protéine

Les cavités droguables observées à la surface des dimères sont pour la plupart situées à l'extérieur de l'interface protéine-protéine (D_i > 2 Å). Leur distribution (**Figure 4.11**) montre néanmoins qu'un nombre significatif d'entre elles (environ 30%) se situe à une distance très proche de l'interface (1.7 > D_i > 2.3 Å). Ces dernières sont soit des cavités interfaciales, soit des cavités de bordure. Les cavités très éloignées de l'interface (D_i > 3 Å) sont des cavités allostériques.

Sur les complexes dissociés (**Figure 4.11**), les cavités sont réparties de manière plus homogène sur l'ensemble de la surface du monomère. Cependant, on observe toujours une concentration de cavités plus élevée à moins de 1.5 Å des interfaces, correspondant soit à des réminiscences de cavités de bordure, soit à des vraies cavités orthostériques.

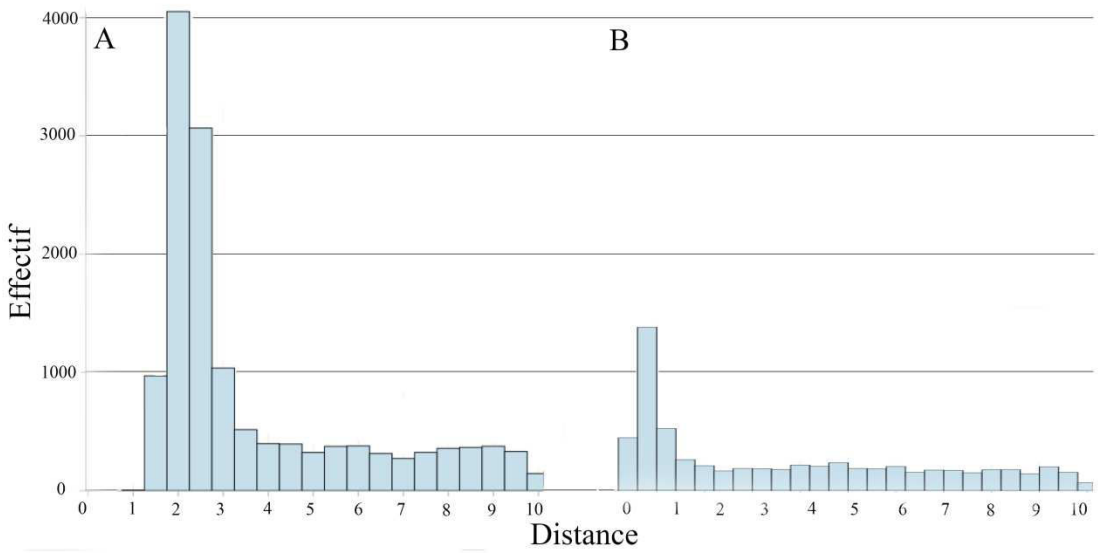


***Figure 4.11**: Distribution des distances à l'interface: A) dimères; B) monomères.*

Enfouissement moyen

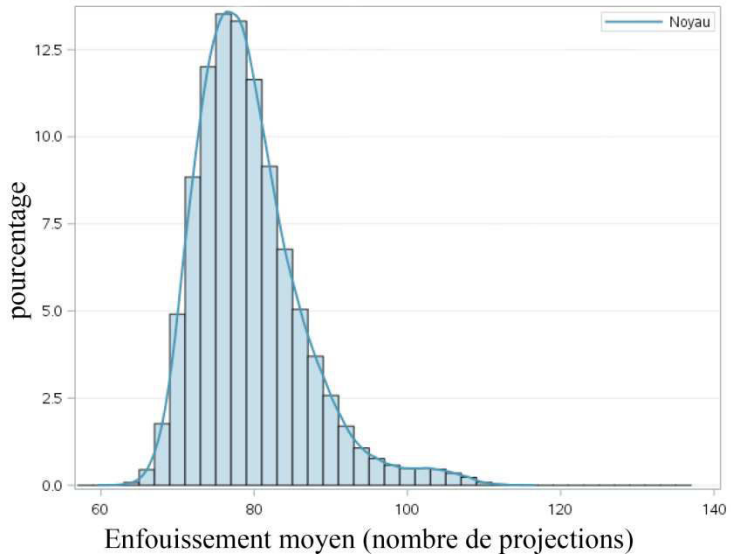


***Figure 4.12**: Distribution des valeurs d'enfouissement moyen de cavités (dimères)*

La distribution observée de l'enfouissment moyen est la même que pour les cavités droguables de la sc-PDB avec un pic à la valeur de 80 (**Figure 4.12**). En plus de cette distribution, on observe une seconde tendance plus légère entre les valeurs de 100 et 110. Ce second épaulement correspond aux cavités intefraciales très enfouies du faite de leur création par agglomération de deux chaines protéiques.

Distribution des propriétés pharmacophoriques

Chaque cavité étant caractérisée par les propriétés pharmacophoriques des voxels la délimitant, nous avons analysé la distribution des propriétés pharmacophoriques des cavités en fonction de leur classification, et par rapport aux cavités de sites de liaison protéine-ligand droguables de la base de données sc-PDB (**Figure 4.13**).

Les cavités interfaciales sont logiquement plus petites (volume moyen de 300 $Å^3$) et possèdent un caractère hydrophobe et aromatique marqué, se traduisant par des ligands interfaciaux de même nature. Les cavités de bordure sont les plus accessibles comme en témoigne la plus grande proportion de voxels à propriété nulle situés à plus de 5 Å de l'atome de protéine le plus proche. Les deux autres types de cavités (orthostériques, allostériques) ont des propriétés pharmacophoriques similaires à celles rencontrées dans les cavités de protéine globulaire (sc-PDB). La seule différence réside en un enrichissement des cavités sc-PDB en voxels donneurs de liaison hydrogène, résultant de la forte représentation en ligands nucléotidiques à forte densité d'atomes accepteurs de liaison hydrogène.

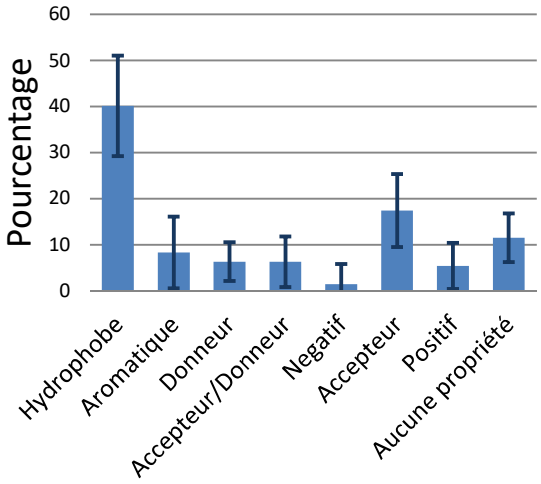


***Figure 4.13**: Distribution des propriétés pharmacophoriques pour les cavités drogguables*

Comparaison des cavités de dimères et des cavités de monomères

Même si nous travaillons sur des structures rigides, nous avons observé des modifications notables de cavités lors de la dissociation du dimère correspondant. Afin de les détecter, nous avons comparé systématiquement, pour chaque entrée, l'ensemble des cavités droguables observées à la surface du dimère, avec celles observées après dissociation sur les deux monomères correspondant (**Figure 4.14**). Les cavités droguables présentes aux

interfaces sont comparées avec Shaper[8], un outil basé sur le recouvrement de formes et de propriétés pharmacophoriques des points de cavités (centre de voxels).

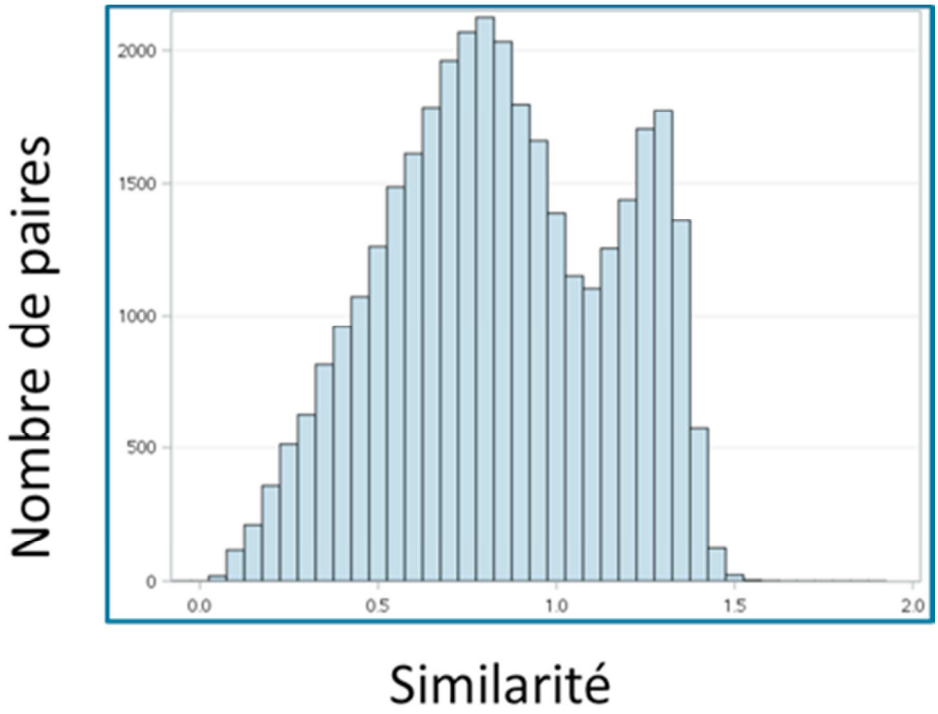


***Figure 4.14**: Similarité des cavités observées (score Tanimoto Combo), pour chaque entrée, à la surface de dimères et des monomères correspondants*

De manière très intéressante, les valeurs de similarité suivent une distribution bimodale avec deux pics centrés sur les valeurs de 0.75 et 1.30, respectivement. 3 203 paires de cavités ont un score de similarité supérieur au seuil empirique de 1,4 que nous avons fixé pour déclarer deux cavités comme similaires, sur la base de notre expérience passée à comparer des sites de liaison protéine-ligand droguables de la base de données sc-PDB. Ces cavités ne varient donc que très peu suite à la dissociation du dimère et correspondent à notre définition de cavité allostérique. Une très forte majorité des cavités présentes à la surface des dimères subissent un réarangement de forme très important suite à la dissociation des deux momonères, ce sont les cavités orthostériques, les cavités interfaciales et les cavités de bordure.

4.3.4. Ligands des cavités droguables

Nous avons enfin analysé systématiquement la localisation des ligands par rapport aux interfaces protéine-protéine sur les structures de dimères (**Figure 4.15**), en se focalisant sur les 700 ligands distants de moins de 8Å d'une interface biologiquement pertinente (plus petite distance entre pseudoatome d'interface et atome du ligand). On observe une distribution bimodale avec un pic à 1Å de distance à l'interface et un second pic plus important à une distance de 5 Å. Il est probable que le premier pic correspond aux ligands interfaciaux et le second aux ligands de bordure ainsi qu'aux ligands allostériques.

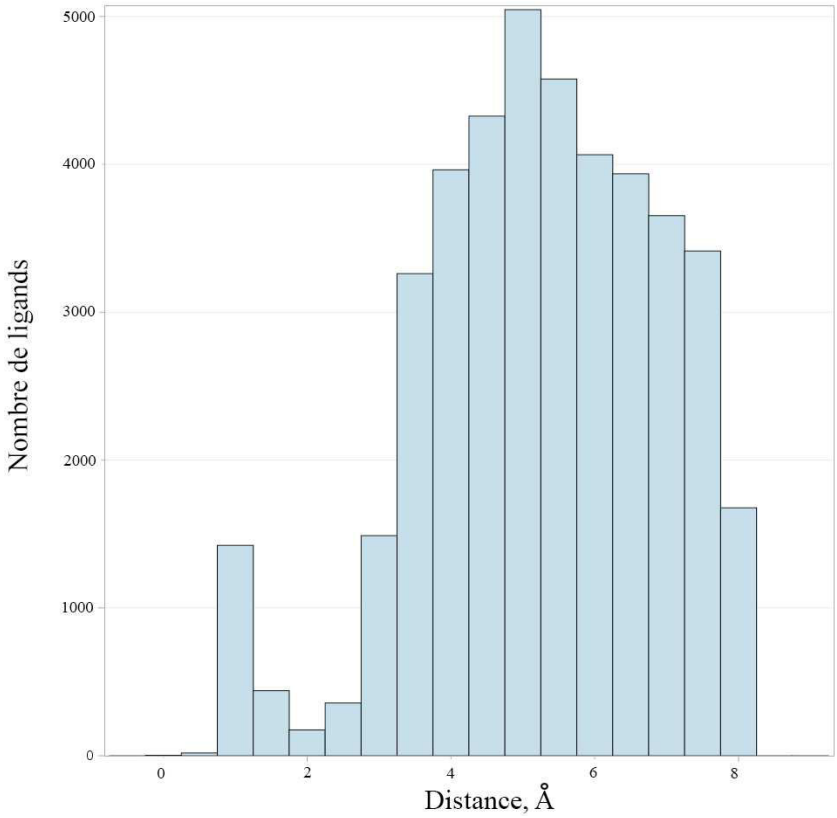


***Figure 4.15**: Distances observées entre ligands et interface protéine-protéine*

Une analyse plus fine nous démontre que les cavités contenant un ligand co-cristallisé sont composées à 90% de cavités de bordures, 9.9% de cavités interfaciales et 1% de cavités allostériques (**Figure 4.16**). La très faible proportion de ligands de cavités allostériques vient

de notre parti pris de ne considérer que des ligands distants de moins de 8 Å de l'interface la plus proche.

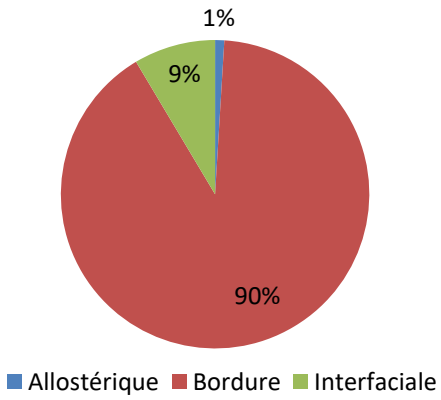


Figure 4.16 Distribution des cavités co-cristallisées avec un ligand à moins de 8 Å de l'interface.

La majorité des ligands observés sont des ions ou des molécules dépouvues de propriétés pharmacologiques (détergeants, agents précipitants). Nous avons recensé 41 824 ligands redondants aux abords des interfaces (< 8Å de l'interface) dont 9 126 sont à une distance de 4 Å ou moins. Les ligands identifiés ont été filtrés selon une liste de code HET pré-établie dans IChem, afin de ne conserver que les ligands pharmacologiques.

Ligands interfaciaux

Après élaguage, on observe 64 ligands uniques situés à moins de 2 Å de l'interface **Annexe 4.2**) et 64 ligands situés entre 2 et 2.5 Å de l'interface (**Annexe 4.3**) dont 5 sont également présents dans la liste précédente (A77, ADE, BT6, PLP, SAM, XFJ). On y retrouve notamment des cosubstrats (ex: s-adénosylméthionine, pyridoxal 5'-phosphate), des inhibiteurs enzymatiques (ex: A77, inhibiteur de la protéase du VIH-1) et des inhibiteurs interfaciaux connus comme la brefeldine (HET = AFB) situé à l'interface du complexe entre Arf1 et Arno (PDB id 1R8Q, **Figure 4.16**).

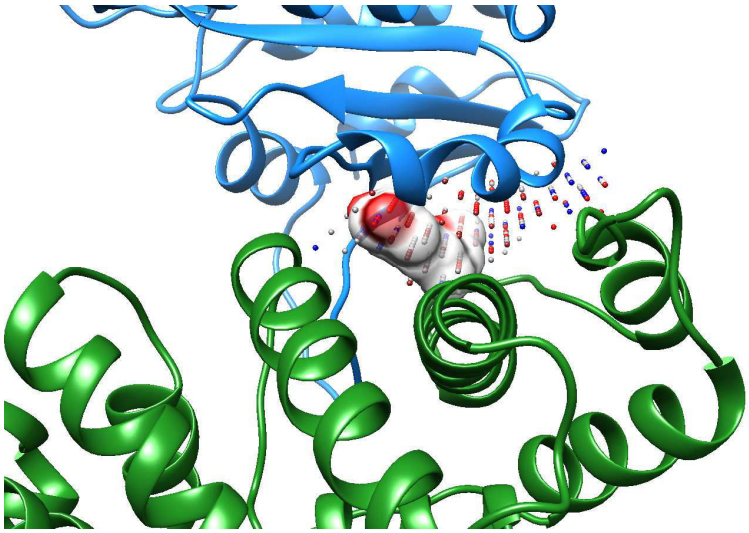


Figure 4.16: Brefeldine (surface translucide) logée dans une cavité droguable (points) à l'interface du complexe entre les protéines Arf1 (rubbans bleus) et le domaine Sec7 de la protéine Arno (rubbans verts). Les points de cavité sont colorés en fonction de leurs propriétés pharmacophoriques (hydrophobe, gris; aromatique, vert; donneur, bleu; positif ionisable, bleu; accepteur; rouge; négatif ionisable, rouge)

Il est à noter que certains ligands interfaciaux connus (ex: rapamycine) n'ont pu être détectés par notre protocole automatisé. En effet, la rapamycine est logée au sein d'une cavité très volumineuse (1400 $Å^3$) délimitant une interface prédite non pertinente par IChemPIC. En effet, les deux protéines (FKBP-12, FRAP) ne présentent que très peu d'interactions directes entre elles, la plupart étant médiées par le ligand d'interface lui-même (**Figure 4.17**)

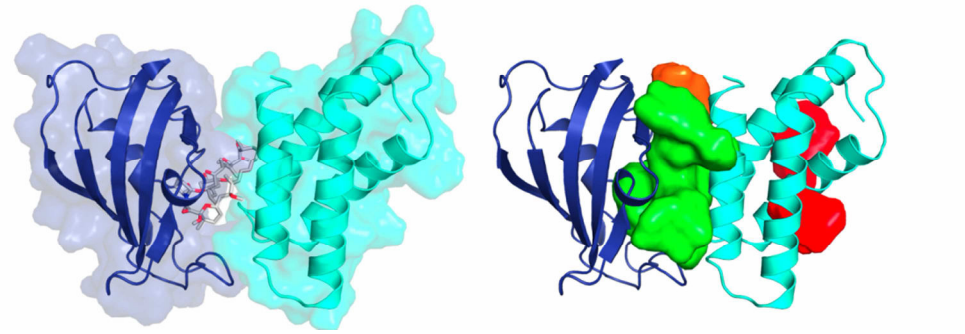


Figure 4.17: A) Complexe entre la rapamycine et l'interface FKBP-12/FRAP (PDB 1FAP). B) La cavité interfaciale contenant la rapamycine est représentée en vert, attenante à une cavité de bordure (surface orange). Une cavité allostérique (surface rouge) est présente sur le monomère FRAP.

Ligands de bordure

Après filtrage, nous n'avons trouvé qu'un seul ligand présent dans une cavité de bordure: il s'agit du dTDP-4-amino-4,6-dideoxyglucose (HET code: 0FX) à l'interface des 3 chaines d'un homotrimère de la galactoside O-acetyltransferase (**Figure 4.18**).

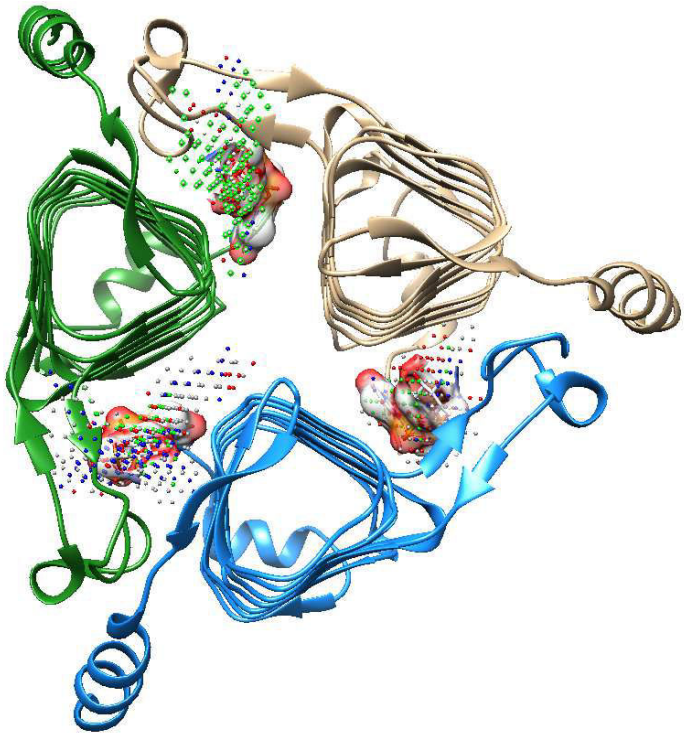


Figure 4.18: *dTDP-4-amino-4,6-dideoxyglucose (surface translucide) à l'interface des 3 chaines (rubbans bleus, verts et jaunes) de la galactoside O-acetyltransferase (PDB ID 3vbi). Les points de cavité sont colorés en fonction de leurs propriétés pharmacophoriques (hydrophobe, gris; aromatique, vert; donneur, bleu; positif ionisable, bleu; accepteur; rouge; négatif ionisable, rouge)*

Ce ligand étant un co-substrat (avec le coenzyme A) d'une acétyltransferase dans la biosynthèse du D-anthrose chez bacillus cereus[23], le site de bordure correspondant est donc un site potentiel dans la recherche d'anti-infectieux. Dans la mesure où la recherche de cavités n'a été effectuée que sur des structures oligomériques, il est normal que nous n'ayons pas encore identifié des sites de bordure co-cristallisés avec un inhibiteur d'interface. Pour ce faire, nous devrons comparer les structures monomériques de la PDB avec le structures oligomériques selon un protocole dejà défini, décrit plus tart (section Ligands orthostériques)

Ligands allostériques

Malgré le fait que les cavités allostériques soient très largement majoritaires parmi toutes les cavités droguable détectées (**Figure 4.9**), nous n'avons pu identifier qu'un seul ligand pharmacologique occupant une de ces cavités, un stabilisateur allostérique de l'interface ubiquitine-Cdc34[24] (**Figure 4.19**)

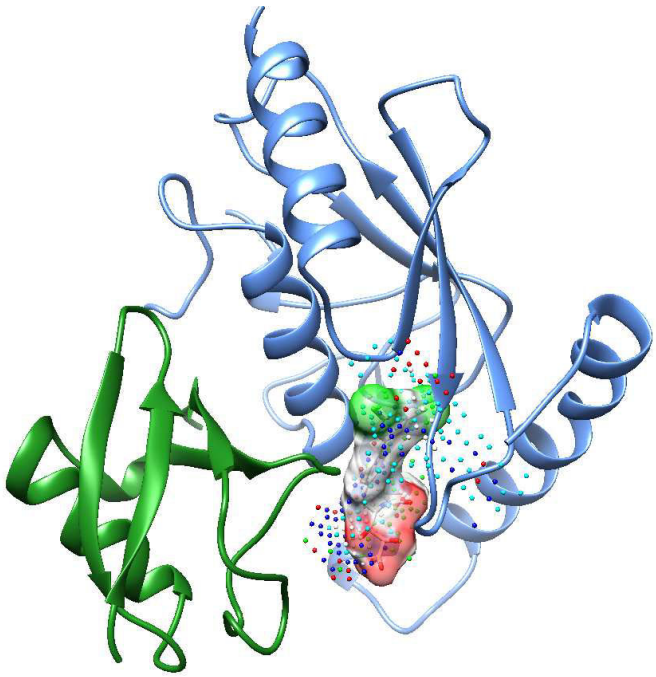


Figure 4.19: *Inhibiteur CC00651 (surface translucide) à l'interface du complexe entre ubiquitine (rubbans verts) et son enzyme de conjugaison Cdc34 (rubbans bleus, PDB ID 4mdk). Les points de cavité sont colorés en fonction de leurs propriétés pharmacophoriques (hydrophobe, gris; aromatique, vert; donneur, bleu; positif ionisable, bleu; accepteur; rouge; négatif ionisable, rouge)*

Nous avons délibérément omis pour le moment toute cavité distante de plus de 8 Å de l'interface. Il est évidemment probable que d'autres ligands aient pu être co-cristallisés dans une cavité droguable à plus grande distance de l'interface

Ligands orthostériques

Le procédé décrit actuellement ne permet pas de détecter les ligands orthostériques[21], principalement constitués d'inhibiteurs d'interfaces qui se fixent à un des partenaires du

complexe afin d'empêcher la fixation de la seconde protéine. Afin de les détecter, il nous reste à mener les étapes suivantes:

- parser les entrées monomériques de la PDB;
- détecter les cavités droguables correspondantes;
- identifier les chaines protéiques présentes à la fois dans ces entrées monomériques et dans nos structures oligomériques déjà analysées;
- aligner les deux chaines protéiques (monomère vs.dimère) afin de les placer dans le même référenciel de coordonnées;
- enfin sélectionner les cavités/ligands à l'interface ou proche de cette dernière.

Nous avons réalisé manuellement cette recherche sur certaines entrées test issues de la base de données 2P2I[21]. Dans le cas d'un complexe où deux protéines se lient sans changement conformationnel notable et après alignement des monomères séparés au dimère cristallographique, nous retrouvons facilement les ligands d'interface dans les cavités orthostériques décrites à partir du dimère dissocié (**Figure 4.20**).

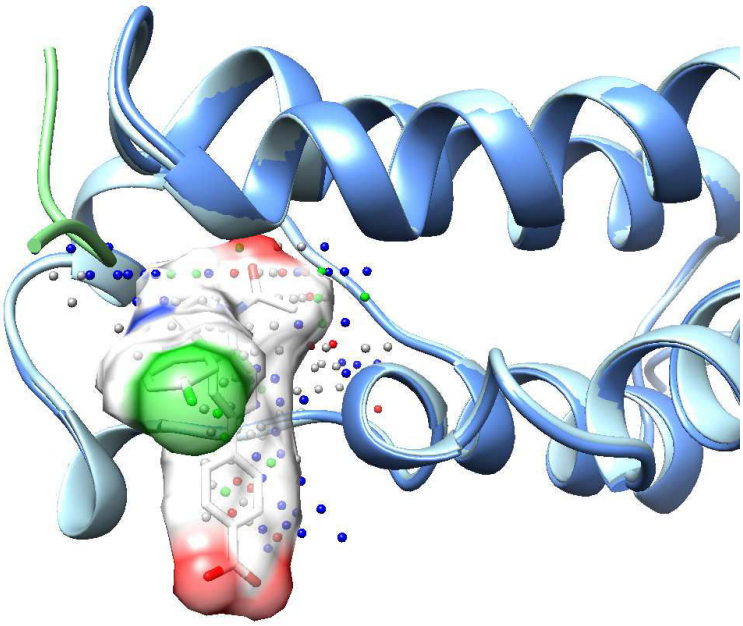


Figure 4.20: *Inhibiteur de Brd2 (73B, surface translucide) co-cristallisé au domaine N-terminal de Brd2 humain (PDB ID 4uyf, rubbans bleu foncé). Le ligand se trouve dans une cavité orthostérique (points) détectée après dissociation du dimère (PDB ID 2dvq) sur la chaine Brd2 (rubbans bleu clair). La cavité délimite exactement le site d'interaction du fragment N-terminal de l'histone H4 (rubban vert). Les points de cavité sont colorés en fonction de leurs propriétés pharmacophoriques (hydrophobe, gris; aromatique, vert; donneur, bleu; positif ionisable, bleu; accepteur; rouge; négatif ionisable, rouge)*

4.4. Conclusions

Nous avons présenté ici une nouvelle méthode de recherche d'interfaces dérivées de complexes cristallographiques afin de découvrir de nouveaux sites potentiels de modulation d'interface protéine-protéine ainsi que leurs ligands. Plus précisément, nous avons montré le développement d'un processus d'analyse des interactions protéine-protéine dans la Protein Data Bank. Le fonctionnement du processus a été détaillé et les résultats préliminaires ont été mis en avant.

Les cavités droguables observées à la surface du dimère ou des monomères dissociés ont été classées en 4 catégories selon leur enfouissment et leur distance à une interface biologiquement relevante. Nous avons pu détecter près de 272 000 cavités non redondantes pour 68 000 interfaces d'intérêt. Ces cavités représentent un gisement extraordinaire pour la recherche de modulateurs (orthostériques, allostériques) de ces interfaces. Leur identification et caractérisation structurale nous permettra dans le futur: (1) de les comparer à des cavités droguables de protéines monomériques globulaires afin d'identifier des ligands potentiels par simple recherche de similarité de cavités, (2) de proposer de manière automatisée un pharmacophore seon le procédé décrit dans le chapitre précédent, (3) d'identifier des ligands potentiels par recherche pharmacophorique ou par arrimage moléculaire.

Afin de prioriser les interfaces les plus prometteuses, il nous faudra encore relier ces cavités et les complexes correspondant soit à des voies métaboliques soit à des maladies déjà identifiées par recherche des identifiants de protéine ou de gène correspondant dans des bases de données spécialisées (ex: KEGG, PharmGKB). Il est à noter que notre détection de cavités n'est pas exhaustive car uniquement opérée soit sur des structure oligomériques, soit sur les monomères dissociés correspondants. Dans la mesure ou les disrupteurs de PPI ne se lient qu'à un des deux partenaires, il nous faudra encore détecter l'ensemble des cavités droguables à la surface des protéines monomériques de la PDB, les placer dans le même référenciel de coordonnées que les complexes existants impliquant cette même protéine, puis localiser ces cavités par rapport aux interfaces biologiquement pertinentes à leur voisinage. Nous aurons ainsi une cartographie complète de l'ensemble des cavités droguables à l'interface et près de cette dernière, ainsi que des ligands inhibiteurs co-cristallisés.

L'ensemble de ces données sera stockée dans une base de données relationelle permettant la saisie de requêtes complexes portant de multiples caractéristiques structurales du complexe, des protéines impliquées, des cavités droguables et des ligands existants.

L'actualisation périodique de cette base sera d'autant plus facilitée que le flux de données présenté dans ce chapitre est entièrement automatisé.

4.5. Bibliographie

1. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. Nat. Rev. Drug Discov. advance online publication**,** (2016).

2. Arkin, M. R., Tang, Y. & Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. Chem. Biol. **21,** 1102–1114 (2014).

3. Bader, G. D., Betel, D. & Hogue, C. W. V. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. **31,** 248–250 (2003).

4. Stumpf, M. P. H. et al. Estimating the size of the human interactome. Proc. Natl. Acad. Sci. **105,** 6959–6964 (2008).

5. Basse, M. J. et al. 2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. Nucleic Acids Res. **41,** D824–D827 (2013).

6. Labbé, C. M., Laconde, G., Kuenemann, M. A., Villoutreix, B. O. & Sperandio, O. iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. Drug Discov. Today **18,** 958–968 (2013).

7. Higueruelo, A. P. et al. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chem. Biol. Drug Des. **74,** 457–467 (2009).

8. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. J. Mol. Biol. **372,** 774–797 (2007).

9. Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S. & Capitani, G. A PDB-wide, evolution-based assessment of protein-protein interfaces. BMC Struct. Biol. **14,** 22 (2014).

10. Da Silva, F., Desaphy, J., Bret, G. & Rognan, D. IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces. J. Chem. Inf. Model. **55,** 2005–2014 (2015).

11. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Continuum Analytics, Nov. 2015. Web. <https://continuum.io>.

12. mariaDB. (mariaDB company).

13. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25,** 1422–1423 (2009).

14. Bietz, S., Urbaczek, S., Schulz, B. & Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J. Cheminformatics **6,** 12 (2014).

15. Desaphy, J., Raimbaud, E., Ducrot, P. & Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. J. Chem. Inf. Model. **53,** 623–637 (2013).

16. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. J. Chem. Inf. Model. **52,** 2287–2299 (2012).

17. Edfeldt, F. N. B., Folmer, R. H. A. & Breeze, A. L. Fragment screening to predict druggability (ligandability) and lead discovery success. Drug Discov. Today **16,** 284–287 (2011).

18. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. J. Med. Chem. **50,** 74–82 (2007).

19. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. Nucleic Acids Res. **43,** D399–404 (2015).

20. Thiel, P., Kaiser, M. & Ottmann, C. Small-molecule stabilization of protein-protein interactions: an underestimated concept in drug discovery? Angew. Chem. Int. Ed Engl. **51,** 2012–2018 (2012).

21. Basse, M.-J., Betzi, S., Morelli, X. & Roche, P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. Database J. Biol. Databases Curation 2016**,** (2016).

22. Gao, M. & Skolnick, J. The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism of formation. Proc. Natl. Acad. Sci. U.S.A. **109**, 3794-3789 (2012).

23. Kubiak, R.L. & Holden, H.M. Structural studies of AntD: an N-acyltransferase involved in the biosynthesis of d-anthrose. Biochemistry, **51**, 867–878 (2012).

24. Ceccarelli, D.F., Orlicky, S., Tyers, M. & Sicheri, F. E2 enzyme inhibition by stabilization of a low-affinity interface with ubiquitin. Nat.Chem.Biol. **10**, 156-163 (2014)

Annexes

Annexe 4.1 Champ de force utilisé pour l'alignement des cavités

```
########################################
#       Preliminary Definitions        #
########################################
########################################
#         Interaction Types            #
########################################
TYPE donor
TYPE acceptor
TYPE cation
TYPE anion
TYPE rings
TYPE hydrophobe
TYPE ringAliph
TYPE test
########################################
#         Type Patterns                #
########################################
# rings
#
PATTERN rings [14C]
#
# ringAliph
#
# hydrophobic
#
# terminal hp
PATTERN hydrophobe [13C]
# non-terminal hp
PATTERN acceptor [17O;X0]
PATTERN acceptor [15O;X0]
PATTERN donor          [14N;X0]
PATTERN donor    [15O;X0]
#
# anion/cation patterns
#
# cations
PATTERN cation [15N;X0]
# anions
PATTERN anion [14O;X0]
########################################
#      Interaction Definitions         #
########################################
INTERACTION donor donor attractive gaussian weight=0.0 radius=1.0
INTERACTION acceptor acceptor attractive gaussian weight=0.0 radius=1.0
INTERACTION rings rings attractive gaussian weight=0.0 radius=1.0
INTERACTION ringAliph ringAliph attractive gaussian weight=0.0 radius=1.0
INTERACTION ringAliph hydrophobe attractive gaussian weight=0.0 radius=1.0
INTERACTION cation cation attractive gaussian weight=0.0 radius=1.0
INTERACTION anion anion attractive gaussian weight=1.0 radius=1.0
INTERACTION anion acceptor attractive gaussian weight=1.0 radius=1.0
```

INTERACTION cation donor attractive gaussian weight=1.0 radius=1.0
INTERACTION hydrophobe hydrophobe attractive gaussian weight=0.0 radius=1.0

Annexe 4.2 Tableau récapitulatif des ligands situés à moins de 2Å d'une interface protéine-protéine

Ligand ID	Ligand Formula	Ligand MW	Ligand Name	Ligand SMILES	Instance PDB IDs (All)
0U7	C14 H7 CL2 F N2 O3 S	373.18	3-[5-(3,5-DICHLOROPHENYL)-1,3,4-OXADIAZOL-2-YL]BENZENESULFONYL FLUORIDE	c1cc(cc(c1)S(=O)(=O)F)c2nnc(o2)c3cc(cc(c3)Cl)Cl	4FI6
16V	C15 H17 F N2 O3	292.31	3-[3-(3,5-DIMETHYL-1H-PYRAZOL-4-YL)PROPOXY]-4-FLUOROBENZOIC ACID	Cc1c(c(n[nH]1)C)CCCOc2cc(ccc2F)C(=O)O	4HIQ
1W3	C20 H18 O4	322.36	7-HYDROXY-3-[(E)-2-(4-HYDROXY-3,5-DIMETHYLPHENYL)ETHENYL]-4-METHYL-2H-CHROMEN-2-ONE	Cc1cc(cc(c1O)C)/C=C/C2=C(c3ccc(cc3OC2=O)O)C	4KY2
1W4	C25 H24 F N O2 S	421.53	S-(4-FLUOROPHENYL) 3-(DIMETHYLAMINO)-5-[(E)-2-(4-HYDROXY-3,5-DIMETHYLPHENYL)ETHENYL]BENZENECARBOTHIOATE	Cc1cc(cc(c1O)C)/C=C/c2cc(cc(c2)N(C)C)C(=O)Sc3ccc(cc3)F	4L1S

1ZK	C45 H62 N7 O7 1	813.03	4-[(2R)-3-{[(1S,2S,3R,4S)-1-(CYCLOHEXYLMETHYL)-2,3-DIHYDROXY-5-METHYL-4-({(1S,2R)-2-METHYL-1-[(PYRIDIN-2-YLMETHYL)CARBAMOYL]BUTYL}CARBAMOYL)HEXYL]AMINO}-2-{[(NAPHTHALEN-1-YLOXY)ACETYL]AMINO}-3-OXOPROPYL]-1H-IMIDAZOL-3-IUM	CC[C@H](C)[C@@H](C(=O)NCc1ccccn1)NC(=O)[C@H](C(C)C)[C@H]([C@@H]([C@H](CC2CCCCC2)NC(=O)[C@H](Cc3c[nH]c[nH+]3)NC(=O)COc4cccc5c4cccc5)O)O	1HIV,1IVP
3EN	C28 H36 N4 O4 S	524.67	N-[(2S,4S,5S)-4-hydroxy-1,6-diphenyl-5-{[(1,3-thiazol-5-ylmethoxy)carbonyl]amino}hexan-2-yl]-L-valinamide	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)C[C@@H]([C@H](Cc2ccccc2)NC(=O)OCc3cncs3)O)N	4U7V
478	C25 H35 N3 O6 S	505.63	{3-[(4-AMINO-BENZENESULFONYL)-ISOBUTYL-AMINO]-1-BENZYL-2-HYDROXY-PROPYL}-CARBAMIC ACID TETRAHYDRO-FURAN-3-YL ESTER	CC(C)C[N@](C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]2CCOC2)O)S(=O)(=O)c3ccc(cc3)N	1HPV,1T7J,3EKP,3EKV,3EM3,3NU3,3NU4,3NU5,3NU6,3NU9,3NUJ,3NUO,3OXV,3S43,3S45,3SM2,4J5J,4JEC,4RVJ
4AJ	C14 H8 Cl2 N2 O3	323.13	2,6-dichloro-4-[5-(3-hydroxyphenyl)-1,3,4-oxadiazol-2-yl]phenol	c1cc(cc(c1O)c2nnc(o2)c3cc(c(c(c3)Cl)O)Cl	4YDM

4V4	C18 H26 N4	298.43	N1-(3-(2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL)PHENYL)-N1,N2-DIMETHYLETHANE-1,2-DIAMINE	Cc1cc(nc(c1)N)CCc2cccc(c2)N(C)CCNC	4UGH,4UGZ,4UH7
6H6	C10 H9 N O4	207.18	3-(2-OXO-1,3-BENZOXAZOL-3(2H)-YL)PROPANOIC ACID	c1ccc2c(c1)N(C(=O)O2)CCC(=O)O	3ZMB,3ZOU
8PC	C18 H13 CL2 N O2	346.21	2-(2,4-DICHLOROPHENOXY)-5-(PYRIDIN-2-YLMETHYL)PHENOL	c1ccnc(c1)Cc2ccc(c(c2)O)Oc3ccc(cc3Cl)Cl	2OP1,3FNE
A77	C44 H58 N8 O6	794.99	N-{1-BENZYL-(2R,3S)-2,3-DIHYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)[C@@H]([C@@H]([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3ccccn3)O)O)NC(=O)N(C)Cc4ccccn4	1HVI,1HVS
AD5	C21 H27 N7 O	393.49	N~6~-CYCLOHEXYL-N~2~-(4-MORPHOLIN-4-YLPHENYL)-9H-PURINE-2,6-DIAMINE	c1cc(ccc1Nc2nc3c(c(n2)NC4CCCCC4)nc[nH]3)N5CCOCC5	2VGO
ADE	C5 H5 N5	135.13	ADENINE	c1[nH]c2c(n1)c(ncn2)N	1AHA,1BJQ,1CB0,1D2A,1GIU,1HQC,1IFS,1J1R,1JH8,1JYS,1LPD,1LU1,1M2T,1MUD,1NLI,1OD2,1OD4,1Q8Y,1QB7,1QCI,1QD2,1S2D,1VRL,1WEI,1WTA,1XE8,1Y26,1YXM,1Z5N,1Z8D,1ZN7,2GA4,2H8G,2ICS,2P8N,2PQJ,2PUB,2PUE,2QES,2QET,2QLU,2QTT,2XOC,2XOY,2YED,3A7I,3BSF,3E9R,3K9W,3KPV,3KU

					0,3LE7,3LGS,3LQV,3MRY,3NG9,3NM6,3ONE,3PAO,3QUI,3RL9,3RYS,3S99,3TPV,3U6Z,3U70,3UJO,3V2K,3W52,3WAZ,4BMX,4D8V,4DAO,4DC2,4F1W,4FBA,4FBB,4G89,4I30,4JOS,4JWT,4KQF,4L0M,4L5C,4L6I,4LNA,4LW0,4M1E,4NSN,4OJT,4P14,4PR3,4QAR,4QEZ,4TRC,4TZX,4TZY,4XGP,4XJ7,4XNR,4YMI,5DK6,5DYX
ANC	C14 H11 N	193.25	ANTHRACEN-1-YLAMINE	c1ccc2cc3c(cccc3N)cc2c1	1GT1,1HN2
BSJ	C36 H62 N11 O23 P3 S	1141.93	(3R,9Z)-17-[(2R,3S,4R,5R,6R)-5-AMINO-6-{[(1R,2R,3S,4R,6S)-4,6-DIAMINO-2,3-DIHYDROXYCYCLOHEXYL]OXY}-3,4-DIHYDROXYTETRAHYDRO-2H-PYRAN-2-YL]-3-HYDROXY-2,2-DIMETHYL-4,8,15-TRIOXO-12-THIA-5,9,16-TRIAZAHEPTADEC-9-EN-1-YL [(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]METHYL DIHYDROGEN DIPHOSPHATE	CC(C)(CO[P@](=O)(O)O[P@](=O)(O)OC[C@@H]1[C@H]([C@H]([C@@H](O1)n2cnc3c2ncnc3N)O)OP(=O)(O)O)[C@H](C(=O)NCCC(=O)/N=C\CSCCC(=O)NC[C@@H]4[C@H]([C@@H]([C@H]([C@@H](O4)O[C@@H]5[C@H](C[C@H]([C@@H]([C@H]5O)O)N)N)N)O)O)O	2VBQ
BT6	C6 H6 S	110.17	BENZENETHIOL	c1ccc(cc1)S	3HSR

CRP	C15 H18 CL3 N O	334.67	((1RS,3SR)-2,2-DICHLORO-N-[(R)-1-(4-CHLOROPHENYL)ETHYL]-1-ETHYL-3-METHYLCYCLOPROPANECARBOXAMIDE	CC[C@]1([C@H](C1(Cl)Cl)C)C(=O)N[C@H](C)c2ccc(cc2)Cl	2STD,7STD
DJR	C28 H38 N2 O8 S	562.68	(3R,3AS,6AR)-HEXAHYDROFURO[2,3-B]FURAN-3-YL [(1S,2R)-1-BENZYL-2-HYDROXY-3-{ISOBUTYL[(4-METHOXYPHENYL)SULFONYL]AMINO}PROPYL]CARBAMATE	CC(C)C[N@@](C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]2CO[C@@H]3[C@H]2CCO3)O)S(=O)(=O)c4ccc(cc4)OC	2I4U,2I4V,3I7E
DP1	C10 H22 N8 O4	318.33	L-N(OMEGA)-NITROARGININE-2,4-L-DIAMINOBUTYRIC AMIDE	C(C[C@@H](C(=O)N[C@@H](CCN)C(=O)N)N)CNC(=N)N[N+](=O)[O-]	1P6H,1P6K,1P6L,1Q2O,1ZZU
DP9	C11 H22 N8 O4	330.35	L-N(OMEGA)-NITROARGININE-(4R)-AMINO-L-PROLINE AMIDE	C1[C@H](CN[C@@H]1C(=O)N)NC(=O)[C@H](CCCNC(=N)N[N+](=O)[O-])N	1P6J,1P6N,1ZZQ,1ZZR,1ZZS,1ZZT
DPF	C4 H11 O4 P	154.10	DIETHYL HYDROGEN PHOSPHATE	CCOP(=O)(O)OCC	2R1K,2R1M,3CAK,3UR5,3URB,4FNM,4QWM,4UBI,4UBJ,4UBK,4UBL,4UBM,4UBN,4UBO,4W1P,4W1Q,4W1R,4W1S,5CH5,5IVI,5IVK

DRS	C19 H26 N2 O5	362.42	(9S,12S)-9-(1-METHYLETHYL)-7,10-DIOXO-2-OXA-8,11-DIAZABICYCLO[12.2.2]OCTADECA-1(16),14,17-TRIENE-12-CARBOXYLIC ACID	CC(C)[C@H]1C(=O)N[C@@H](Cc2ccc(cc2)OCCCCC(=O)N1)C(=O)O	3BXS
DTB	C10 H18 N2 O3	214.26	6-(5-METHYL-2-OXO-IMIDAZOLIDIN-4-YL)-HEXANOIC ACID	C[C@H]1[C@H](NC(=O)N1)CCCCCC(=O)O	1DAM,1R30,3FPA,3RDQ,4A0R,4DNE,5IWK,5IWP,5IWR,5IWT
DUC	C4 H6 N2 O2	114.10	DIHYDROPYRIMIDINE-2,4(1H,3H)-DIONE	C1CNC(=O)NC1=O	1UAQ,2FVK,4XK4
DZ2	C13 H7 BR2 CL2 N O2	439.92	2,5-DICHLORO-N-(3,5-DIBROMO-4-HYDROXYPHENYL)BENZAMIDE	c1cc(c(cc1Cl)C(=O)Nc2cc(c(c(c2)Br)O)Br)Cl	3ESO
EXI	C17 H25 N5	299.41	N1-(5-(2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL)PYRIDIN-3-YL)-N1,N2-DIMETHYLETHANE-1,2-DIAMINE	Cc1cc(nc(c1)N)CCc2cc(cnc2)N(C)CCNC	4UGY,4UH1,4UH5,4UH8,5G0N
F1P	C6 H13 O9 P	260.14	1-O-PHOSPHONO-BETA-D-FRUCTOPYRANOSE	C1[C@H]([C@H]([C@@H]([C@](O1)(COP(=O)(O)O)O)O)O)O	2DF8,4BB9
FNN	C6 H4 F N O3	157.10	3-FLUORO-4-NITROPHENOL	c1cc(c(cc1O)F)[N+](=O)[O-]	3VGN
HOZ	C5 H12 N2 O2	132.16	(4S)-4,5-DIAMINOPENTANOIC ACID	C(CC(=O)O)[C@@H](CN)N	2HOZ,2HP1,2HP2,3USF

I1E	C9 H10 O	134.18	1-PHENYLPROPAN-1-ONE	CCC(=O)c1ccccc1	3SZB
I5U	C27 H29 N5 O2 S2	519.68	N',N'-{[(2R)-3-AMINOPROPANE-1,2-DIYL]BIS(OXYMETHANEDIYLBENZENE-3,1-DIYL)}DITHIOPHENE-2-CARBOXIMIDAMIDE	[H]/N=C(\c1ccccs1)/Nc2cccc(c2)COC[C@@H](CN)OCc3cccc(c3)/N=C(\c4cccs4)/N	4UGP,4UPM,4UPQ
IMD	C3 H5 N2 1	69.09	IMIDAZOLE	c1c[nH+]c[nH]1	1AES,1AKY,1CCG,1CXA,1DP9,1DS4,1DSE,1DSG,1DSO,1DSP,1F4U,1FI7,1FI9,1GW6,1GZ5,1H19,1HEZ,1HS6,1HX3,1I12,1I1D,1IKJ,1IRC,1JRL,1K6Z,1KAE,1L5N,1L9E,1MBI,1MRC,1MRD,1MRE,1MRF,1MUN,1MUY,1N1I,1NMI,1NOC,1NOS,1NU7,1NU9,1OAU,1OD8,1OUW,1PEE,1PM1,1PZN,1QPI,1R8E,1RKY,1RUW,1SQM,1SXU,1T8K,1U17,1U7R,1UGI,1UUX,1UUY,1V0M,1V0N,1V2G,1VDD,1W7C,1W81,1W8K,1WEG,1WKQ,1WWJ,1Y59,1Y5A,1Y5B,1Y5E,1Y5U,1YWV,1YXS,1YXU,1YXV,1YXX,2A0I,2AH8,2ASN,2AT3,2C4W,2D6C,2DDF,2DKK,2E18,2E2E,2F5T,2FMY,2G5F,2G8C,2GO3,2GPE,2H7Q,2HBA,2HRC,2HVF,2I6O,2KIL,2NOP,2NOS,2NPC,2NPD,2NPJ,2NPK,2NQC,2NUH,2O3P,2O63,2O64,2O65,2OBL,2OD5,2OGJ,2OH3,2OPC,2OWQ,2PSD,2PSE,2QCU,2QD1,2QD2,2QLE,2QZC,2R45,2R46,2R4E,2R4J,2R5R,2RD9,2UGI,2V08,2VD3,2VJ8,2VNQ,2VYO,2VZ5,2W14,2W8X,2WCR,2WIY,2WVT,2X32,2X7Z,2X9X,2XMS,2XN2,2Y24,2Y73,2Y74,2YGT,2YMV,2YNA,2Z3T,2ZBX,2ZC0,2ZEG,2ZQN,2ZQO,3AKY,3B7R,3B7T,3B7U,3BQZ,3BR0,3C1H,3C1J,3C2Q,3C4E,3C8L,3CC1,3CGX,3CHP,3CHQ,3CHR,3CHS,3D70,3D71,3D85,3DEE,3DKQ,3E2T,3E49,3E6Q,3EBZ,3EC0,3ECG,3EER,3EJ8,3F0Y,3FAV,3FCS,3FH5,3FH7,3FH8,3FHE,3FPC,3FTS,3FTU,3FTV,3FTW,3FTX,3FTY,3FTZ,3FU0,3FU3,3FU5,3FU6,3FUD,3FUE,3FUF,3FUH,3FUI,3FUJ,3FUK,3FUL,3FUM,3FUN,3FVB,3FXU,3FY4,3G5O,3GFF,3GGD,3GZ2,3H87,3HCN,3HCO,3HCP,3HCR,3HDJ,3HMZ,3HQ8,3HQ9,3HTA,3HX8,3IG9,3IGE,3IHQ,3IUU,3JTJ,3JW4,3K1M,3K1N,3K43,3K9T,3KKC,3KNZ,3L07,3L1O,3L5O,3LAH,3LAI,3LAT,3LRG,3LYG,3M1J,3MBJ,3MGC,3MOM,3MVU,3MYR,3NN1,3NO6,3NQ2,3NRB,3OA6,3P0K,3P76,3P9Y,3PA1,3PA2,3PFE,3PGF,3PH2,3PMM,3POH,3PS4,3PVK,3Q10,3Q38,3Q39,3Q3A,3Q6Q,3Q6R,3Q93,3QB8,3QK1,3QMA,3QZY,3R00,3R01,3R02,3R04,3R3Q,3RF4,3RF5,3S42,3S45,3S67,3SBP,3SBQ,3SBR,3SI0,3SL9,3SMV,3SOY,3T5T,3T7D,3TCL,3TE7,3TE

					M,3TGB,3TKB,3TQL,3TS4,3U24,3U7Q,3U9W,3UGX,3UR2,3URA,3URB,3URN,3URQ,3UVC,3V13,3V7M,3VBY,3VC4,3VDP,3VE5,3VHB,3VIV,3WFX,3WG5,3WS6,3ZJP,4A7C,4ADN,4AIQ,4AJX,4ALU,4ALV,4ALW,4AOR,4ASM,4AU8,4BCX,4BF7,4BOE,4C5B,4D6X,4D6Z,4DM3,4E0U,4E6R,4EFK,4EFS,4EHR,4EJ5,4EL4,4ELC,4EP6,4ESE,4EWG,4FCA,4FDP,4FF6,4FRZ,4FXX,4G84,4GFC,4GK8,4GK9,4GR8,4GYU,4HF3,4HHR,4HJI,4HPN,4HVL,4IJ1,4IRP,4J27,4J28,4J2K,4J3G,4JBS,4JFS,4JFT,4JFU,4JFV,4JFW,4JL1,4JL2,4JQ9,4JYS,4KF2,4KKD,4KLR,4KNK,4KNL,4KQO,4KQQ,4KQR,4KRY,4KSI,4KVL,4KW5,4L36,4LJ3,4LJO,4LSV,4LUE,4LZD,4M08,4M09,4M0R,4M7S,4MB3,4MBA,4MCW,4ME4,4MGR,4MPZ,4MQQ,4MTB,4MU1,4N5B,4NCR,4NCY,4O2A,4O2B,4O6H,4O6I,4O6O,4OC9,4ONG,4OP7,4OWM,4OWN,4OWO,4OWQ,4OWS,4OWU,4P8C,4P8L,4P8M,4P8N,4P8P,4P8T,4P8Y,4PAC,4PCS,4PCT,4PE3,4PEE,4PFB,4PFD,4PII,4PKD,4PSN,4PW4,4Q2B,4Q40,4Q7I,4QC4,4QHP,4QIR,4QKR,4QME,4QO3,4QP1,4QP2,4QP6,4QT2,4QUO,4QVE,4R7L,4RN0,4RN1,4RS3,4RSY,4RVB,4TKU,4TKV,4U3B,4U3D,4U89,4UD1,4UG1,4UGI,4UIQ,4UY5,4UZI,4V06,4W6F,4W73,4W7E,4W9N,4WAC,4WAT,4WI1,4WSK,4X2V,4X58,4X59,4X5D,4X7F,4XAB,4XAC,4XH7,4XMG,4XXB,4XXT,4YCX,4YIC,4YJ2,4YJ3,4YOL,4YTB,4YXB,4YZ0,4YZA,4YZO,4YZX,4Z3V,4ZE4,4ZE5,4ZEH,4ZEN,4ZEP,4ZFM,4ZLY,4ZLZ,4ZOF,4ZOJ,5A1G,5AEN,5AIF,5BN1,5BNE,5BO2,5BO3,5BPD,5BVG,5BVH,5C2L,5C6F,5D2C,5D2N,5DC9,5DDK,5DF1,5DF7,5DGC,5DKF,5DLE,5DW4,5DW5,5DW6,5DWM,5DYF,5E5H,5EXK,5FJI,5HA5,5HDR,5HKO,5HSQ,5I5R,5INA,5IOA,5IOE,5IOG,5IZ3,5JH7,5K1J,5K1N
J5L	C32 H43 N5 O10 S	689.78	(3R,3AS,4R,6AR)-4-[2-(METHYLAMINO)-2-OXOETHOXY]HEXAHYDROFURO[2,3-B]FURAN-3-YL {(2S,3R)-3-HYDROXY-4-[{[(2Z)-2-(METHYLIMINO)-2,3-DIHYDRO-1,3-BENZOXAZOL-6-YL]SULFONYL}(2-	CC(C)CN(C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]2CO[C@@H]3[C@H]2[C@H](CO3)OCC(=O)NC)O)S(=O)(=O)c4ccc5c(c4)O/C(=N\C)/N5	5AHB

			METHYLPROPYL)AMINO]-1-PHENYLBUTAN-2-YL}CARBAMATE		
K5Q	C22 H19 N O7	409.39	(2S)-6-[(E)-[(2E)-2-HYDROXYIMINO-3H-INDEN-1-YLIDENE]METHYL]-2-(3-HYDROXY-3-OXOPROPYL)-2,3-DIHYDRO-1,4-BENZODIOXINE-5-CARBOXYLIC ACID	c1ccc\2c(c1)C/C(=N\O)/C2=C/c3ccc4c(c3C(=O)O)OC[C@@H](O4)CCC(=O)O	4CGG
KPC	C5 H10 O4 S2	198.25	(2-[2-KETOPROPYLTHIO]ETHANESULFONATE	CC(=O)CSCCS(=O)(=O)O	1MO9,2CFC,3Q6J
LJ4	C12 H8 BR2 O2	344.00	2,6-DIBROMO-4-PHENOXYPHENOL	c1ccc(cc1)Oc2cc(c(c(c2)Br)O)Br	3CN3
M2T	C12 H19 N5 O3 S	313.37	5'-DEOXY-5'-(DIMETHYL-LAMBDA~4~-SULFANYL)ADENOSINE	CS(C)C[C@@H]1[C@H]([C@H]([C@@H](O1)n2cnc3c2ncnc3N)O)O	3H0V,3IWD
MGB	C5 H12 N8	184.20	METHYLGLYOXAL BIS-(GUANYLHYDRAZONE)	[H]/N=C(\N)/N/N=C/C(=NNC(=N)N)C	1I7C
MR5	C13 H7 BR2 N O2	369.01	4-(1,3-BENZOXAZOL-2-YL)-2,6-	c1ccc2c(c1)nc(o2)c3cc(c(c(c	2QGD

			DIBROMOPHENOL	3)Br)O)Br	
N4M	C30 H45 N6 O16 P	776.69	1-[4-({(1R)-1-[(6S,7S)-2-AMINO-7-METHYL-4-OXO-1,4,5,6,7,8-HEXAHYDROPTERIDIN-6-YL]ETHYL}AMINO)PHENYL]-1-DEOXY-5-O-{5-O-[(R)-{[(1R)-1,3-DICARBOXYPROPYL]OXY}(HYDROXY)PHOSPHORYL]-ALPHA-D-RIBOFURANOSYL}-D-XYLITOL	C[C@H]1[C@@H](NC2=C(N1)NC(=NC2=O)N)[C@@H](C)Nc3ccc(cc3)C[C@@H]([C@H]([C@@H](CO[C@@H]4[C@@H]([C@@H]([C@H](O4)COP(=O)(O)O[C@H](CCC(=O)O)C(=O)O)O)O)O)O)O	4GVQ
N7I	C10 H12 O3	180.20	4-[(1E)-3-HYDROXYPROP-1-EN-1-YL]-2-METHOXYPHENOL	COc1cc(ccc1O)/C=C/CO	3TKY,4E70,4EVI,5CVJ,5CVV
N8M	C13 H20 N6 O3	308.34	5'-DEOXY-5'-(DIMETHYLAMINO)-8-METHYLADENOSINE	Cc1nc2c(ncnc2n1[C@H]3[C@@H]([C@@H]([C@H](O3)CN(C)C)O)O)N	3H0W
PEP	C3 H5 O6 P	168.04	PHOSPHOENOLPYRUVATE	C=C(C(=O)O)OP(=O)(O)O	1FWN,1FWS,1FWT,1FWW,1FXQ,1G7U,1HFB,1JCY,1KFL,1KHF,1LRO,1LRQ,1N8F,1NHX,1OAB,1OF8,1OFA,1ONE,1P48,1Q3N,1QR7,1RZM,1T8X,1T96,1VBH,1VS1,1XUZ,1ZCO,1ZHA,2A21,2A2I,2AL1,2AL2,2B7O,2DWO,2EF9,2GMV,2NWR,2NWS,2NX1,2NX3,2NXG,2NXH,2NXI,2O0E,2ONE,2OX3,2PTY,2QZY,2R46,2XGZ,2XH0,2XZ7,3E0I,3FYO,3FYP,3NV8,3QPW,3TFC,3UCD,3UJE,3UJF,3UJR,3UND,4C1K,4EGR,4GMW,4GNL,4GNP,4HSN,4HSO,4HYV,4I4I,4I7E,4OWG,4UCG,4WPT,4WPU,4Z17,4Z1D,5BOE,5CZ0,5CZS,5D02,5D03,5D05,5D0

					9,5J04
PI7	C36 H53 N5 O7	667.84	N-[3-(8-SEC-BUTYL-7,10-DIOXO-2-OXA-6,9-DIAZA-BICYCLO[11.2.2] HEPTADECA-1(16),13(17),14-TRIEN-11-YAMINO)-2-HYDROXY-1-(4-HYDROXY-BENZYL) -PROPYL]-3-METHYL-2-PROPIONYLAMINO-BUTYRAMIDE	CC[C@H](C)[C@H]1C(=O)NCCCOc2ccc(cc2)C[C@@H](C(=O)N1)NC[C@H]([C@H](Cc3ccc(cc3)O)NC(=O)[C@H](C(C)C)NC(=O)CC)O	1B6P
PI9	C32 H48 N4 O6 S	616.81	(10S,13S,1'R)-13-[1'-HYDROXY-2'-(N-P-AMINOBENZENESULFONYL-1''-AMINO-3''-METHYLBUTYL)ETHYL]-8,11-DIOXO-10-ISOPROPYL-2-OXA-9,12-DIAZABICYCLO [13.2.2]NONADECA-15,17,18-TRIENE	CC(C)CC[N@@](C[C@H]([C@@H]1Cc2ccc(cc2)OCCCCCC(=O)N[C@H](C(=O)N1)C(C)C)O)S(=O)(=O)c3ccc(cc3)N	1D4L

PLP	C8 H10 N O6 P	247.14	PYRIDOXAL-5'-PHOSPHATE	Cc1c(c(c(cn1)COP(=O)(O)O)C=O)O	1A3G,1A50,1A5A,1A5B,1A5S,1AAM,1AAW,1AHE,1AHF,1AHG,1AHP,1AHX,1AHY,1AKA,1ARI,1ARS,1ART,1ASA,1ASB,1ASF,1ASG,1ASM,1ASN,1AXR,1AY4,1AY5,1AY8,1B4D,1B4X,1B54,1B5O,1B5P,1B8G,1B9H,1BJ4,1BJO,1BKS,1BT4,1BX3,1C0N,1C29,1C4K,1C50,1C7G,1C7N,1C8K,1C8L,1C8V,1C9D,1CJ0,1CT5,1CW2,1CX9,1CZC,1CZE,1D2F,1D6S,1DAA,1DGD,1DGE,1DJE,1DJU,1DKA,1DTY,1E1Y,1E4O,1E5F,1ECX,1EG5,1EKF,1EKP,1EKV,1ELQ,1EM6,1EQB,1ET0,1EXV,1F2D,1F3T,1FA9,1FC0,1FC4,1FCJ,1FG3,1FS4,1FTQ,1FTW,1FTY,1FU4,1FU7,1FU8,1FUY,1G2W,1G4V,1G4X,1G76,1G77,1G78,1G79,1G7W,1G7X,1GBN,1GC3,1GC4,1GCK,1GD9,1GDE,1GEW,1GEX,1GFZ,1GG8,1GGN,1GPA,1GPB,1GPY,1H0C,1H1C,1H5U,1HKV,1HLF,1I1K,1I1L,1I1M,1I29,1I2K,1I43,1I48,1IAX,1IAY,1IBJ,1IJI,1IX6,1IX7,1IX8,1IYD,1J0A,1J0C,1J0E,1J32,1JBQ,1JF9,1JNW,1JS3,1JS6,1K06,1K08,1K3U,1K7E,1K7F,1K7X,1K8X,1K8Y,1K8Z,1KFB,1KFC,1KFK,1KKJ,1KKP,1KL1,1KL2,1KL7,1KMJ,1KMK,1KNW,1KO0,1KTI,1L5Q,1L5R,1L5S,1L5V,1L5W,1L6I,1L7X,1LK9,1LKC,1LS3,1LW4,1LW5,1LWN,1LWO,1M32,1M4N,1M54,1MDX,1MDZ,1MGV,1MLY,1MLZ,1N2T,1N31,1N8P,1NOI,1NOJ,1NOK,1NRG,1O4S,1O61,1OAS,1OAT,1OHV,1OHW,1OHY,1ORD,1P29,1P2B,1P2D,1P2G,1P3W,1P4G,1P4H,1P4J,1P5J,1PG8,1PMM,1QGN,1QIR,1QIS,1QIT,1QJ3,1QJ5,1QM5,1QOP,1QOQ,1QU4,1QZ9,1RCQ,1RFU,1RQX,1RV3,1RV4,1RVU,1RVY,1S07,1SF2,1SFT,1SZS,1SZU,1T3I,1TAR,1TAT,1TDJ,1TJP,1TTP,1TTQ,1TWI,1TYZ,1TZ2,1TZJ,1TZK,1TZM,1U08,1UBS,1UIM,1UIN,1UZU,1V2D,1V2E,1V2F,1V71,1V72,1V8Z,1VE1,1VE5,1VEF,1VFH,1VJO,1VP4,1W23,1W3U,1W7L,1W7M,1W8G,1WBJ,1WDW,1WKV,1WRV,1WST,1WTC,1WUT,1WUY,1WV0,1WV1,1WW2,1WW3,1WYU,1WYV,1XC7,1XEY,1XFC,1XI9,1XKX,1XL0,1XL1,1XOI,1XQL,1XRS,1YAA,1YGP,1YJS,1YJY,1YJZ,1YOO,1Z3Z,1Z62,1Z7W,1ZOB,1ZOD,2A1H,2A5H,2ABJ,2AMV,2AQ6,2ASV,2ATI,2AV6,2AW3,2AY1,2AY2,2AY3,2AY4,2AY5,2AY6,2AY7,2AY8,2AY9,2AZD,2BHS,2BHT,2BHX,2BI1,2BI2,2BI3,2BI5,2BI9,2BIA,2BIE,2BIG,2BKW,2BWN,2BWO,2BYJ,2BYL,2C0R,2C2B,2C2G,2C4M,2C7T,2CAN,2CB1,2CFT,2CH1,2CH2,2CIN,2CJD,2CJH,2CLE,2CLF,2CLH,2CLI,2CLK,2COG,2COI,2COJ,2CST,2CTZ,2D1F,2D5Y,2D61,2D63,2D64,2D65,2D66,2D7Y,2D7Z,2DAB,2DGK,2DGL,2DGM,2DH5,2DKB,2DKJ,2DR1,2DY3,2E54,2E7I,2E7J,2ECO,2ECP,2ECQ,2EFY,2EGY,2EH6,2EIY,2EJ3,2EO5,2F3P,2F3Q,2F3S,2F3U,2FET,2FF5,2FFR,2FNI,2FYF,2GPA,2GPB,2GSA,2HDK,2HG8,2HGW,2HGX,2HHF,2HP1,2HP2,2HZP,2ISQ,2J66,2J9Z,2JC3,2JG2,2J

					IS,2JJE,2JJG,2JJH,2NMP,2NV9,2O1B,2OO0,2ORD,2P69,2PB0,2PB2,2PQM,2PRI,2PRJ,2QGH,2QLL,2QN7,2QN8,2QN9,2RH9,2RHG,2RJG,2RKB,2SFP,2SKC,2SKD,2SKE,2TOD,2UZP,2V03,2VD8,2VGS,2VGT,2VGU,2VGV,2VGW,2VGZ,2VI8,2VI9,2VIA,2VIB,2VLH,2VMN,2VMO,2VMP,2VMQ,2VMR,2VMS,2VMT,2VMU,2VMV,2VMW,2VMX,2VMY,2VMZ,2VYC,2W7D,2W7E,2W7F,2W7G,2W7H,2W7I,2W7J,2W7K,2W7L,2W7M,2W8T,2W8U,2W8V,2WK8,2WK9,2WSY,2X3L,2X5D,2X5F,2X8U,2XH1,2Y4R,2YCT,2YHK,2YKU,2YKX,2YKY,2YOB,2YRR,2YXX,2Z1Y,2Z1Z,2Z20,2Z67,2ZB2,2ZGI,2ZP7,2ZSJ,2ZUK,2ZY2,2ZY3,2ZY4,2ZY5,3A2B,3A8U,3A9X,3A9Y,3A9Z,3AAT,3AMV,3ANU,3ANV,3AOV,3AOW,3ATH,3AWN,3AWO,3B1C,3B1D,3B8T,3B8U,3B8V,3B8W,3BB8,3BD7,3BD8,3BDA,3BM5,3BV0,3C5Q,3CEH,3CEJ,3CEM,3CEP,3CO8,3COG,3CSW,3DD1,3DDS,3DDW,3DOD,3DTG,3DU4,3DWG,3DWI,3DXV,3DXW,3DYD,3E5P,3E77,3E9K,3EI8,3EIB,3ELE,3F9T,3FCR,3FDD,3FHX,3G8M,3GJU,3GPB,3HMK,3HQT,3HY8,3I16,3I5T,3IF2,3IHJ,3II0,3ISL,3JZ6,3K28,3K7Y,3KEU,3KGW,3KKI,3KOW,3KP1,3L6B,3L6C,3L8A,3LV2,3LVJ,3LVK,3LVL,3LVM,3LY1,3MEB,3N29,3N2O,3NYS,3NZP,3O05,3PC2,3PPL,3QBO,3R79,3RBF,3RCH,3SS7,3SS9,3TAT,3TFT,3TQX,3UYY,3UZB,3UZO,3VAX,3VBE,3VOM,3VSA,3VSC,3VSD,3WQC,3WQD,3WQE,3WQF,3WQG,3WWH,3WWI,3WWJ,3X43,3ZCP,3ZCQ,3ZCR,3ZCS,3ZCT,3ZCU,3ZCV,3ZEI,3ZRP,3ZZJ,4A0F,4A0G,4A0H,4A0R,4A3Q,4A6T,4A72,4ADB,4ADC,4ADD,4AEC,4AH3,4AO9,4ATP,4ATQ,4AZJ,4AZK,4B98,4B9B,4BEQ,4BEU,4BF5,4BMK,4BQ0,4BQE,4BQF,4BQI,4CBR,4CBS,4CE5,4CHI,4CMD,4COO,4CTM,4CTN,4CTO,4CVQ,4CXQ,4CXR,4DAA,4DGT,4DQ6,4E1O,4EB5,4EB7,4EMY,4F4F,4FL0,4GPB,4GSA,4H67,4H6D,4HT3,4JE5,4JEY,4JF1,4K6N,4L0D,4L0O,4L27,4L28,4L3V,4LMA,4LMB,4LW2,4LW4,4M2J,4M2K,4M2M,4N0W,4NOG,4O6Z,4OBU,4OT8,4OTL,4PB3,4PB4,4PB5,4PCU,4PFF,4PFN,4PPM,4QGR,4QYS,4R2N,4R8D,4RKD,4S1I,4UHM,4UHN,4UHO,4UOX,4UOY,4UQV,4V15,4W1V,4W1W,4W1X,4W5K,4WBT,4WR3,4WX2,4WYA,4WYC,4WYD,4WYE,4WYF,4WYG,4XAU,4XEW,4XJL,4XJM,4XJO,4XJP,4XUG,4Y0H,4Y6G,4YI3,4YI5,4YSN,4YUA,4YWR,4Z5X,4ZGY,4ZLV,4ZM3,4ZM4,4ZQC,4ZU6,4ZWM,5B36,5BJ3,5BJ4,5BW6,5BWA,5BWR,5BWT,5BWU,5BWV,5BWW,5BWX,5C3U,5C6U,5CE8,5CGQ,5CR5,5CVC,5D5S,5D84,5D86,5D87,5DAA,5DDS,5DX5,5EAA,5EQC,5EY5,5G0A,5G2P,5GPB,5HDM,5HNE,5HXX,5I5S,5I5T,5I5U,5I5V,5I5W,5I5Y,5I60,5I6D,5I7A,5I7H,5I7O,5I7R,5I90,5IKO,5IKP,5IW8,5IWC,5IWQ,5J

					8Q,6GPB,7AAT,7GPB,7ODC,8AAT,8GPB,9GPB
PRZ	C9 H14 N2 O	166.22	2-ISOBUTYL-3-METHOXYPYRAZINE	CC(C)Cc1c(nccn1)OC	1DZK,1GT1,1HQP,1QY1,1YP6,2NND,2P70
Q14	C18 H27 N5 O2	345.44	6-[({(2S)-1-AMINO-4-[(6-AMINO-4-METHYLPYRIDIN-2-YL)METHOXY]BUTAN-2-YL}OXY)METHYL]-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)COCC[C@@H](CN)OCc2cc(cc(n2)N)C	4K5F,4K5J,4UGD

Q1T	C19 H25 N5	323.43	3-(2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL)-5-(METHYL(2-(METHYLAMINO)ETHYL)AMINO)BENZONITRILE	Cc1cc(nc(c1)N)CCc2cc(cc(c2)N(C)CCNC)C#N	4UGJ,4UH4,4UH6
QSO	C16 H12 O5	284.27	5,7-DIHYDROXY-3-(4-METHOXYPHENYL)-4H-CHROMEN-4-ONE	COc1ccc(cc1)C2=COc3cc(cc(c3C2=O)O)O	2QYO,4FJ2
RCO	C6 H6 O2	110.11	RESORCINOL	c1cc(cc(c1)O)O	1EVR,1QIZ,2OLY,2OLZ,2OM0,2OM1,2OMH,2OMI,2W44,3AQT,3ZU1,4AJX,4DM3,4E49,4QOO,4Z10
RIT	C37 H48 N6 O5 S2	720.94	RITONAVIR	CC(C)c1nc(cs1)CN(C)C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](Cc2ccccc2)C[C@@H]([C@H](Cc3ccccc3)NC(=O)OCc4cncs4)O	1HXW,1N49,1RL8,1SH9,2B60,3NDW,3NDX,3NXU,3PRS,3Q70,3TNE,4EYR,4NJV
RW1	C10 H8 N2	156.19	4-PHENYLPYRIMIDINE	c1ccc(cc1)c2ccncn2	3B9S,3KAN,3KER
S42	C23 H29 N5	375.52	6-[(2R)-3-AMINO-2-{3-[2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL]PHENYL}PROPYL]-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)CCc2cccc(c2)[C@@H](Cc3cc(cc(n3)N)C)CN	4CTU

S4M	C14 H24 N6 O3 S	356.44	5'-[(S)-(3-AMINOPROPYL)(METHYL)-LAMBDA~4~-SULFANYL]-5'-DEOXYADENOSINE	C[S@@H](CCCN)C[C@@H]1[C@H]([C@H]([C@@H](O1)n2cnc3c2ncnc3N)O)O	2O0L,2PT6,2PT9,4BP3,4YUV,4YUW,4YUX,4YUY,4YUZ,4YV0,4YV1,4YV2
S85	C18 H25 F N4	316.42	N1-(3-(2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL)-5-FLUOROPHENYL)-N1,N2-DIMETHYLETHANE-1,2-DIAMINE	Cc1cc(nc(c1)N)CCc2cc(cc(c2)F)N(C)CCNC	4UH3,4UH9
URO	C6 H6 N2 O2	138.13	(2E)-3-(1H-IMIDAZOL-4-YL)ACRYLIC ACID	c1c(nc[nH]1)C=CC(=O)O	1UWK,1W1U
VNJ	C8 H9 N O	135.16	2-AMINOACETOPHENONE	CC(=O)c1ccccc1N	4CZ1
VXL	C31 H43 N3 O10 S	649.75	(3R,3AS,4R,6AR)-4-[2-(METHYLAMINO)-2-OXOETHOXY]HEXAHYDROFURO[2,3-B]FURAN-3-YL [(2S,3R)-3-HYDROXY-4-{[(4-METHOXYPHENYL)SULFONYL](2-METHYLPROPYL)AMINO}-1-PHENYLBUTAN-2-YL]CARBAMATE	CC(C)CN(C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]2CO[C@@H]3[C@H]2[C@H](CO3)OCC(=O)NC)O)S(=O)(=O)c4ccc(cc4)OC	5AHC

XFJ	C21 H25 N5	347.46	6,6'-(PYRIDINE-3,5-DIYLDIETHANE-2,1-DIYL)BIS(4-METHYLPYRIDIN-2-AMINE)	Cc1cc(nc(c1)N)CCc2cc(cnc2)CCc3cc(cc(n3)N)C	3N5T,3N5W,3N61,3N62,3N65,3N69,3N6B,3N6F,4UG6
XFK	C21 H25 N5	347.46	6-(2-{5-[2-(2-AMINO-6-METHYLPYRIDIN-4-YL)ETHYL]PYRIDIN-3-YL}ETHYL)-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)CCc2cc(cnc2)CCc3cc(nc(c3)N)C	3N5S,3N5X,3N6A,3N6G,4UG5
XJH	C22 H23 N3 O2	361.44	6-{[(3R,4R)-4-(3-PHENOXYPHENOXY)PYRROLIDIN-3-YL]METHYL}PYRIDIN-2-AMINE	c1ccc(cc1)Oc2cccc(c2)O[C@H]3CNC[C@H]3Cc4cccc(n4)N	3N2R
ZLP	C32 H44 N4 O10 S	676.78	(3R,3AS,4R,6AR)-4-(2-METHOXYETHOXY)HEXAHYDROFURO[2,3-B]FURAN-3-YL {(2S,3R)-3-HYDROXY-4-[{[(2Z)-2-(METHYLIMINO)-2,3-DIHYDRO-1,3-BENZOXAZOL-6-YL]SULFONYL}(2-METHYLPROPYL)AMINO]-1-PHENYLBUTAN-2-YL}CARBAMATE	CC(C)CN(C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]2CO[C@@H]3[C@H]2[C@H](CO3)OCCOC)O)S(=O)(=O)c4ccc5c(c4)O/C(=N\C)/N5	5AHA

Annexe 4.3 Tableau récapitulatif des ligands situés entre 2 et 2.5Å d'une interface protéine-protéine

Ligand ID	Ligand Formula	Ligand MW	Ligand Name	Ligand SMILES	Instance PDB IDs (All)
08R	C21 H30 N4 O2	370.49	6-{[(3S,4S)-4-{2-[(2-METHOXYBENZYL)AMINO]ETHOXY}PYRROLIDIN-3-YL]METHYL}-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)C[C@H]2CNC[C@H]2OCCNCc3ccccc3OC	3TYM
0NX	C10 H9 N O2	175.19	(5-PHENYL-1,2-OXAZOL-3-YL)METHANOL	c1ccc(cc1)c2cc(no2)CO	3VQ4
0Q4	C40 H70 N11 O8 1	833.06	N-[(2R)-2-({N~5~-[AMINO(IMINIO)METHYL]-L-ORNITHYL-L-VALYL}AMINO)-4-METHYLPENTYL]-L-PHENYLALANYL-L-ALPHA-GLUTAMYL-L-ALANYL-L-NORLEUCINAMIDE	CCCC[C@@H](C(=O)N)NC(=O)[C@H](C)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](Cc1ccccc1)NC[C@H](CC(C)C)NC(=O)[C@H](C(C)C)NC(=O)[C@H](CCCNC(=[NH2+])N)N	1A8K,1A94,1BAI,1DAZ,1DW6,1EBK,1FFF,1FG8,1K1T,1K1U,1K2B,1K2C,2AOE
0ZT	C38 H49 N5 O8	703.83	N-{(2S,3S)-3-[(TERT-BUTOXYCARBONYL)AMINO]-2-HYDROXY-4-PHENYLBUTYL}-L-PHENYLALANYL-L-ALPHA-GLUTAMYL-L-PHENYLALANINAMIDE	CC(C)(C)OC(=O)N[C@@H](Cc1ccccc1)[C@H](CN[C@@H](Cc2ccccc2)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc3ccccc3)C(=O)N)O	1FQX,1ZJ7,1ZSR

17Z	C11 H10 O2	174.20	2-METHYLNAPHTHALENE-1,4-DIOL	Cc1cc(c2ccccc2c1O)O	4HQM
1FG	C21 H22 N4 O2	362.43	2-({[2-(3,4-DIHYDROQUINOLIN-1(2H)-YL)-2-OXOETHYL](METHYL)AMINO}METHYL)QUINAZOLIN-4(1H)-ONE	CN(CC1=NC(=O)c2ccccc2N1)CC(=O)N3CCCc4c3cccc4	4IG0
1KM	C6 H14 N O6 P S	259.21	(2R)-2-AZANYL-3-[(1R,2S)-2-OXIDANYL-1-PHOSPHONO-PROPYL]SULFANYL-PROPANOIC ACID	C[C@@H]([C@H](P(=O)(O)O)SC[C@@H](C(=O)O)N)O	4JH7,4JH9
1SQ	C9 H8 N2	144.18	ISOQUINOLIN-1-AMINE	c1ccc2c(c1)ccnc2N	2OHK,3KPW,4YUY
215	C27 H27 N5 O2	453.54	(1Z)-5-(2-{4-[2-(DIMETHYLAMINO)ETHOXY]PHENYL}-5-PYRIDIN-4-YL-1H-IMIDAZOL-4-YL)INDAN-1-ONE OXIME	CN(C)CCOc1ccc(cc1)c2[nH]c(c(n2)c3ccc\4c(c3)CC/C4=N/O)c5ccncc5	2FB8
3EM	C43 H51 N5 O8 S	797.96	N~2~-({[7-(diethylamino)-2-oxo-2H-chromen-4-yl]methoxy}carbonyl)-N-[(2S,4S,5S)-4-hydroxy-1,6-diphenyl-5-{[(1,3-thiazol-5-ylmethoxy)carbonyl]amino}hexan-2-yl]-L-valinamide	CCN(CC)c1ccc2c(c1)OC(=O)C=C2COC(=O)N[C@@H](C(C)C)C(=O)N[C@@H](Cc3ccccc3)C[C@@H]([C@H](Cc4ccccc4)NC(=O)OCc5cncs5)O	4U7Q

3MI	C14 H7 CL2 N O3	308.12	2-(3,5-DICHLOROPHENYL)-1,3-BENZOXAZOLE-6-CARBOXYLIC ACID	c1cc2c(cc1C(=O)O)oc(n2)c3cc(cc(c3)Cl)Cl	3TCT,4HIS
77F	C31 H38 N2 O6 S	566.71	N-[(2S,3R)-4-{(CYCLOHEXYLMETHYL)[(4-METHOXYPHENYL)SULFONYL]AMINO}-3-HYDROXY-1-PHENYLBUTAN-2-YL]-3-HYDROXYBENZAMIDE	COc1ccc(cc1)S(=O)(=O)[N@@](CC2CCCCC2)C[C@H]([C@H](Cc3ccccc3)NC(=O)c4cccc(c4)O)O	3SAA
7S7	C22 H28 N6	376.50	6-(3-AMINO-2-(6-(2-(6-AMINO-4-METHYLPYRIDIN-2-YL)ETHYL)PYRIDIN-2-YL)PROPYL)-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)CCc2cccc(n2)[C@H](Cc3cc(cc(n3)N)C)CN	4CTV
A76	C44 H58 N8 O6	794.99	N-{1-BENZYL-(2R,3R)-2,3-DIHYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)[C@H]([C@@H]([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3ccccn3)O)O)NC(=O)N(C)Cc4ccccn4	1HVL

A77	C44 H58 N8 O6	794.99	N-{1-BENZYL-(2R,3S)-2,3-DIHYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)[C@@H]([C@@H]([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3ccccn3)O)O)NC(=O)N(C)Cc4ccccn4	1HVI,1HVS
A78	C44 H60 N8 O5 2	781.01	N-{1-BENZYL-3-HYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)C[C@@H]([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3cccc[nH+]3)O)NC(=O)N(C)Cc4cccc[nH+]4	1HVJ
A79	C44 H58 N8 O6	794.99	N-{1-BENZYL-(2S,3S)-2,3-DIHYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)[C@@H]([C@H]([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3ccccn3)O)O)NC(=O)N(C)Cc4ccccn4	1HVC,1HVK

A85	C44 H56 F2 N8 O6	830.97	N-{1-BENZYL-2,2-DIFLUORO-3,3-DIHYDROXY-4-[3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRYLAMINO]-5-PHENYL-PENTYL}-3-METHYL-2-(3-METHYL-3-PYRIDIN-2-YLMETHYL-UREIDO)-BUTYRAMIDE	CC(C)[C@@H](C(=O)N[C@@H](Cc1ccccc1)C(C([C@H](Cc2ccccc2)NC(=O)[C@H](C(C)C)NC(=O)N(C)Cc3ccccn3)(F)F)(O)O)NC(=O)N(C)Cc4ccccn4	1DIF
ADE	C5 H5 N5	135.13	ADENINE	c1[nH]c2c(n1)c(ncn2)N	1AHA,1BJQ,1CB0,1D2A,1GIU,1HQC,1IFS,1J1R,1JH8,1JYS,1LPD,1LU1,1M2T,1MUD,1NLI,1OD2,1OD4,1Q8Y,1QB7,1QCI,1QD2,1S2D,1VRL,1WEI,1WTA,1XE8,1Y26,1YXM,1Z5N,1Z8D,1ZN7,2GA4,2H8G,2ICS,2P8N,2PQJ,2PUB,2PUE,2QES,2QET,2QLU,2QTT,2XOC,2XOY,2YED,3A7I,3BSF,3E9R,3K9W,3KPV,3KU0,3LE7,3LGS,3LQV,3MRY,3NG9,3NM6,3ONE,3PAO,3QUI,3RL9,3RYS,3S99,3TPV,3U6Z,3U70,3UJO,3V2K,3W52,3WAZ,4BMX,4D8V,4DAO,4DC2,4F1W,4FBA,4FBB,4G89,4I30,4JOS,4JWT,4KQF,4L0M,4L5C,4L6I,4LNA,4LW0,4M1E,4NSN,4OJT,4P14,4PR3,4QAR,4QEZ,4TRC,4TZX,4TZY,4XGP,4XJ7,4XNR,4YMI,5DK6,5DYX
B87	C22 H25 N5 O S	407.53	4-(1-BENZOTHIOPHEN-2-YL)-6-[4-(2-OXO-2-PYRROLIDIN-1-YLETHYL)PIPERAZIN-1-YL]PYRIMIDINE	c1ccc2c(c1)cc(s2)c3cc(ncn3)N4CCN(CC4)CC(=O)N5CCCC5	3IPY
BE2	C7 H7 N O2	137.14	2-AMINOBENZOIC ACID	c1ccc(c(c1)C(=O)O)N	1AN9,1C0I,1E8N,1F8S,1ZFP,1ZYK,2E4A,2GVQ,2HU8,2JB3,2YR6,3H78,3T44,4OWV,4X5D,4X5E,4YI7,5FJN

BEN	C7 H8 N2	120.15	BENZAMIDINE	[H]/N=C(\c1ccccc1)/N	1A0J,1ANB,1ANC,1AND,1ANE,1BIT,1BRA,1BTY,1CC7,1CC8,1CE5,1DPO,1DWB,1EAX,1H4W,1J14,1J15,1J16,1J8A,1JBU,1KLI,1L2E,1LO6,1LPU,1LR4,1MBQ,1NKZ,1NSA,1OSS,1RTF,1S0R,1TRM,1V11,1V16,1V1M,1V2J,1V2L,1V2M,1V2S,1V2U,1V2V,1W80,1WRI,1ZHM,1ZHP,1ZHR,2AER,2AIQ,2ASS,2AST,2AYW,2BLV,2BLW,2BMV,2BPQ,2CKR,2CKS,2EEK,2GLL,2GLM,2GLP,2GNN,2J9N,2O8U,2OQ5,2OXS,2PKA,2TBS,2TIO,2TRM,2VJ0,2VZB,2Y46,2Y5Z,2Y8D,2ZFQ,2ZI2,2ZIQ,2ZPQ,2ZPR,2ZPS,3ATL,3B3J,3B7J,3BB8,3BCX,3BEU,3BG8,3CF8,3CF9,3D04,3D49,3DOY,3DOZ,3DP0,3DP1,3DP2,3DP3,3ED0,3GY7,3I78,3ITI,3M7O,3MFJ,3MI4,3MXN,3NQ8,3NQV,3P70,3P8G,3PLB,3PTB,3PWB,3QK1,3RXE,3RXQ,3RXU,3RXV,3T25,3T26,3T27,3T28,3T29,3TAY,3TH2,3TPK,3UNQ,3UNR,3UY9,3VYW,3W5S,3WJP,3WJQ,3WJR,3ZSN,4COF,4D8T,4D8W,4D92,4D96,4D97,4D99,4D9B,4D9C,4D9E,4D9F,4DSO,4E2K,4E3Q,4EDG,4EDK,4EDR,4EDT,4EDV,4EE1,4EMN,4EQM,4HZE,4I06,4I8G,4I8H,4I8J,4I8K,4I8L,4IBL,4IE2,4IE3,4IXU,4IXV,4JNN,4JPU,4KPM,4N8Z,4NCY,4NFE,4NVC,4P1H,4TPY,4UEH,4UQU,4UQW,4UR0,4UR1,4XV8,4YTA,5AUK,5C50,5CAJ,5F6M,5FXL
BMC	C12 H10 N2 O4	246.22	3-(1,3-BENZODIOXOL-5-YL)-1-METHYL-1H-PYRAZOLE-5-CARBOXYLIC ACID	Cn1c(cc(n1)c2ccc3c(c2)OCO3)C(=O)O	3AO3
BT6	C6 H6 S	110.17	BENZENETHIOL	c1ccc(cc1)S	3HSR
BTN	C10 H16 N2 O3 S	244.31	BIOTIN	C1[C@H]2[C@@H]([C@@H](S1)CCCCC(=O)O)NC(=O)N2	1AVD,1BDO,1BIB,1DF8,1F27,1HXD,1KQS,1LUQ,1MEP,1MK5,1N43,1N9M,1NDJ,1NQM,1STP,1SWD,1SWE,1SWG,1SWK,1SWN,1SWP,1SWR,1SWT,1VQ6,1VQN,1WBI,1WPY,1XNY,1Y52,1Y55,2AVI,2B8G,2BDO,2C1Q,2C4I,2DTH,2DTO,2DXT,2EJ9,2EJF,2EJG,2F01,2FYK,2GH7,2IZF,2IZG,2IZH,2IZI,2IZJ,2JGS,2RTD,2RTE,2RTF,2RTG,2UYW,2UZ2,2Y3F,2ZGW,2ZSC,3D9L,3EFR,3EFS,3EW2,3G8C,3IB9,3MG5,3O34,3RDM,3RDO,3RKY,3RY2,3SZJ,3T2W,3T6F,3V8K,3WYP,3WYQ,3WZN,4BBO,4BCS,4BJ8,4DVE,4EKV,4GD9,4GDA,4GGZ,4IRW,4JNJ,4JXT,4LOC,4MFE,4Q6S,4Q94,4Q96,4WVP,4YVB,4Z28
BZQ	C13 H10 O	182.22	DIPHENYLMETHANONE	c1ccc(cc1)C(=O)c2ccccc2	1DZP,1GT5

CHO	C26 H43 N O5	449.63	GLYCOCHENODEOXYCHOLIC ACID	C[C@H](CCC(=O)NCC(=O)O)[C@H]1CC[C@@H]2[C@@]1(CC[C@H]3[C@H]2C(C[C@H]4[C@@]3(CC[C@H](C4)O)C)O)C	1AHI,1FMC,2B04,2LBA,2LFO,2MM3
D28	C17 H13 F N2 O3 S	344.36	4-{[4-(4-FLUORO-3-METHYLPHENYL)-1,3-THIAZOL-2-YL]AMINO}-2-HYDROXYBENZOIC ACID	Cc1cc(ccc1F)c2csc(n2)Nc3ccc(c(c3)O)C(=O)O	2VD1
DIQ	C11 H19 N O2	197.28	2-METHYL-DECAHYDRO-ISOQUINOLINE-3-CARBOXYLIC ACID	C[N@@]1C[C@H]2CCCC[C@H]2C[C@H]1C(=O)O	1MTB,2FGU,2FGV
DMC	C14 H19 N O3	249.31	3-(4-DIETHYLAMINO-2-HYDROXY-PHENYL)-2-METHYL-PROPIONIC ACID	CCN(CC)c1ccc(c(c1)O)\C=C(/C)\C(=O)O	4GCH
DUP	C9 H16 N3 O13 P3	467.16	2'-DEOXYURIDINE 5'-ALPHA,BETA-IMIDO-TRIPHOSPHATE	C1[C@@H]([C@H](O[C@H]1N2C=CC(=O)NC2=O)CO[P@](=O)(N[P@@](=O)(O)OP(=O)(O)O)O)O	1RN8,1RNJ,1SIX,1SJN,2BT1,2CIC,2D4N,2FMQ,2FMS,2HQU,2HXD,2OKE,2PFN,2PFO,2PY4,2XCE,2XY3,2YAY,3C2K,3EHW,3H6D,3HZA,3I93,3LOJ,3P48,3S9H,3SCX,3SI6,3SJJ,3SQ0,3SQ1,3TP1,3TPN,3TPS,3TPW,3TPY,3TQ3,3TQ4,3UIQ,3ZEZ,3ZF0,3ZF1,3ZF4,3ZF5,4GCY,4GV8,4JWM,4JWN,4KHQ,4M04,4M9J,4OOP,4R65,4R66,4WRK,4YD1,5CCT,5IIN

E21	C45 H74 O12	807.06	21-hydroxy-oligomycin	CC[C@H]1/C=C/C=C/C[C@H]([C@@H]([C@@](C(=O)[C@H]([C@@H]([C@H](C(=O)[C@H]([C@@H]([C@H](/C=C/C(=O)O[C@@H]2[C@@H]([C@H](C[C@H]1O)O[C@@]3([C@H]2C)CC[C@@H]([C@@H](O3)C[C@@H](C)O)C)C)C)O)C)C)O)C)(C)O)O)C	5BQJ
EFO	C45 H74 O11	791.07	OLIGOMYCIN A	CC[C@@H]\1CC[C@H]2[C@H]([C@H]([C@@H]([C@@H]([C@]3(O2)CC[C@@H]([C@@H](O3)C[C@@H](C)O)C)C)OC(=O)/C=C/[C@@H]([C@H]([C@@H](C(=O)[C@@H]([C@H]([C@@H](C(=O)[C@]([C@H]([C@@H](C/C=C/C=C1)C)O)(C)O)C)O)C)C)O)C)C	3WGV,4F4S,4WQ0,5BPS

FBG	C9 H7 F O4	198.15	6-FLUORO-4H-1,3-BENZODIOXINE-8-CARBOXYLIC ACID	c1c(cc(c2c1COCO2)C(=O)O)F	3VQB
G5A	C12 H17 N7 O7 S	403.37	5'-O-(GLYCYLSULFAMOYL) ADENOSINE	c1nc(c2c(n1)n(cn2)[C@H]3[C@@H]([C@@H]([C@H](O3)COS(=O)(=O)NC(=O)CN)O)O)N	3HXV,3HY0,3MF1,4H2T,4H2X,5F5W
GDE	C7 H6 O5	170.12	3,4,5-TRIHYDROXYBENZOIC ACID	c1c(cc(c(c1O)O)O)C(=O)O	3WKU,3WPM,3WR3,3WR4,3WR9,3WRB,4IC0,4J0H,4Z5X
GDN	C16 H19 N5 O10 S	473.41	GLUTATHIONE S-(2,4 DINITROBENZENE)	c1cc(c(cc1[N+](=O)[O-])[N+](=O)[O-])SC[C@@H](C(=O)NCC(=O)O)NC(=O)CC[C@@H](C(=O)O)N	18GS,1GSQ,1HNA,1HNB,1HNC,1VF3,1XWK,4ZBB,5GST
GEN	C15 H10 O5	270.24	GENISTEIN	c1cc(ccc1C2=COc3cc(cc(c3C2=O)O)O)O	1QKM,1X7J,1X7R,2QA8,3KGT,3KGU,4FJ1,5AUZ,5AV4
GG6	C3 H9 O5 P	156.07	[(1S,2S)-1,2-DIHYDROXYPROPYL]PHOSPHONIC ACID	C[C@@H]([C@@H](O)P(=O)(O)O)O	2P7Q,2RL2
HW0	C21 H31 N5 O	369.51	6-{[(3R,4R)-4-{[5-(6-AMINOPYRIDIN-2-YL)PENTYL]OXY}PYRROLIDIN-3-YL]METHYL}-4-METHYLPYRIDIN-2-AMINE	Cc1cc(nc(c1)N)C[C@@H]2CNC[C@@H]2OCCCCCc3cccc(n3)N	3UFW,4CWX

IPH	C6 H6 O	94.11	PHENOL	c1ccc(cc1)O	1AI0,1AI7,1AIY,1FJW,1FOH,1JHX,1JHY,1LI2,1LPH,1MPJ,1PN0,1Q4V,1QIY,1RWE,1V03,1W8P,1WAV,1XDA,1XU5,1XW7,1ZEG,1ZNJ,2AIY,2AS3,2J9N,2OLD,2OMB,2OMN,2PZV,2VE7,2WS6,2WS7,3AIY,3F39,3GKY,3GUO,3JSD,3KMH,3KQ6,3NC0,3NX8,3P2X,3P33,3Q3O,3ROV,3U3E,3V19,3V1G,3ZQR,3ZS2,4A71,4AIY,4AJZ,4AKJ,4F3T,4H07,4HDS,4I7L,4JMW,4OLA,4P65,4W5N,4W5O,4W5Q,4W5R,4W5T,4Z4C,4Z4D,4Z4E,4Z4F,4Z4G,4Z4H,4Z4I,5AIY,5BQQ,5FRW,5JS1,5JS2,5KBE,5KI6
JI1	C13 H22 N4 O	250.34	3-({(3S,4S)-4-[(6-AMINOPYRIDIN-2-YL)METHYL]PYRROLIDIN-3-YL}AMINO)PROPAN-1-OL	c1cc(nc(c1)N)C[C@H]2CNC[C@H]2NCCCO	3B3M
JI2	C12 H21 N5	235.33	N-{(3S,4S)-4-[(6-AMINOPYRIDIN-2-YL)METHYL]PYRROLIDIN-3-YL}ETHANE-1,2-DIAMINE	c1cc(nc(c1)N)C[C@H]2CNC[C@H]2NCCN	3B3N,3DQR
M7K	C18 H18 F N3	295.36	7-{2-[(3-FLUOROBENZYL)AMINO]ETHYL}QUINOLIN-2-AMINE	c1cc(cc(c1)F)CNCCc2ccc3ccc(nc3c2)N	4CAN
MJZ	C17 H26 N4 O17 P2	620.36	(2S,3S,4R,5R,6R)-5-(ACETYLAMINO)-4-AMINO-6-{[(R)-{[(R)-{[(2R,3S,4R,5R)-5-(2,4-DIOXO-3,4-DIHYDROPYRIMIDIN-1(2H)-YL)-3,4-DIHYDROXYTETRAHYDROFURAN-2-YL]METHOXY}(HYDROXY)PHOSPHORYL]OXY	CC(=O)N[C@@H]1[C@H]([C@@H]([C@H](O[C@@H]1O[P@@](=O)(O)O[P@](=O)(O)OC[C@@H]2[C@H]([C@H]([C@@H](O2)N3C=CC(=O)NC3=O)O)O)C(=O)O)O)N	3MQH

			}(HYDROXY)PHOSPHORYL]OXY}-3-HYDROXYTETRAHYDRO-2H-PYRAN-2-CARBOXYLIC ACID		
ML6	C19 H20 F N3	309.39	7-({[3-(3-FLUOROPHENYL)PROPYL]AMINO}METHYL)QUINOLIN-2-AMINE	c1cc(cc(c1)F)CCCNCc2ccc3ccc(nc3c2)N	4CAP
NGV	C21 H14 O7	378.34	METHYL 5,7-DIHYDROXY-2-METHYL-4,6,11-TRIOXO-3,4,6,11-TETRAHYDROTETRACENE-1-CARBOXYLATE	CC1=C(c2cc3c(c(c2C(=O)C1)O)C(=O)c4c(cccc4O)C3=O)C(=O)OC	1SJW,2F98
NVU	C9 H7 N O3	177.16	2-(1,2-BENZOXAZOL-3-YL)ETHANOIC ACID	c1ccc2c(c1)c(no2)CC(=O)O	3ZL6,4CK2,5I5S
OAC	C10 H12 O3	180.20	TRANS-O-HYDROXY-ALPHA-METHYL CINNAMATE	CC(Cc1ccccc1O)C(=O)O	3GCH
OXM	C2 H3 N O3	89.05	OXAMIC ACID	C(=O)(C(=O)O)N	1A5Z,1H17,1I0Z,1I10,1LDG,1LDM,1LDN,1LTH,1OC4,1T2E,2DLD,2V5K,2V7P,2XXJ,3H3F,3PFL,3UQN,3VPH,4AJ1,4KNL,4LOC,4ND1,4OL9,4PLG,4PLH,4PLT,4PLZ,5A1T,5ES3,5H9O,5HJR,5K9F,9LDB,9LDT
PH4	C5 H13 N O10 P2	309.11	3-{HYDROXY[(PHOSPHONOOXY)ACETYL]AMINO}PROPYL DIHYDROGEN PHOSPHATE	C(CN(C(=O)COP(=O)(O)O)O)COP(=O)(O)O	3C56

PI8	C41 H54 N4 O6	698.90	N-13-[(10S,13S)-9,12-DIOXO-10-(2-BUTYL)-2-OXA-8,11-DIAZABICYCLO [13.2.2] NONADECA-15,17,18-TRIENE] (2R)-BENZYL-(4S)-HYDROXY-5-AMINOPENTANOIC (1R)-HYDROXY-(2S)-INDANEAMIDE	CC[C@H](C)[C@H]1C(=O)NCCCCCOc2ccc(cc2)C[C@@H](C(=O)N1)NC[C@H](C[C@@H](Cc3ccccc3)C(=O)N[C@H]4c5ccccc5C[C@H]4O)O	1D4K
PLP	C8 H10 N O6 P	247.14	PYRIDOXAL-5'-PHOSPHATE	Cc1c(c(c(cn1)COP(=O)(O)O)C=O)O	1A3G,1A50,1A5A,1A5B,1A5S,1AAM,1AAW,1AHE,1AHF,1AHG,1AHP,1AHX,1AHY,1AKA,1ARI,1ARS,1ART,1ASA,1ASB,1ASF,1ASG,1ASM,1ASN,1AXR,1AY4,1AY5,1AY8,1B4D,1B4X,1B54,1B5O,1B5P,1B8G,1B9H,1BJ4,1BJO,1BKS,1BT4,1BX3,1C0N,1C29,1C4K,1C50,1C7G,1C7N,1C8K,1C8L,1C8V,1C9D,1CJ0,1CT5,1CW2,1CX9,1CZC,1CZE,1D2F,1D6S,1DAA,1DGD,1DGE,1DJE,1DJU,1DKA,1DTY,1E1Y,1E4O,1E5F,1ECX,1EG5,1EKF,1EKP,1EKV,1ELQ,1EM6,1EQB,1ET0,1EXV,1F2D,1F3T,1FA9,1FC0,1FC4,1FCJ,1FG3,1FS4,1FTQ,1FTW,1FTY,1FU4,1FU7,1FU8,1FUY,1G2W,1G4V,1G4X,1G76,1G77,1G78,1G79,1G7W,1G7X,1GBN,1GC3,1GC4,1GCK,1GD9,1GDE,1GEW,1GEX,1GFZ,1GG8,1GGN,1GPA,1GPB,1GPY,1H0C,1H1C,1H5U,1HKV,1HLF,1I1K,1I1L,1I1M,1I29,1I2K,1I43,1I48,1IAX,1IAY,1IBJ,1IJI,1IX6,1IX7,1IX8,1IYD,1J0A,1J0C,1J0E,1J32,1JBQ,1JF9,1JNW,1JS3,1JS6,1K06,1K08,1K3U,1K7E,1K7F,1K7X,1K8X,1K8Y,1K8Z,1KFB,1KFC,1KFK,1KKJ,1KKP,1KL1,1KL2,1KL7,1KMJ,1KMK,1KNW,1KO0,1KTI,1L5Q,1L5R,1L5S,1L5V,1L5W,1L6I,1L7X,1LK9,1LKC,1LS3,1LW4,1LW5,1LWN,1LWO,1M32,1M4N,1M54,1MDX,1MDZ,1MGV,1MLY,1MLZ,1N2T,1N31,1N8P,1NOI,1NOJ,1NOK,1NRG,1O4S,1O61,1OAS,1OAT,1OHV,1OHW,1OHY,1ORD,1P29,1P2B,1P2D,1P2G,1P3W,1P4G,1P4H,1P4J,1P5J,1PG8,1PMM,1QGN,1QIR,1QIS,1QIT,1QJ3,1QJ5,1QM5,1QOP,1QOQ,1QU4,1QZ9,1RCQ,1RFU,1RQX,1RV3,1RV4,1RVU,1RVY,1S07,1SF2,1SFT,1SZS,1SZU,1T3I,1TAR,1TAT,1TDJ,1TJP,1TTP,1TTQ,1TWI,1TYZ,1TZ2,1TZJ,1TZK,1TZM,1U08,1UBS,1UIM,1UIN,1UZU,1V2D,1V2E,1V2F,1V71,1V72,1V8Z,1VE1,1VE5,1VEF,1VFH,1VJO,1VP4,1W23,1W3U,1W7L,1W7M,1W8G,1WBJ,1WDW,1WK

					V,1WRV,1WST,1WTC,1WUT,1WUY,1WV0,1WV1,1WW2,1WW3,1WYU,1WYV,1XC7,1XEY,1XFC,1XI9,1XKX,1XL0,1XL1,1XOI,1XQL,1XRS,1YAA,1YGP,1YJS,1YJY,1YJZ,1YOO,1Z3Z,1Z62,1Z7W,1ZOB,1ZOD,2A1H,2A5H,2ABJ,2AMV,2AQ6,2ASV,2ATI,2AV6,2AW3,2AY1,2AY2,2AY3,2AY4,2AY5,2AY6,2AY7,2AY8,2AY9,2AZD,2BHS,2BHT,2BHX,2BI1,2BI2,2BI3,2BI5,2BI9,2BIA,2BIE,2BIG,2BKW,2BWN,2BWO,2BYJ,2BYL,2C0R,2C2B,2C2G,2C4M,2C7T,2CAN,2CB1,2CFT,2CH1,2CH2,2CIN,2CJD,2CJH,2CLE,2CLF,2CLH,2CLI,2CLK,2COG,2COI,2COJ,2CST,2CTZ,2D1F,2D5Y,2D61,2D63,2D64,2D65,2D66,2D7Y,2D7Z,2DAB,2DGK,2DGL,2DGM,2DH5,2DKB,2DKJ,2DR1,2DY3,2E54,2E7I,2E7J,2ECO,2ECP,2ECQ,2EFY,2EGY,2EH6,2EIY,2EJ3,2EO5,2F3P,2F3Q,2F3S,2F3U,2FET,2FF5,2FFR,2FNI,2FYF,2GPA,2GPB,2GSA,2HDK,2HG8,2HGW,2HGX,2HHF,2HP1,2HP2,2HZP,2ISQ,2J66,2J9Z,2JC3,2JG2,2JIS,2JJE,2JJG,2JJH,2NMP,2NV9,2O1B,2OO0,2ORD,2P69,2PB0,2PB2,2PQM,2PRI,2PRJ,2QGH,2QLL,2QN7,2QN8,2QN9,2RH9,2RHG,2RJG,2RKB,2SFP,2SKC,2SKD,2SKE,2TOD,2UZP,2V03,2VD8,2VGS,2VGT,2VGU,2VGV,2VGW,2VGZ,2VI8,2VI9,2VIA,2VIB,2VLH,2VMN,2VMO,2VMP,2VMQ,2VMR,2VMS,2VMT,2VMU,2VMV,2VMW,2VMX,2VMY,2VMZ,2VYC,2W7D,2W7E,2W7F,2W7G,2W7H,2W7I,2W7J,2W7K,2W7L,2W7M,2W8T,2W8U,2W8V,2WK8,2WK9,2WSY,2X3L,2X5D,2X5F,2X8U,2XH1,2Y4R,2YCT,2YHK,2YKU,2YKX,2YKY,2YOB,2YRR,2YXX,2Z1Y,2Z1Z,2Z20,2Z67,2ZB2,2ZGI,2ZP7,2ZSJ,2ZUK,2ZY2,2ZY3,2ZY4,2ZY5,3A2B,3A8U,3A9X,3A9Y,3A9Z,3AAT,3AMV,3ANU,3ANV,3AOV,3AOW,3ATH,3AWN,3AWO,3B1C,3B1D,3B8T,3B8U,3B8V,3B8W,3BB8,3BD7,3BD8,3BDA,3BM5,3BV0,3C5Q,3CEH,3CEJ,3CEM,3CEP,3CO8,3COG,3CSW,3DD1,3DDS,3DDW,3DOD,3DTG,3DU4,3DWG,3DWI,3DXV,3DXW,3DYD,3E5P,3E77,3E9K,3EI8,3EIB,3ELE,3F9T,3FCR,3FDD,3FHX,3G8M,3GJU,3GPB,3HMK,3HQT,3HY8,3I16,3I5T,3IF2,3IHJ,3II0,3ISL,3JZ6,3K28,3K7Y,3KEU,3KGW,3KKI,3KOW,3KP1,3L6B,3L6C,3L8A,3LV2,3LVJ,3LVK,3LVL,3LVM,3LY1,3MEB,3N29,3N2O,3NYS,3NZP,3O05,3PC2,3PPL,3QBO,3R79,3RBF,3RCH,3SS7,3SS9,3TAT,3TFT,3TQX,3UYY,3UZB,3UZO,3VAX,3VBE,3VOM,3VSA,3VSC,3VSD,3WQC,3WQD,3WQE,3WQF,3WQG,3WWH,3WWI,3WWJ,3X43,3ZCP,3ZCQ,3ZCR,3ZCS,3ZCT,3ZCU,3ZCV,3ZEI,3ZRP,3ZZJ,4A0F,4A0G,4A0H,4A0R,4A3Q,4A6T,4A72,4ADB,4ADC,4ADD,4AEC,4AH3,4AO9,4ATP,4ATQ,4AZJ,4AZK,4B98,4B9B,4BEQ,4BEU,4BF5,4BMK,4BQ0,4BQE,4BQF,4BQI,4CBR,4CBS,4CE5,4CHI,4CMD,4COO,4

					CTM,4CTN,4CTO,4CVQ,4CXQ,4CXR,4DAA,4DGT,4DQ6,4E1O,4EB5,4EB7,4EMY,4F4F,4FL0,4GPB,4GSA,4H67,4H6D,4HT3,4JE5,4JEY,4JF1,4K6N,4L0D,4L0O,4L27,4L28,4L3V,4LMA,4LMB,4LW2,4LW4,4M2J,4M2K,4M2M,4N0W,4NOG,4O6Z,4OBU,4OT8,4OTL,4PB3,4PB4,4PB5,4PCU,4PFF,4PFN,4PPM,4QGR,4QYS,4R2N,4R8D,4RKD,4S1I,4UHM,4UHN,4UHO,4UOX,4UOY,4UQV,4V15,4W1V,4W1W,4W1X,4W5K,4WBT,4WR3,4WX2,4WYA,4WYC,4WYD,4WYE,4WYF,4WYG,4XAU,4XEW,4XJL,4XJM,4XJO,4XJP,4XUG,4Y0H,4Y6G,4YI3,4YI5,4YSN,4YUA,4YWR,4Z5X,4ZGY,4ZLV,4ZM3,4ZM4,4ZQC,4ZU6,4ZWM,5B36,5BJ3,5BJ4,5BW6,5BWA,5BWR,5BWT,5BWU,5BWV,5BWW,5BWX,5C3U,5C6U,5CE8,5CGQ,5CR5,5CVC,5D5S,5D84,5D86,5D87,5DAA,5DDS,5DX5,5EAA,5EQC,5EY5,5G0A,5G2P,5GPB,5HDM,5HNE,5HXX,5I5S,5I5T,5I5U,5I5V,5I5W,5I5Y,5I60,5I6D,5I7A,5I7H,5I7O,5I7R,5I90,5IKO,5IKP,5IW8,5IWC,5IWQ,5J8Q,6GPB,7AAT,7GPB,7ODC,8AAT,8GPB,9GPB
PLR	C8 H12 N O5 P	233.16	(5-HYDROXY-4,6-DIMETHYLPYRIDIN-3-YL)METHYL DIHYDROGEN PHOSPHATE	Cc1c(cnc(c1O)C)COP(=O)(O)O	1PMO,2CFB,2GJ4,2GM9,2IEG,2IEI,3FSL,3FZ8,3GZC,3GZD,3HL2,3MAU,3PIU,3USF,4LNJ,4LNM,4QYS,4ZDL,4ZDO,4ZDP
PTU	C9 H12 N2 S	180.27	2-ETHYL-1-PHENYL-ISOTHIOUREA	CCSC(=Nc1ccccc1)N	1D1V,1K2T
S4M	C14 H24 N6 O3 S	356.44	5'-[(S)-(3-AMINOPROPYL)(METHYL)-LAMBDA~4~-SULFANYL]-5'-DEOXYADENOSINE	C[S@@H](CCCN)C[C@@H]1[C@H]([C@H]([C@@H](O1)n2cnc3c2ncnc3N)O)O	2O0L,2PT6,2PT9,4BP3,4YUV,4YUW,4YUX,4YUY,4YUZ,4YV0,4YV1,4YV2

SKO	C17 H25 N5	299.41	N'-[6-[2-(6-AZANYL-4-METHYL-PYRIDIN-2-YL)ETHYL]PYRIDIN-2-YL]-N,N'-DIMETHYL-ETHANE-1,2-DIAMINE	Cc1cc(nc(c1)N)CCc2cccc(n2)N(C)CCNC	4UGI,4UH0
TCU	C19 H24 O2	284.40	5-HEXYL-2-(2-METHYLPHENOXY)PHENOL	CCCCCCc1ccc(c(c1)O)Oc2ccccc2C	2X22,2X23,4BNM,5COQ,5CP8
THM	C10 H14 N2 O5	242.23	THYMIDINE	CC1=CN(C(=O)NC1=O)[C@H]2C[C@@H]([C@H](O2)CO)O	1E2J,1G0R,1H5R,1KIM,1OT3,1P6X,1P72,1P7C,1RXU,1TLW,1W2G,1ZMX,2B8T,2J9R,2QQ0,2QQE,2VTK,2Y1I,2Y1J,2Z1A,3BCU,3H5Q,3N2I,3ROE,4ESH,4G8J,4HN1,4HO2,4HO4,4HO8,4LCA,4LZW,4OGK,4QSV,4R8J,4TXJ,4UXI,4YEK,4ZU5,5BSZ,5FUV,5FUW,5IDT
TMC	C12 H16 N2 O4	252.27	1-[4-HYDROXY-5-(HYDROXYMETHYL)BICYCLO[3.1.0]HEX-2-YL]-5-METHYLPYRIMIDINE-2,4(1H,3H)-DIONE	CC1=CN(C(=O)NC1=O)[C@H]2C[C@@H]([C@]3([C@@H]2C3)CO)O	1E2K,1E2L
TPV	C31 H33 F3 N2 O5 S	602.67	N-(3-{(1R)-1-[(6R)-4-HYDROXY-2-OXO-6-PHENETHYL-6-PROPYL-5,6-DIHYDRO-2H-PYRAN-3-YL]PROPYL}PHENYL)-5-(TRIFLUOROMETHYL)-2-PYRIDINESULFONAMIDE	CCC[C@]1(CC(=C(C(=O)O1)[C@H](CC)c2cccc(c2)NS(=O)(=O)c3ccc(cn3)C(F)(F)F)O)CCc4ccccc4	1D4S,1D4Y,2O4L,2O4N,2O4P,3SPK,4NJU
TQ9	C5 H7 CL O3	150.56	5-CHLORO-4-OXOPENTANOIC ACID	C(CC(=O)O)C(=O)CCl	2XYH

XFJ	C21 H25 N5	347.46	6,6'-(PYRIDINE-3,5-DIYLDIETHANE-2,1-DIYL)BIS(4-METHYLPYRIDIN-2-AMINE)	Cc1cc(nc(c1)N)CCc2cc(cnc2)CCc3cc(cc(n3)N)C	3N5T,3N5W,3N61,3N62,3N65,3N69,3N6B,3N6F,4UG6
XVE	C18 H18 CL N O3	331.80	PHENYLMETHYL N-[(2S)-4-CHLORO-3-OXO-1-PHENYL-BUTAN-2-YL]CARBAMATE	c1ccc(cc1)C[C@@H](C(=O)CCl)NC(=O)OCc2ccccc2	2XYP,4Q24

Conclusion

L'objectif de cette thèse était de développer des outils cheminformatiques spécifiquemment dédiés aux interfaces protéine-protéine longtemps négligées par la recherche pharmaceutique. Pour ce faire, ce travail décrit la création de plusieurs logiciels permettant de mieux caractériser les interfaces protéine-protéine et de les utiliser à des fins d'identification de modulateurs de petit poids moléculaire. L'avantage de ces développements est qu'ils sont applicables à l'ensemble des structures présentes dans la *Protein Data Bank* (PDB).

Originellement, cette thèse visait à poursuivre le travail de développement d'analyse de sites de liaisons protéine-ligand afin l'appliquer aux interactions protéine-protéine. Ce travail consistait au développement de plusieurs outils, le premier (detectPPI) permet de détecter les zones d'interaction entre les protéines. Nous souhaitions être capables de définir précisément l'ensemble des interactions (liaison hydrogène, liaison hydrophobe, liaison ionique et métallique) entre deux chaines peptidiques. Nous nous sommes rendus compte que toutes les interfaces détectées n'était pas biologiquement pertinentes et avons ainsi développé un classifieur d'interfaces protéine-protéine (IChemPIC). De part leur nature, les structures de protéine obtenues par diffraction des rayons X ne reflètent pas forcément leur état en milieux aqueux phsiologiques. La diffraction des rayons X utilisant la redondance d'informations présente au sein d'un cristal pur de protéine, l'état cristallin entraine une compaction des protéines et donc la création de contacts cristallins artéfactuels en plus des vraies interfaces biologiquement pertinentes. Il est important de distinguer ces contacts cristallins des interfaces biologiquemes car les premiers n'ont aucun intérêt pharmacologique. Le troisième outil (VolSite) développé dutant ma thèse permet de détecter et d'analyser l'ensemble des interfaces présentes à la surface d'une ou plusieurs protéines ; il s'agit de l'évolution d'un outil préalablement développé au laboratoire mais qui s'est affranchi de la présence d'un ligand centré sur la cavité à détecter. Ces routines ont été incorporées dans la suite logicielle IChem, un outil complet de détection et d'analyse des interfaces protéine-protéine.

Dans un second temps, j'ai exploité notre outil de détection de cavité afin générer automatiquement des pharmacophores déduits de la structure de la cavité cible. Les pharmacophores sont le plus souvent utilisés de deux manières distinctes, soit à partir de ligands soi à partir de complexes protéine-ligand. Des pharmacophores basés uniquement sur

la structure de la protéine cible sont possibles mais restent difficilement utilisables car beaucoup trop complexes. J'ai ainsi développé une méthode de création de pharmacophore dont les propriétés sont déduites de la structure 3D de la cavité. Je me suis focalisé sur la méthode à employer afin de sélectionner et optimiser le nombre d'éléments pharamcophoriques ainsi que leurs types. Cette méthode permet la création de pharmacophore de taille raisonnable (35 éléments) avec lesquels il est maintenant possible de rechercher de nouveaux ligands par criblage virtuel. Pour compléter l'étude des pharmacophore, j'ai testé plusieurs méthodes de d'alignement ligand-pharmacophores et ai décidé d'employer une routine basée sur l'alignement de formes. Cette méthode donne de bonnnes poses de ligands que nous ne sommes malheureusement pas encore capables d'évaluer correctement de manière quantitative afin de prioriser une pose unique par ligand.

Dans un troisième temps, j'ai utilisé l'ensemble des outils préalablement développés et les ai appliqués à l'ensemble des structures de la PDB. Nous souhaitions obtenir une cartographie précise de l'ensemble des PPIs d'intérêt pharmacologique de structure 3D connue. Nous avons mis au point un premier flux de travail automatisé et reproductible permettant l'analyse des structures, la détection des interfaces, la prédiction de leur pertinence biologique puis la détection de cavités droguables à l'interface ainsi qu'à sa proximité immédiate. En parallèlele, flux d'analyse des ligands co-cristallisé a aussi été mis au point. Grâce à l'ensemble des données présentes nous avons pu déterminer des statistiques sur la taille des interfaces, leurs propriétés mais aussi sur les cavités et les ligands situés à proximité. Nous avons identifié un grand nombre de cavités droguables allostériques potentielles à proximité des interfaces ainsi que de nombreuses cavités orthostériques directement situés à interface. Nous proposons une nouvelle ontologie des cavités en 4 catégories (interfaciale, bordure, orthostérique, allostérique) selon leur localisation par rapport à l'interface ainsi que leur accessibilité.

Ce travail prouve que beaucoup d'informations disponibles dans la PDB sont restées jusqu'à présent inexploitées. Le developpement d'outils chemiformatiques dédiés a permis l'extraction automatique de données cruciales à la mise au point d'une cartographie 3D précise des PPIs à l'échelle de la PDB. Ce travail ouvre la possibilité d'identifier prochainement des modulateurs de petit poids moléculaire pour un très grand nombre d'interfaces.

Publications

Biological and Crystallographic Protein-Protein Interfaces

Franck Da Silva, Jérémy Desaphy, Guillaume Bret and Didier Rognan

Journal of Chemical Information and Modelling, 2015, 55, 2005-2014

Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015

Inna Slynko, Franck Da Silva, Guillaume Bret and Didier Rognan

Journal of Computer-Aided Molecular Design, 2016, DOI:10.1007/s10822-016-9930-3

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

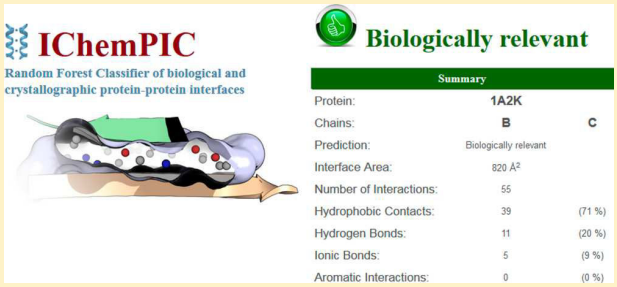
IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein−Protein Interfaces

Franck Da Silva, Jérémy Desaphy,[†] Guillaume Bret, and Didier Rognan*

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS−Université de Strasbourg, 67400 Illkirch, France

Supporting Information

ABSTRACT: Protein−protein interactions are becoming a major focus of academic and pharmaceutical research to identify low molecular weight compounds able to modulate oligomeric signaling complexes. As the number of protein complexes of known three-dimensional structure is constantly increasing, there is a need to discard biologically irrelevant interfaces and prioritize those of high value for potential druggability assessment. A Random Forest model has been trained on a set of 300 protein−protein interfaces using 45 molecular interaction descriptors as input. It is able to predict the nature of external test interfaces (crystallographic vs biological) with accuracy at least equal to that of the best state-of-the-art methods. However, our method presents unique advantages in the early prioritization of potentially ligandable protein−protein interfaces: (i) it is equally robust in predicting either crystallographic or biological contacts and (ii) it can be applied to a wide array of oligomeric complexes ranging from small-sized biological interfaces to large crystallographic contacts.



INTRODUCTION

Protein−protein interactions (PPI) stand at the heart of most pathophysiological situations in living cells and therefore have attracted more and more attention in drug discovery.[1−3] Among the many strategies to identify low molecular weight PPI modulators, rational structure-based approaches have historically played an important role, notably because of the possible integration of biophysical screening of fragment libraries (surface plasmon resonance, isothermal titration calorimetry, nuclear magnetic resonance spectroscopy, mass spectrometry) with X-ray structure determination.[4] To fully exploit the current structural knowledge on druggable targets, it is desirable to ascertain their true oligomeric state as well as their biological relevance. Throughout this article, we will consider as biological any protein−protein complex with a true biological relevance and function (e.g., cell adhesion, cell signaling, immune recognition, transcription). Homo- or hetero-oligomeric complexes resulting either from crystal packing or lacking any known biological function will be considered crystallographic. Unfortunately, inferring the quaternary structure and biological relevance from atomic coordinates in the Protein Data Bank (PDB)[5] is not straightforward. For example, the contents of the asymmetric unit (ASU) deposited in the PDB (the fraction of the crystallographic unit cell that has no crystallographic symmetry) can describe one or several copies of a macromolecule but with no particular indication on which oligomeric state (e.g., monomer, dimer) is the most relevant. Likewise, the ASU may need crystallographic symmetry operations to be applied before reconstituting the beforehand known biologically relevant macromolecular assembly (biological unit). Automated procedures to discriminate, from 3D structures, crystal from biologically relevant and stable interfaces are, therefore, needed to avoid long and costly biochemical experiments such as gel filtration, light scattering, or equilibrium sedimentation.

As a rule of thumb, crystallographic interfaces are generally much smaller (<1000 Å^2) than biologically relevant ones.[6] However, this simple rule suffers from many exceptions since some very important interfaces, like those involving α-helix recognition sites, may be quite small in size (e.g., 780 Å^2 for the p53−mdm2 complex). Many classification methods have been designed, therefore, to directly predict the oligomeric status of protein complexes from atomic coordinates.[7] The very first approach, reported in 1998 as PQS (protein quaternary structure file server),[8] used an empirical scoring function based on several contributions (interface contact area; number of interfacial buried residues, salt bridges, and disulfide bonds; solvation energy of quaternary structure formation). Although it is not perfect (at least 20% of misclassifications were reported by the authors themselves), the PQS server paved the way for many methods that can be grouped in two categories.

A first type of approach, of which PISA[9] is representative, relies on first-principles physics to predict the stability of protein assemblies in solution. For example, PISA explicitly computes Gibbs dissociation free energies to predict the biological relevance of a macromolecular assembly. When applied to a dataset of 218 PDB structures, it achieved a

Received: April 6, 2015
Published: September 7, 2015

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

remarkable success rate of 90% in predicting true biological interfaces.[9] PISA can be considered to be a reference method, as it is currently used to predict quaternary structures of every entry of the RCSB PDB web site. A second group of methods[7,10−17] applies linear or nonlinear regression/classification models to predefined training sets (crystallographic, biological) in order to predict the quaternary structure of external test sets. Many geometrical and chemical complementarity descriptors of the interface can be used to discriminate, with comparable accuracies (ca. 85−90%), crystal from biological contacts. Very often, these methods (e.g., IPAC,[7] DiMoVo,[12] or NOXClass[13]) utilize a machine learning algorithm (support vector machine, decision trees, Bayesian inference) trained on atom or residue-based contact vectors to decide which parameter set is the most adequate for an optimal classification. Residue conservation of interface core residues[18,19] can be added to the above-cited descriptors, as, for example, in EPPIC,[20] to highlight the importance of highly buried core residues at biological interfaces.

To allow them to be compared, most studies have relied on a limited number of benchmarking datasets,[10,21,22] which turned out to be biased toward small crystal and high-affinity large biological interfaces.[7,12,20] As a consequence, most current classification methods have a much lower accuracy when applied to a set of interfaces (biological, crystal) with an equivalent distribution of interface areas. A recently designed dataset[20] paid attention to select true biological and crystallographic interfaces with a hard cutoff with respect to interface areas (mostly above 1000 Å^2). Since we finally aim at identifying potentially ligandable protein−protein interfaces that may be small in size,[23] none of the existing datasets appears to be satisfactory. We therefore designed a hand-curated dataset (FDS set) of 200 biologically relevant nonredundant protein−protein complexes of known X-ray structure, which was further supplemented by an equivalent number of 200 crystal interfaces filtered to span a comparable interface area range. We next used a machine learning algorithm (Random Forest) and 45 molecular interaction descriptors to train a model that, when applied to several external test sets, achieves good accuracy and robustness in distinguishing between crystallographic and biologically relevant interfaces, whatever their size.

COMPUTATIONAL METHODS

Datasets. *FDS Dataset.* Crystallographic interfaces were retrieved from two previously reported datasets.[20,21] First, 141 known monomeric proteins from the Bahadur dataset[21] with a crystalline interface area in the 400−1200 Å^2 range were retrieved as follows. Atomic coordinates of the asymmetric unit were retrieved from the RCSB Protein Data Bank, and one unit cell was reconstructed for each entry using AmberTools14.[24] For each structure and all possible pairs of chains, the interface area IA (eq 1) was measured with MSMS[25] after removing nonprotein atoms (solvent, ligands, ions) and using a probe radius and vertex density of 1.4 Å and 2.0/Å^2, respectively.

$$IA_{A,B} = \frac{(ASA_A + ASA_B) - ASA_{AB}}{2} \quad (1)$$

where $IA_{A,B}$ is the interface area between chains A and B, ASA_A is the solvent-accessible surface area of isolated chain A, ASA_B is the solvent-accessible surface area of isolated chain B, and ASA_{AB} is the solvent-accessible surface area of complex AB.

The largest interface area was kept for each PDB entry. The Bahadur set was then complemented by 82 interfaces from the DCxtal dataset[20] selected to present an interface area in the 1000−1500 Å^2 range for proteins reported to be monomeric in solution. The corresponding PDB files were directly retrieved from the EPPIC web site (http://www.eppic-web.org/ewui/#downloads). Protein redundancy was removed by keeping only one entry in cases where sequence identity between two different entries was above 70%, according to RCSB redundancy rules.[26] The final set of 200 nonredundant crystallographic interfaces (PDB identifier, protein name, chains, interface area, resolution, protein classification) is given in Table S1.

A collection of 200 biologically relevant nonredundant protein−protein interfaces (<70% pairwise sequence identity between any two chains) was manually assembled from the literature according to the following sources: (i) the recently described DCbio dataset[20] of homodimeric biological interfaces (74 PPIs); (ii) the 2P2I database[23] that archives heterodimers for which an X-ray structure exists for the complex, each individual monomer in the free state, and one partner is bound to a low molecular weight inhibitor (10 PPIs); (iii) existing small molecular weight inhibitors[27] for biologically relevant PPIs of known X-ray structure (five PPIs); (iv) cancer-related PPIs[1] of known X-ray structure (eight PPIs); (v) the PPI affinity database[28] of biologically relevant PPIs with available X-ray structures (complex, free state) and known experimental binding constant (54 PPIs); (vi) the dataset of "hot loops" mediated PPIs[29] of known X-ray structures (20 PPIs) and (vi) diverse biologically relevant heterodimeric proteins (18 PPIs). The biological relevance of all of these 200 complexes was checked manually according to known literature data.[20,23,27−29]

The corresponding structures were downloaded from the Protein Data Bank (PDB). Chains participating in the interface were selected manually according to the above-cited sources. After removing nonprotein atoms (solvent, ligands, ions), the buried interface area for the selected chains was computed as previously described in eq 1. The full set of 200 biological interfaces (PDB identifier, protein names, chains, interface area, selection mode, resolution, protein classification) is given Table S2.

The above-described 400 interfaces (crystallographic, biological) were randomly split into two sets (75% in the training set; 25% in the test set), maintaining an equal proportion of crystallographic and biological interfaces in each subset. Caution was also given to ascertain an equivalent distribution of interface area sizes in both sets. Following the above-described procedure, different random splits (75/25) do not influence the obtained results (best RF parameters, F-measure of the best RF models on the validation and external test sets; data not shown). Training or test set membership is given in Tables S1 and S2.

IPAC,[7] Ponstigl,[10] and Bahadur[21] datasets were retrieved from the PDB according to PDB identifiers and chain names described in the literature.

Atomic Coordinates. For each input PDB file, hydrogen atoms were added with Protoss,[30] a recently described method for the placement of hydrogen coordinates in protein−ligand complexes that takes tautomers and protonation states of both protein chains into account. The method generates the most probable hydrogen positions on the basis of an optimal hydrogen-bonding network using an empirical scoring function.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

Full atomic coordinates of FDS dataset entries can be downloaded at http://bioinfo-pharma.u-strasbg.fr/IChemPIC.

Protein−Protein Interface Descriptors. Interfaces between protein chains are detected following a three-step procedure as follows. First, the interface is broadly defined by counting pairwise distances between all atoms of different chains and keeping only patches for which at least 20 interatomic distances are shorter than 5 Å. In a second step, all intermolecular interactions (hydrophobic, aromatic, hydrogen bond, ionic bond) between the two selected chains are precisely defined using standard parameters of the in-house developed IChem toolkit.[31] The set of topological rules, used to define interactions based on atom pair-dependent distances and angles, is explicitly described in the previous report describing the IChem toolkit.[31] In the third step, an interaction pseudoatom (IPA) is placed at the mid-distance of each atom pair in an interaction according to IChem. Please note that hydrophobic IPAs are clustered if they are less than 1.0 Å apart.[31] If the total number of IPAs is higher than or equal to 5, then the interface is saved; otherwise, it is discarded. Last, a vector of 45 real numbers is generated for each remaining interface describing its size, chemical complementarity, and buriedness (Table S3). The final vector has the following form:

- the total number of IPAs (one parameter);
- the percentage of each of the four interaction types (four parameters);
- the distribution (in counts), for each interaction type, of the buriedness of the corresponding IPAs, binned in 10 intervals from 25 to 100% burial (4 × 10 = 40 parameters). Buriedness of each IPA was inferred as previously reported[32] by computing the proportion of 120 regularly spaced 8 Å long rays having their origin at the IPA 3D coordinates and intersecting the surrounding protein surface.

Altogether, the complete process comprising hydrogen atom addition, interaction detection, and descriptor generation is fast enough (a few seconds per PDB entry) to be applied to the entire PDB.

Random Forest Model. Random Forest (RF) models were built using the RandomForest 4.6-7 library[33] in the R statistical package.[34] A total of 500 decision trees (ntree parameter) were trained on all descriptors of the training set (n = 300), but the number of variables (mtry parameter) at each splitting node was varied. A 5-fold cross-validation procedure was used to randomly split the training set five times into an internal training (four-fifths of the dataset) and an internal test set (one-fifth of the dataset) and analyze the predictive power of RF models on the nonoverlapping five internal test sets. For each mtry value (integer between 2 and 10), the corresponding cross-validated model was assessed according to the following statistical parameters

$$\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{F-measure} = 2 \times (\text{sensitivity} \times \text{precision})/(\text{sensitivity} + \text{precision})$$

where TP are true positives (biological interfaces predicted biological), FP are false positives (crystallographic interfaces predicted biological), TN are true negatives (crystallographic interfaces predicted crystallographic), and FN are false negatives (biological interfaces predicted crystallographic)

The best mtry value was used (i) to derive 10 RF models from the full training set (300 complexes) by varying the random seed number and (ii) these 10 RF models were applied to predict the nature of interfaces in the 100 entries of the external test set.

Comparison to Other Methods. IChemPIC predictions were done using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) and rely on the majority of the 10 independent RF predictions (biological or crystallographic) to annotate an input protein−protein interface. In cases where there are an equal number of predictions for both types, the interface is predicted to be crystallographic. IChemPIC was compared to four state-of-the-art methods (NOXClass,[13] PISA,[9] DiMoVo,[12] and EPPIC)[20] on three external test sets. For each of these tools, standard parameters available at their online version were chosen, giving as input either the PDB code and chain letters (biological interfaces) or the above-prepared atomic coordinates for the two chains (crystallographic interfaces). For the NOXClass multistage SVM classification (http://noxclass.bioinf.mpi-inf.mpg.de/), the pairwise class probabilities (biological, crystallographic) were retained for each pair of protein chains, using three descriptors (interface area, interface area ratio, area-based amino acid composition). In PISA (http://www.ebi.ac.uk/pdbe/pisa/), the interface was defined as biological if the corresponding interface was predicted to be stable among all proposed assemblies. Otherwise, the interface was predicted to be crystallographic. Using the DiMoVo prediction method (http://albios.saclay.inria.fr/dimovo), a score above 0.50 was used to assign a potential biological function to an interface. Last, EPPIC predictions (Bio or Xtal) were done on a web server (http://www.eppic-web.org/ewui/) and based on the consensus voting scheme (final score) based on four descriptors (core sizes, geometry, core-rim, core-surface) of the input interface.

RESULTS AND DISCUSSION

Setting Up the FDS Dataset of Ligandable Protein−Protein Interfaces. None of the existing benchmark datasets is suitable for the purpose of discriminating crystallographic from biologically relevant protein−protein interfaces. On one hand, historical datasets[10,13,21,22] are biased by having a majority of crystallographic entries of much lower interface areas (500−1000 Å^2) than those of true biologically relevant entries (1000−3000 Å^2). On the other hand, the DC dataset[20] corrected this discrepancy by selecting entries with an homogeneous distribution of interface areas (1000−1500 Å^2) that, however, still falls outside the applicability domain of many biologically important PPIs (e.g., p53−mdm2 interface of 780 Å^2; PDB ID: 1YCR) modulated by low molecular weight inhibitors.[35] Both the Bahadur[21] and DC datasets, which do not overlap much with respect to the interface area range, were therefore merged to enlarge the applicability domain of our next predictions. Lastly, we manually collected an additional set of 115 biologically relevant PPIs to yield a final number of 400 interfaces, which was split into training (75% of data) and test (25% of data) sets. Inspecting the respective distribution of interface area sizes in both sets reveals no major bias, although biological interfaces remains, on average, slightly larger than

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

crystallographic ones (Figure 1). We will demonstrate later that the size of the interface has no major influence in discriminating crystallographic from biological interfaces.

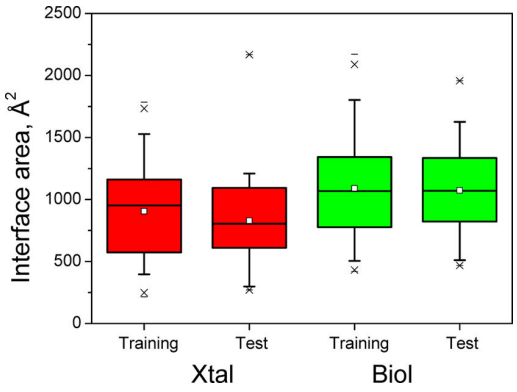


Figure 1. Distribution of interface areas in the designed FDS training and test sets (Xtal, crystallographic interface; Biol, biological interface). Boxes delimit the 25th and 75th percentiles; whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

About 80% of crystallographic complexes (both in the training and test sets) relate to enzymes with well-defined catalytic sites, with the rest being dominated by protein transporters. The proportion of enzymes in the biologically relevant set is lower (about 55% in both the training and test sets), with the biologically relevant set exhibiting more examples of recognition complexes involved in important biological processes (e.g., immune recognition, cell signaling, cell adhesion, transcription).

The average resolution of crystallographic and biologically relevant complexes was 1.84 ± 0.35 and 2.10 ± 0.58 Å, respectively. A large majority of structures (ca. 75%) in both sets was solved at high resolution (<2.5 Å). In the protein preparation step, we ascertained that all side chains participating in the interface were fully described. None of the herein described 400 PDB complexes present an incomplete side chain at the selected protein−protein interface. Along the same line, we checked that no small ions could stabilize the interface of the herein described complexes. Handling water molecules at the interface is tricky since water molecules are absent in 184 of the 400 complexes. We therefore decided to remove all water molecules, whatever their location, resulting in a unique preparation protocol and a fair comparison to other methods that do not take bound waters into consideration. Coming back to the 400 raw PDB files, it appears that bound waters are not frequently present at protein−protein interfaces. When this is the case (30% of 216 entries with explicit water molecules), bound waters typically engage in no more than a single hydrogen bond. We therefore do not believe that the decision to remove bound waters really affects the accuracy of our classifier.

PPI Detection and Descriptor Generation. We first detect the interface between two protein chains and then explicitly compute all nonbonded interactions (hydrophobic contacts, aromatic interactions, hydrogen, and ionic bonds) and generate a pseudoatom (IPA) for describing each interaction (Figure 2). A complex molecular assembly of thousands of

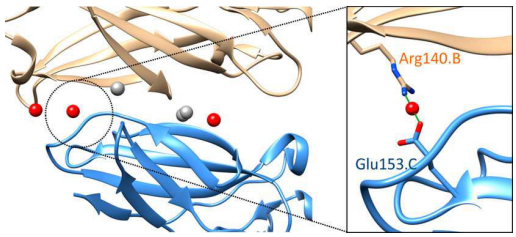


Figure 2. Interface (PDB ID: 4NNY) between interleukin-7 receptor subunit alpha (tan, chain C) and cytokine receptor-like factor 2 (blue, chain C). Six interaction pseudoatoms (spheres) are placed at the mid-distance of each pair of atoms in the interaction and assigned a property corresponding to the interaction (hydrophobic, aromatic, hydrogen bond, ionic bond). The right inset is a zoomed view of a single ionic bond that explicitly displays the interacting side chains.

atoms can be represented, therefore, by a much simpler set of a few IPAs (60−70 on average) describing both the nature and buriedness of the corresponding interaction. Since we explicitly consider hydrogen bonds, it is worth noting that all hydrogen atoms are added to the raw PDB files while optimizing the tautomeric and protonation states protein's amino acids.[30]

Although biological interfaces exhibit, on average, many more IPAs (86 ± 40) than crystal packing contacts (50 ± 30), the average percentages of interaction types are rather similar in both sets (Table 1). As expected, hydrophobic contacts are

Table 1. Average Percentage of Interaction Types at Crystallographic and Biological Interfaces

	protein−protein interface[a]	
interaction type	crystallographic	biological
hydrophobic	78.06 ± 15.70	83.32 ± 9.71
aromatic	0.24 ± 1.14	0.10 ± 0.32
hydrogen bond	17.97 ± 12.11	13.51 ± 7.24
ionic bond	3.65 ± 5.80	3.00 ± 3.87

[a]Statistics from 27 186 protein−protein interactions (200 crystallographic and 200 biological interfaces from the FDS set) detected by IChem.[31]

dominant and represent about 80% of all interactions, although they are more populated in biological interfaces (Table 1). Aromatic interactions (edge-to-face and face-to-face) are quite rare, but they are more populated in the crystallographic set, thereby confirming previous observations.[21] Hydrogen bonds are more frequently found in crystallographic interfaces than in biological assemblies. However, the quality of these hydrogen bonds (e.g., strength, accessibility) is not taken into account in the current analysis. Lastly, the frequency of ionic bonds is rather constant for the two interface categories (3%). Since metal chelation is rarely found at protein−protein interfaces, this interaction type was discarded from the descriptor set to define the shortest possible input vector for RF modeling.

Random Forest Binary Classification Model. Random Forest (RF) is a highly versatile ensemble machine learning method for classification and regression that relies on many independent decision trees.[36] Each tree is created by bootstrap samples of the original training data using a randomly selected subset of features. Then individual trees are combined through a voting process to provide an unbiased prediction. In contrast with single-decision trees, random forests have a low variance and very few biases. Considering that random forests have few

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

parameters to tune (number of trees, number of variables at each split), the method is easy to use in order to produce a reasonable model fast and efficiently. Among its many potential applications, RF is increasingly used in life sciences as either a classifier or nonlinear regression method.[37]

In our application, the number of trees (ntree) was fixed to 500. Besides it having a clear influence on the overall computing time, variations of this parameter did not influence the herein presented results. The number of variables randomly sampled as candidates at each split (mtry parameter) was systematically varied from 2 to 10, and each model was repeated five times by varying a random seed number. Using a mtry value equal to 4, Random Forest modeling leads to a stable and robust 5-fold cross-validated model (F-measure = 0.776 ± 0.09) when applied to the FDS training set (Table 2).

Table 2. Statistics of the Best RF Model on the FDS Dataset

parameter	training set (n = 300)[a]	external set (n = 100)[b]
sensitivity	0.794 ± 0.017	0.728 ± 0.014
precision	0.759 ± 0.010	0.745 ± 0.018
specificity	0.747 ± 0.014	0.750 ± 0.025
accuracy	0.771 ± 0.009	0.739 ± 0.012
F-measure	0.776 ± 0.009	0.736 ± 0.010

[a]Mean and standard deviation of the best 5-fold cross-validated model (ntree = 500, mtry = 4), repeated five times using different random seed numbers. [b]Mean and standard deviation of the best RF model (ntree = 500, mtry = 4) at predicting the nature of 100 protein−protein interfaces, repeated 10 times using different random seed numbers.

The model is equally good at predicting either biological contacts (sensitivity) or crystallographic interfaces (specificity). When applied to the FDS external set of 100 PPIs, a moderate drop in accuracy (0.739 ± 0.012) and F-measure (0.736 ± 0.010) is observed, but the model is still robust at predicting the two categories of PPIs equally well (sensitivity = 0.728 ± 0.014; specificity = 0.750 ± 0.025; Table 2).

To be sure that observed data are neither the result of overtraining nor chance correlation, we first performed a y-scrambling test by randomly assigning the dependent variable (crystallographic or biological) to each of the 400 protein−protein interfaces of the FDS training and test sets. As expected, the F-measure of the corresponding RF models (same parameters as above) significantly dropped to mean values of 0.515 and 0.502 when applied to the training and external test sets, respectively. We next computed 45 RF models (ten runs/model) in which the values of the 45 descriptors were iteratively scrambled for each entry of the training set. For all 45 descriptors, the previously computed 300 values of descriptor d_i were just randomly assigned to the 300 objects (training interfaces). Analyzing the variations in the mean F-measure for the training set permits the identification of the most important parameters among our 45 descriptors (Figure 3).

Out of the 45 descriptors, 11 have a real contribution to the global model (>1% decrease in F-measure) when their respective values are scrambled. The most important parameters are clearly the number of interaction pseudoatoms (nPTS) and the percentage of fully buried hydrophobic contacts (hydro7−hydro10 descriptors; Table S3).

Permutating the values taken by the total number of IPAs (nPTS) decreases the overall F-measure of the model by 1.6%

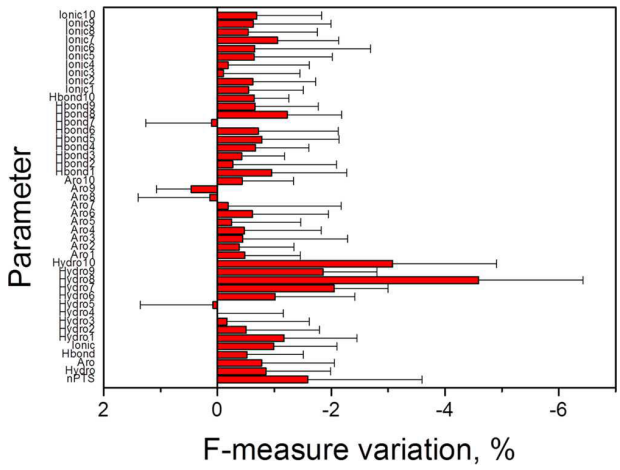


Figure 3. Influence of the permutation of descriptor values on the mean F-measure of 10 RF models obtained with the best cross-validated parameters (ntree = 500, mtry = 4) and trained on the FDS training set.

(Figure 3). Although accessible hydrophobic contacts (hydro1−hydro6 parameters) do not really contribute to the overall F-measure, the more buried hydrophobic interactions (hydro7−hydro10 parameters) are truly critical. Remarkably, permutating the value of the hydro10 parameter (percentage of 100% buried hydrophobic contacts) decreases the F-measure of the RF model by almost 3% (Figure 3). Accordingly, hydrophobic core interface residues, defined as buried by at least 95%, have recently been described as key determinants of biological interfaces.[20] Of less importance, but still significant, is the percentage of other interactions (hydrogen bonds, ionic bonds) and their buriedness, which tends to be higher in biological interfaces than in crystal packing contacts (Figure 3). Scrambling the values of four out of the 45 parameters (hydro5, Aro8, Aro9, Hbond7) leads to slightly better RF models. The largest observed decrease in F-measure (scrambling values of hydro8 parameter) is only by 5% and is probably explained by compensatory effects upon removal of the most critical descriptor. To demonstrate this assumption, we suppressed the hydro8 descriptor from the vector, recomputed a RF model on the set of $n - 1$ descriptors (F-measure of 0.705 on the training set), and iteratively scrambled again the $n - 1$ descriptor values. This time, the most critical descriptor is hydro10 (the former second most important descriptor starting from the full set of descriptors), with a much higher mean decrease in the F-measure (11.3 ± 3.3%). This observation perfectly illustrates our hypothesis and the compensatory effect of the hydro10 parameter upon removing the influence of the hydro8 descriptor.

The weaker contribution of the hydro9 parameter (count of hydrophobic IPAs buried is between 91.6 and 100%) with respect to that of the related hydro8 (count of hydrophobic IPAs buried is between 83.3 and 91.6%) and hydro10 (count of 100% buried hydrophobic IPAs) parameters is intriguing and explained by a peculiar distribution of these parameter values when comparing crystallographic and biologically relevant interfaces (Figure 4). Hence, the distributions of hydro8 and hydro10 counts are clearly different when examining the two subsets of interfaces (shift to higher hydro8 values in crystallographic contacts; shift to higher hydro10 values in biological interfaces). Intriguingly, the hydro9 parameter values

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

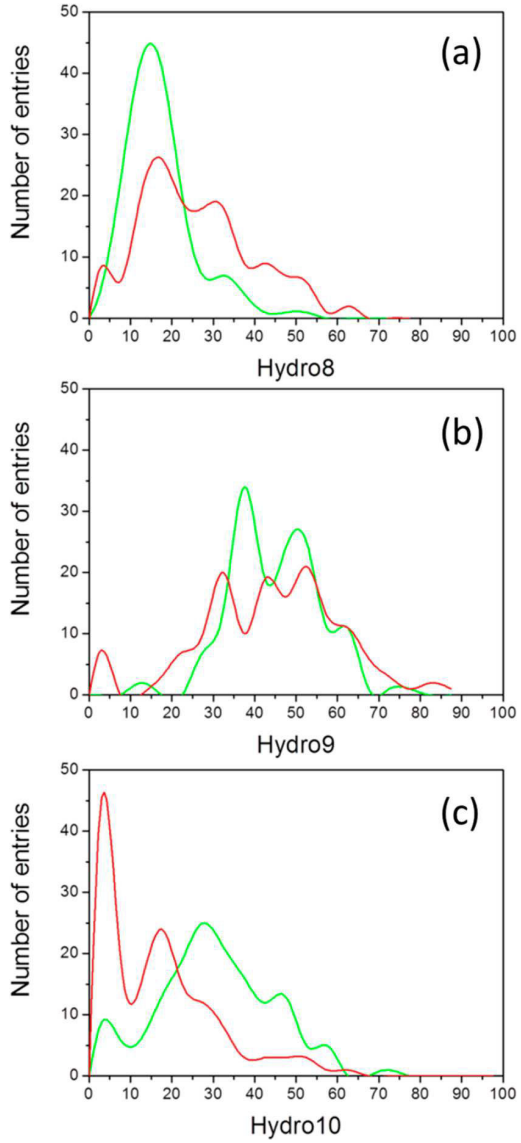


Figure 4. Distribution of the hydro8–hydro10 parameter counts across the FDS training set (green: biological interfaces; red: crystallographic interfaces).

are similarly distributed (Figure 4), thereby explaining why it contributes less to the RF cross-validated model.

To confirm the above-suggested importance of some interface parameters (nPTS, hydro7–hydro10), we ranked the 300 training interfaces by decreasing value of each descriptor (45 lists of PDB entries ranked by decreasing counts for the current descriptor under investigation). We next performed a binary classification of the 300 entries (crystallographic, biological) from the ranks obtained in these 45 lists. A perfect descriptor would lead to a classification (ROC AUC = 1) in which all 150 biological interfaces are ranked higher than the first ranked crystallographic interface. Using the ROC classification, we can estimate the relative importance of each descriptor in discriminating the two categories. Any single descriptor-based classification with an AUC higher than 0.7 (Figure 5) indicates that this descriptor is particularly efficient. This analysis confirms the crucial role of two parameters

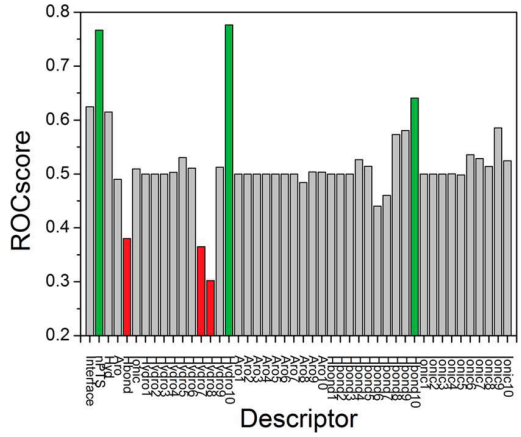


Figure 5. Area under the ROC curve for a binary classification (crystallographic or biological) of 300 interfaces (FDS training set) based on decreasing counts of each of the 45 IChemPIC descriptors.

(nPTS, hydro10) in the discrimination of the two interface subsets. This complementary analysis also shows that counts observed for three descriptors (Hbond, hydro7, and hydro8) are, indeed, higher for crystallographic contacts (ROCscore < 0.50) and therefore also helps to discriminate the two sets of PDB entries. Importantly, using the interface area as a descriptor does not lead to a good binary classification (ROCscore = 0.59), confirming that the FDS training set is really well-balanced with respect to this important criterion, which has been overlooked in the past.

Comparison of IChemPIC to Existing Methods. IChemPIC was compared to four state-of-the-art methods (NOXClass,[13] PISA,[9] DiMoVo,[12] and EPPIC[20]) with regard to their ability to predict the nature of PPIs originating from three different external test sets.

Regarding the diversity of interfaces in the herein presented FDS dataset, it is not surprising that the observed accuracy of existing methods is significantly lower than that reported in the seminal reports describing them.[9,13,12,20] NOXClass is remarkably sensitive (good true positive rate), but this comes at the cost of a much lower specificity (low true negative rate). In contrast, EPPIC and, to a lesser extent, DiMoVo are specific in detecting crystal assemblies (specificity = 0.949), but they perform less well in recognizing biological contacts (low sensitivity), notably when the interface area is small (<750 Å^2; Table S4). PISA, the method currently used in predicting macromolecular assemblies in the RCSB PDB, is the most stable with respect to all statistical parameters taken into account (Table 3).

Altogether, IChemPIC still appears to be the method of choice for a binary classification of protein–protein interfaces since it provides constant and robust performance at predicting both biological and crystal interfaces, whatever their size (see full predictions in Table S4). On one hand, it is inferior to NoxClass and PISA at detecting biological contacts, but it is much more accurate at predicting crystallographic interfaces. On the other hand, IChemPIC is less accurate than EPPIC at predicting crystallographic contacts, but it is significantly better at predicting biologically relevant interfaces (Table 3). Out of the five methods tested here, IChemPIC is notably the only method able to predict, with a similar good accuracy, either small biological interfaces or large crystal contacts.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005–2014

Table 3. Comparison of IChemPIC and State-of-the-Art Methods' Abilities To Predict the Status (Crystallographic, Biological) of the FDS External Test Set (n = 100)

	method				
statistics	IChemPIC[a]	NOXClass	DiMoVo[b]	PISA[c]	EPPIC[d]
sensitivity	0.740	0.878	0.480	0.771	0.667
precision	0.755	0.694	0.857	0.725	0.909
specificity	0.760	0.525	0.733	0.674	0.949
accuracy	0.750	0.719	0.538	0.725	0.826
F-measure	0.747	0.775	0.615	0.747	0.769

[a]Consensus predictions (biological or crystallographic) out of 10 independent RF models. In cases where there is an equal number of predictions for both properties, the interface is predicted to be crystallographic. Predictions were obtained using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC). [b]Thirty five entries common to the DiMoVo and FDS training sets were excluded. [c]Two entries (1i5h, 1y7q) could not be predicted by PISA (crystallographic data is either absent or incomplete); seven additional entries common to the PISA and FDS training sets were excluded. [b]Twenty nine entries common to the EPPIC and FDS training sets were excluded.

Since IChemPIC has been trained on the FDS dataset, it is fair to compare its performance on totally independent external test sets. We therefore chose three additional external datasets (IPAC,[7] Ponstigl,[10] and Bahadur[21]) containing PDB entries not present in the FDS training set. The first two sets have notably been used for benchmarking most tools similar to IChemPic. As stated before,[12,20] the Bahadur and Ponstingl datasets are not very informative because they have a strong bias toward small crystallographic and large biological contacts. As a consequence, all programs including IChemPIC achieve an excellent accuracy (0.85−0.95) at predicting the nature of these entries (Table 4). IChemPic notably exhibits the highest F-measure (0.932 and 0.870, respectively) on these two external sets, which indicates its robustness at predicting biological and crystallographic interfaces equally well (see full predictions in Tables S5 and S6).

The last external set (IPAC validation set 3)[7] is composed of 66 heterodimeric proteins of known crystal structure and experimentally determined binding constants. It notably permits the sensitivity of the method at predicting biological interfaces of quite different strengths to be evaluated. Out of the five methods, NOXClass presents the highest performance (only sensitivity is reported since crystallographic interfaces are lacking in this set) when applied to this dataset (Table 4). Surprisingly, this method never fails when it is applied to the lowest affinity complexes ($K_d < 10^{-5}$ M; Table S7). Given the propensity of NoxClass to overpredict biological interfaces in the previously examined external test sets (sensitivity ≫ precision; Tables 3 and 4), its excellent performance should be considered with extreme caution. Other methods are indeed sensitive to the strength of the corresponding complexes and logically failed more often for low-affinity than for high-affinity complexes (Table S7). Among these methods, IChemPic clearly exhibits the highest accuracy (Table 4).

Practical Application of IChemPIC to PDB Biological Unit Files and Reasons for Its Failure. IChemPIC was next applied to classify 4950 nonredundant interfaces from the Dockground resource of protein−protein X-ray structures.[38] All of these structures are based on the proposed biological unit file (Biounit) inferred from PISA predictions and provided online by the RCSB PDB. About 30% (1493 in total) of these interfaces are, nevertheless, predicted as crystallographic by IChemPIC (Table S8). These discrepancies result from three major causes (Figure 6).

First, our method, like any other, is far from being perfect and fails in ca. 25% of test cases (recall Tables 3 and 4). In many of these cases, the error occurs because IChemPIC predicts interfaces and not quaternary structures. Hence, two chains may form a stable interface depending on the precise context of a much larger oligomeric state. For example, the isolated interface between CTLA-4 (chain B) and B7-2 (chain

Table 4. Performance of IChemPIC with Respect to State-of-the-Art Methods at Predicting the Status (Crystallographic, Biological) of Three Independent Benchmark Sets

	no. of interfaces			method				
test set	crystallographic	biological	statistics	IChemPIC[a]	NOXClass	DiMoVo	PISA	EPPIC
Bahadur[b]	20	122	sensitivity	0.902	0.938	n.a.[c]	0.918	0.885
			precision	0.965	0.892	n.a.	0.875	0.973
			specificity	0.800	0.450	n.a.	0.556	0.850
			accuracy	0.887	0.855	n.a.	0.835	0.880
			F-measure	0.932	0.915	n.a.	0.896	0.927
Ponstingl[d]	67	76	sensitivity	0.882	0.919	0.714	n.a.[e]	0.895
			precision	0.859	0.760	0.714	n.a.	0.840
			specificity	0.831	0.731	0.930	n.a.	0.806
			accuracy	0.858	0.822	0.887	n.a.	0.853
			F-measure	0.870	0.832	0.714	n.a.	0.866
IPACdb[f]	0	66	sensitivity	0.706	0.946	0.394	0.682	0.636

[a]Consensus predictions (biological or crystallographic) out of 10 independent RF models. In cases where there is an equal number of predictions for both properties, the interface is predicted to be crystallographic. Predictions were obtained using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC). [b]One-hundred forty two PDB structures (122 biological, 20 crystallographic) not present in the IChemPIC training set. Entries present in NoxClass (n = 25), DiMoVo (n = 142), and PISA (n = 63) training sets were removed when the corresponding method was used for prediction. [c]Not applicable because DiMoVo was trained on the Bahadur set. [d]One-hundred forty three PDB structures (76 biological, 67 crystallographic) not present in the IChemPIC training set. Entries present in NoxClass (n = 14), DiMoVo (n = 72), and PISA (n = 109) training sets were removed when the corresponding method was used for prediction. [e]Not applicable because PISA was trained on the Ponstingl set. [f]Sixty six PDB heterodimeric structures (validation set 3) of known binding constants. Entries present in IChemPic (n = 15) and NoxClass (n = 10) training sets were removed when the corresponding method was used for prediction.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

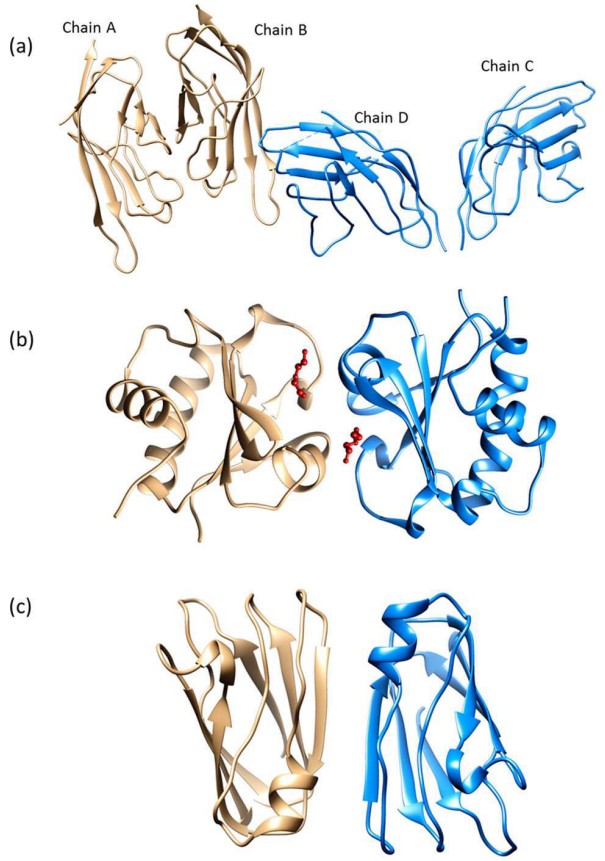


Figure 6. Typical examples of Dockground biological assemblies predicted crystallographic by IChemPIC. (A) CTLA-4 (chains A, B) /B7-2 (chains C, D) complex (PDB ID 1i85); (B) human phosducin-like protein 2 with bound PEG molecules (red ball and sticks) at the interface (PDB ID 3evi); (C) plastocyanin from the cyanobacterium Synechocystis sp. PCC 6803 (PDB ID 1pcs).

D) is predicted to be crystallographic (PDB ID 1I85; interface = 621 Å^2, nPTS = 42) since it exists only within a larger network (Figure 6A), explaining the periodic organization of these molecules within the immunological synapse at the cell surface.[39] Second, many of the small-sized interfaces (149 are smaller than 500 Å^2) are a clear consequence of the crystallization conditions, with either a salt or a precipitant-facilitating molecule at the interface. This case is nicely exemplified by the X-ray structure of human phosducin-like protein 2[40] (PDB ID 3EVI; interface = 422 Å^2, nPTS = 21), which presents two diethylene glycol molecules stabilizing an artifactual homodimeric interface (Figure 6B). Lastly, strong crystal packing may produce artificial interfaces, as shown here by the predicted biological assembly of a cyanobacterium plastocyanin (Figure 6C) with a perfect C_2 symmetry (PDB ID 1PCS; interface = 395 Å^2, nPTS = 6) but no biological relevance.[41]

From the present exercise, we estimate that ca. 15% of PDB biological units have a proposed oligomeric state that is likely to be biologically irrelevant. We therefore strongly suggest the usage of an accurate classifier like IChemPIC to reduce the number of such erroneous biological assemblies and enable the design of PPI inhibitors on biologically relevant interfaces.

CONCLUSIONS

We present a novel computational approach (IChemPIC) to distinguish between biologically relevant and crystallographic protein−protein interfaces. Since none of the existing benchmark datasets are satisfactory at this, notably for predicting small-sized ligandable biological interfaces, novel training and external test sets (FDS set) were defined and manually curated to afford (i) a comparable coverage of interface areas for existing crystallographic and biological interfaces and (ii) an application to small-sized protein−protein interfaces known to be modulated by low molecular weight drug-like compounds.

By describing the interface with a simple vector of 45 real numbers focusing on intermolecular interactions, machine learning methods can be used to classify interfaces as either crystallographic or biological. Due to its simplicity and low parametrization level, the Random Forest machine learning method was chosen to derive a model that distinguishes crystallographic from biological interfaces with a robust accuracy close to 80%. With respect to current state-of-the art methods, IChemPIC is the only approach able to predict with the same good accuracy the two categories of protein−protein interfaces, whatever the external test set. There are many advantages of using IChemPIC with respect to other methods: (i) the implicit inclusion of hydrogen atoms allows for using hydrogen bonds as descriptors for model development; (ii) the method can be applied to interfaces presenting post-translational modifications; (iii) the performance is independent of the size of the interface; and (iv) the applicability domain is large, ranging from small biological protein−protein interfaces (500 Å^2) to larger crystallographic contact (1500 Å^2).

We should acknowledge, however, that IChemPIC is currently parametrized to treat interfaces between two protein chains. For example, the three possible interfaces (AB, BC, AC) of an ABC heterotrimer will be predicted to be either crystallographic or biological, but no prediction will be made for interfaces between one chain and the two others. In other words, no prediction is made for the entire assembly, as in PISA, for example. This drawback explains some of the false negatives observed by IChemPIC and could be easily corrected by enabling the detection of all possible interactions between a single chain and its full protein environment. However, since our method is primarily aimed at further detecting all PDB biologically relevant interfaces amenable to small molecule disruption or stabilization, we prefer to restrict IChemPIC to treat only two-chain interfaces in order to exactly localize the interface to be targeted by a potential PPI modulator. IChemPIC can be used online (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) starting from either a PDB identifier or a user-provided PDB input file.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00190.

Set of 200 crystallographic interfaces (FDS dataset); set of 200 biological interfaces (FDS dataset); descriptors of protein−protein interfaces; prediction of 100 protein−protein interfaces (FDS external set) by various methods; prediction of 142 protein−protein interfaces (Bahadur external set) by various methods; prediction of 143 protein−protein interfaces (Ponstingl external set) by

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

various methods; prediction of 66 protein−protein interfaces (IPAC validation set 3) by various methods; list of Dockground interfaces predicted by PPI-Ichem (PDF).

AUTHOR INFORMATION

Corresponding Author

*Phone: +33 3 68 85 42 35. Fax: +33 3 68 85 43 10. E-mail: rognan@unistra.fr.

Present Address

†(J.D.) Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the National Center for Scientific Research (CNRS, Institut de Chimie) and the Alsace Region for the doctoral fellowship to F.D.S. The High-Performance Computing Center (University of Strasbourg, France) and the Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) are acknowledged for allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg, Germany) for providing an executable version of Protoss. O. Sperandio, X. Morelli, P. Roche, and E. Kellenberger are acknowledged for critical reading of the manuscript and helpful discussions.

REFERENCES

(1) Ivanov, A. A.; Khuri, F. R.; Fu, H. Targeting Protein-Protein Interactions as an Anticancer Strategy. *Trends Pharmacol. Sci.* **2013**, *34*, 393−400.

(2) Villoutreix, B. O.; Kuenemann, M. A.; Poyet, J.-L.; Bruzzoni-Giovanelli, H.; Labbé, C.; Lagorce, D.; Sperandio, O.; Miteva, M. A. Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology. *Mol. Inf.* **2014**, *33*, 414−437.

(3) Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **2007**, *450*, 1001−1009.

(4) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2012**, *33*, 224−232.

(5) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic acids research* **2000**, *28*, 235−242.

(6) Janin, J. Specific Versus Non-Specific Contacts in Protein Crystals. *Nat. Struct. Biol.* **1997**, *4*, 973−974.

(7) Mitra, P.; Pal, D. Combining Bayes Classification and Point Group Symmetry under Boolean Framework for Enhanced Protein Quaternary Structure Inference. *Structure* **2011**, *19*, 304−312.

(8) Henrick, K.; Thornton, J. M. PQS: A Protein Quaternary Structure File Server. *Trends Biochem. Sci.* **1998**, *23*, 358−361.

(9) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.

(10) Ponstingl, H.; Henrick, K.; Thornton, J. M. Discriminating between Homodimeric and Monomeric Proteins in the Crystalline State. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 47−57.

(11) Ponstingl, H.; Kabir, T.; Thornton, J. M. Discriminating Between Homodimeric and Monomeric Proteins in the Cristalline State. *J. Appl. Crystallogr.* **2003**, *36*, 1116−1122.

(12) Bernauer, J.; Bahadur, R. P.; Rodier, F.; Janin, J.; Poupon, A. DiMoVo: A Voronoi Tessellation-Based Method For Discriminating Crystallographic and Biological Protein-Protein Interactions. *Bioinformatics* **2008**, *24*, 652−658.

(13) Zhu, H.; Domingues, F. S.; Sommer, I.; Lengauer, T. Noxclass: Prediction of Protein-Protein Interaction Types. *BMC Bioinf.* **2006**, *7*, 27.

(14) Liu, Q.; Kwoh, C. K.; Li, J. Binding Affinity Prediction for Protein-Ligand Complexes Based on Beta Contacts and B Factor. *J. Chem. Inf. Model.* **2013**, *53*, 3076−3085.

(15) Tsuchiya, Y.; Kinoshita, K.; Naiamura, H. Analyses of Homo-Oligomer Interfaces of Proteins from the Complementarity of Molecular Surface, Electrostatic Potential and Hydrophobicity. *Protein Eng., Des. Sel.* **2006**, *19*, 421−429.

(16) Block, P.; Paern, J.; Hullermeier, E.; Sanschagrin, P.; Sotriffer, C. A.; Klebe, G. Physicochemical Descriptors to Discriminate Protein-Protein Interactions in Permanent and Transient Complexes Selected by Means of Machine Learning Algorithms. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 607−622.

(17) Mintseris, J.; Weng, Z. Atomic Contact Vectors in Protein-Protein Recognition. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 629−639.

(18) Valdar, W. S.; Thornton, J. M. Protein-Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 108−124.

(19) Elcock, A. H.; McCammon, J. A. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 2990−2994.

(20) Duarte, J. M.; Srebniak, A.; Scharer, M. A.; Capitani, G. Protein Interface Classification by Evolutionary Analysis. *BMC Bioinf.* **2012**, *13*, 334.

(21) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A Dissection of Specific and Non-Specific Protein-Protein Interfaces. *J. Mol. Biol.* **2004**, *336*, 943−955.

(22) Chakrabarti, P.; Janin, J. Dissecting Protein-Protein Recognition Sites. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 334−343.

(23) Bourgeas, R.; Basse, M. J.; Morelli, X.; Roche, P. Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2p2i Database. *PLoS One* **2010**, *5*, e9598.

(24) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. A., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, W.; Wolf, R. M.; Wu, X.; Kollman, P. A. *Amber*, version 14; University of California: San Francisco, CA. http://ambermd.org/.

(25) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305−320.

(26) Redundancy in the Protein Data Bank. http://www.rcsb.org/pdb/statistics/clusterStatistics.do.

(27) Rognan, D. Rational Design of Protein-Protein Interaction Inhibitors. *MedChemComm* **2015**, *6*, 51−60.

(28) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. A Structure-Based Benchmark for Protein-Protein Binding Affinity. *Protein Sci.* **2011**, *20*, 482−491.

(29) Gavenonis, J.; Sheneman, B. A.; Siegert, T. R.; Eshelman, M. R.; Kritzer, J. A. Comprehensive Analysis of Loops at Protein-Protein Interfaces for Macrocycle Design. *Nat. Chem. Biol.* **2014**, *10*, 716−722.

(30) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.

(31) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.

(32) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.

(33) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R news* **2002**, *2*, 18−22.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

(34) *R: A language and environment for statistical computing*, version 3.2.0; R Foundation for Statistical Computing: Vienna, Austria. http://www.r-project.org/.
(35) Chang, Y. S.; Graves, B.; Guerlavais, V.; Tovar, C.; Packman, K.; To, K. H.; Olson, K. A.; Kesavan, K.; Gangurde, P.; Mukherjee, A.; Baker, T.; Darlak, K.; Elkin, C.; Filipovic, Z.; Qureshi, F. Z.; Cai, H.; Berry, P.; Feyfant, E.; Shi, X. E.; Horstick, J.; Annis, D. A.; Manning, A. M.; Fotouhi, N.; Nash, H.; Vassilev, L. T.; Sawyer, T. K. Stapled Alpha-Helical Peptide Drug Development: A Potent Dual Inhibitor of Mdm2 and Mdmx for P53-Dependent Cancer Therapy. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E3445−3454.
(36) Breiman, L. Random Forests. *Mach Learn* **2001**, *45*, 5−32.
(37) Touw, W. G.; Bayjanov, J. R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; van Hijum, S. A. F. T. Data Mining In the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle? *Briefings Bioinf.* **2013**, *14*, 315−326.
(38) Douguet, D.; Chen, H. C.; Tovchigrechko, A.; Vakser, I. A. DOCKGROUND Resource for Studying Protein-Protein Interfaces. *Bioinformatics* **2006**, *22*, 2612−2618.
(39) Schwartz, J. C.; Zhang, X.; Fedorov, A. A.; Nathenson, S. G.; Almo, S. C. Structural Basis for Co-Stimulation by The Human Ctla-4/B7−2 Complex. *Nature* **2001**, *410*, 604−608.
(40) Lou, X.; Bao, R.; Zhou, C. Z.; Chen, Y. Structure of the Thioredoxin-Fold Domain of Human Phosducin-Like Protein 2. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2009**, *65*, 67−70.
(41) Romero, A.; De la Cerda, B.; Varela, P. F.; Navarro, J. A.; Hervas, M.; De la Rosa, M. A. The 2.15 Å Crystal Structure of a Triple Mutant Plastocyanin from the Cyanobacterium Synechocystis sp. pcc 6803. *J. Mol. Biol.* **1998**, *275*, 327−336.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

J Comput Aided Mol Des
DOI 10.1007/s10822-016-9930-3

Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015

Inna Slynko[1] · Franck Da Silva[1] · Guillaume Bret[1] · Didier Rognan[1]

Received: 23 April 2016 / Accepted: 25 July 2016

Abstract High affinity ligands for a given target tend to share key molecular interactions with important anchoring amino acids and therefore often present quite conserved interaction patterns. This simple concept was formalized in a topological knowledge-based scoring function (GRIM) for selecting the most appropriate docking poses from previously X-rayed interaction patterns. GRIM first converts protein–ligand atomic coordinates (docking poses) into a simple 3D graph describing the corresponding interaction pattern. In a second step, proposed graphs are compared to that found from template structures in the Protein Data Bank. Last, all docking poses are rescored according to an empirical score (GRIMscore) accounting for overlap of maximum common subgraphs. Taking the opportunity of the public D3R Grand Challenge 2015, GRIM was used to rescore docking poses for 36 ligands (6 HSP90α inhibitors, 30 MAP4K4 inhibitors) prior to the release of the corresponding protein–ligand X-ray structures. When applied to the HSP90α dataset, for which many protein–ligand X-ray structures are already available, GRIM provided very high quality solutions (mean rmsd = 1.06 Å, n = 6) as top-ranked poses, and significantly outperformed a state-of-the-art scoring function. In the case of MAP4K4 inhibitors, for which preexisting 3D knowledge is scarce and chemical diversity is much larger, the accuracy of GRIM poses decays (mean rmsd = 3.18 Å, n = 30) although GRIM still outperforms an energy-based scoring function. GRIM rescoring appears to be quite robust with comparison to the other approaches competing for the same challenge (42 submissions for the HSP90 dataset, 27 for the MAP4K4 dataset) as it ranked 3rd and 2nd respectively, for the two investigated datasets. The rescoring method is quite simple to implement, independent on a docking engine, and applicable to any target for which at least one holo X-ray structure is available.

Keywords Docking · D3R · Drug discovery data resource · Grand challenge

Introduction

In absence of structural data on protein ligand complexes (X-ray diffraction, nuclear magnetic resonance spectroscopy, electron microscopy), molecular docking remains the computational method of choice to predict ligand binding modes [1]. Since the pioneering work of Kuntz et al. [2], over 100 different docking software have been reported [1, 3–6] that progressively addressed most of the issues related to this computational exercise: full ligand flexibility, accurate configurational sampling of the ligand in the protein binding site, partial protein flexibility, implicit or explicit solvation, prediction of relative or absolute binding (free) energies. Many benchmarking studies [7–11] comparing different algorithms across diverse datasets of protein–ligand X-ray structures, agree on the point that state-of-the-art docking algorithms are very efficient in predicting ligand poses: a relative position of a ligand with respect to a protein and a conformation of a protein-bound ligand. Unfortunately, these good solutions are hardly distinguishable from a much larger set of

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9930-3) contains supplementary material, which is available to authorized users.

✉ Didier Rognan
rognan@unistra.fr

[1] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France

Published online: 01 August 2016

incorrect proposals (decoys) using any predicted energy criterion. Considering success in pose prediction as the ability to predict poses with root-mean square deviation (rmsd) from X-ray solution below 2 Å, most docking tools present in the best cases (self-docking) a success rate close to 70 % when considering the top-1 (best scored) solution [12]. Considering all possible solutions, this rate raises typically up to 85–90 % [12] thereby demonstrating that the best scored pose is not always the most reliable one. If handling multiple docking solutions is feasible albeit cumbersome for one particular ligand, this approach cannot be followed upon in silico screening a large compound library. Reasons accounting for repetitive failures in predicting either binding free energies or relative potency ranks [9, 12–14] are numerous: target flexibility, incorrect protonation/tautomeric states, incorrect treatment of many energy terms (solvation, entropy, metal chelation and weak non-covalent interactions).

Three main approaches have been followed to rescue the inability of fast scoring functions to prioritize the best docking poses: (1) develop more sophisticated first-principle scoring functions, (2) use supervised machine learning (ML) algorithms to predict the likelihood of docking poses, (3) apply knowledge-based (chemical and topological) rules to filter out unreliable solutions. The first approach uses CPU-intensive energy calculations (e.g. MM-PBSA, MM-GBSA) to refine early docking results. Unfortunately, the benefit of this extra computational cost is controversial as it appears to be target-dependent and hardly predictable [15–17]. The second approach consists in training machine learning algorithms (e.g. support vector machines [18], random forest [19, 20]) with 3D protein–ligand structural descriptors in order to discriminate good from bad poses. If remarkable results in predicting binding affinities from protein–ligand X-ray structures have been recently published [20], such scoring functions have rarely been applied to prospective virtual screening campaigns and their true utility in virtual screening remains unknown. In any case, docking/ML combinations [21] must be regarded with great care due to the tendency of machine learning methods to be overtrained [22]. The third strategy, which is currently experiencing a revival, utilizes various knowledge-based approaches to rescore docking poses. The main idea is to use non-energetical topological criteria to address the quality of docking poses, notably by comparing docking solutions with protein–ligand complexes of known X-ray structure. Among knowledge-based approaches, we can clearly distinguish those methods aimed at constraining the docking algorithms towards expected poses (pharmacophore-constrained docking [23], shape-guided docking [24, 25], template matching [26]) from computational protocols that just restrain the analysis of docking poses to reward user-defined features. Both methods have proven useful in many examples for enhancing the quality of top-ranked poses as well as enriching virtual hit lists in true actives. Constrained docking may however be dangerous in forcing known inactive compounds to properly dock in a binding site. It is therefore common practice to conduct a totally free docking calculation and further apply simple cheminformatics descriptors (1D fingerprints [27], 3D similarity [28]) to enable the selection of docking solutions that look the most similar to experimentally-determined poses of known ligands. For example, we [29] and others [27, 30, 31] proposed several years ago, the concept of molecular interaction fingerprints (IFPs) [29] to post-process docking data and pick-up poses producing IFPs similar to that of known actives. Computing IFPs from docking poses is a robust and very efficient manner to predict ligand binding modes [32], propose reliable scaffold hops [33], and enrich virtual hits in true actives upon docking a compound library [34, 35]. The success of this post-processing approach relies on the fact that true ligands of a same target often share key interactions with key anchoring residues and thereby produce quite similar IFPs. A drawback of this method lies in the definition of a consensus binding site (fixed set of target residues) in order to generate fixed-sized and comparable interaction fingerprints. To overcome this limitation and extend the concept of interaction fingerprints to binding site-independent and coordinates frame-invariant fingerprints, we recently proposed to encode interaction patterns (sets of protein–ligand interactions) into either generic 1D fingerprints or 3D graphs [36]. Our GRIM algorithm for matching interaction pattern graphs has been described in details elsewhere [36], and here it will be just briefly summarized. Starting from 3D coordinates of a protein–ligand complex, molecular interactions (hydrophobic contacts, aromatic interactions, hydrogen bonds, ionic bonds, metal chelation) are first computed from a set of topological and chemical rules. Every detected interaction is then labelled by three interaction pseudoatoms (IPAs) located on (1) the ligand-interacting atom, (2) the protein-interacting atom and (3) the geometric barycenter of protein and ligand-interacting atoms (Fig. 1a, b). Starting from two sets of IPAs (reference, target), GRIM first creates a list of possible IPA matches. A pair is made if reference and target IPAs have the same label (same interaction type) and represent the same point of view (ligand, protein, barycenter). A product graph is created from the two reference and target graphs in which each successfully matched pair is consequently a vertex. A weight is added to each pair which is inversely proportional to the observed frequency among the 284,186 IPAs generated from the 9877 protein–ligand complexes of the sc-PDB dataset [37]. Assigned weights were as follows: hydrophobic IPA (0.299), aromatic IPA (0.990), h-bond

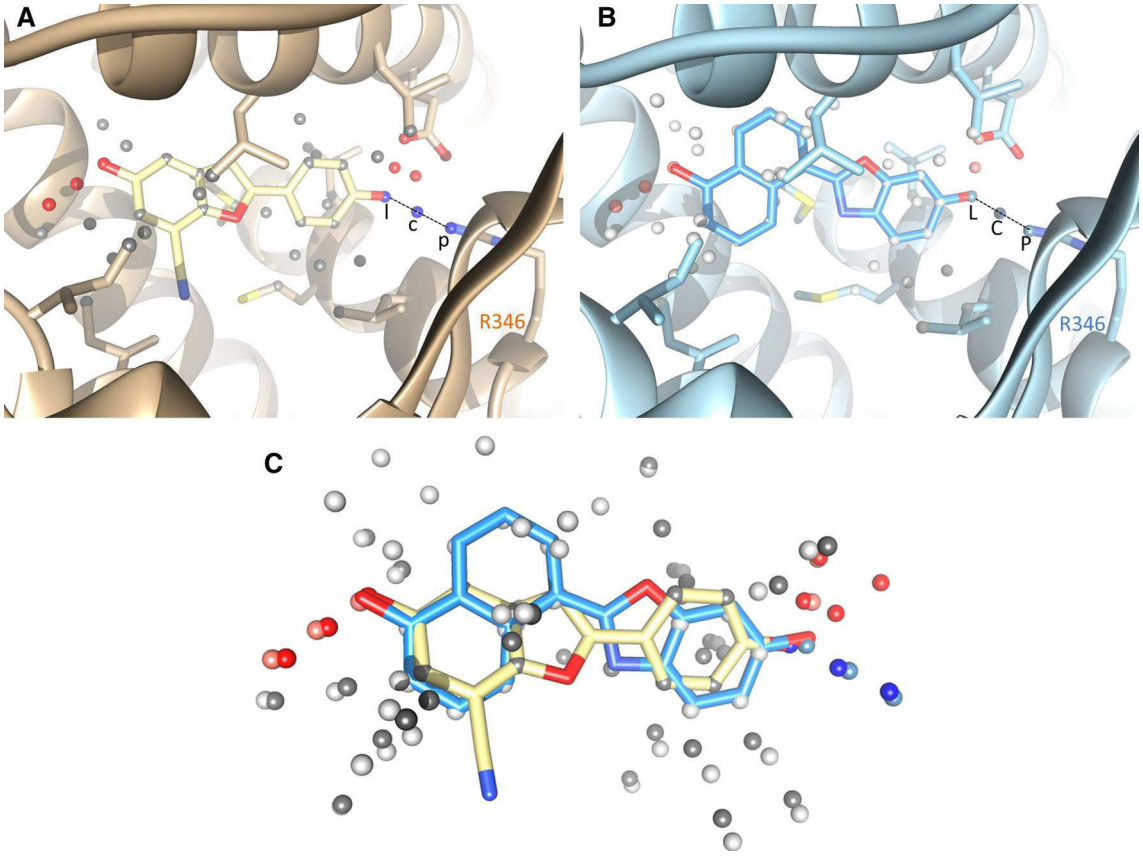


Fig. 1 Principle of the graph alignment method for matching interaction patterns (GRIM). **a** The estrogen receptor β-WAY697 complex (PDB ID 1 × 76) is converted into a set of interaction pseudoatoms (IPAs) describing intermolecular interactions. For each interaction (e.g. hydrogen-bond between Arg346 of the protein and a phenolic oxygen atom of the ligand, displayed by a *dashed black line*), IPAs are placed on the ligand-interacting atom (l), the protein-interacting atom (p) and the geometric barycenter of both interacting atoms (c). IPAs are color-coded according the described interaction (*gray*, apolar and aromatic interactions; *red*, hydrogen bond (protein acceptor) and ionic bond (protein negatively charged); *blue*, hydrogen bond (protein donor) and ionic bond (protein positively charged). **b** Same procedure as above for a second complex between the same receptor and WAY-338. IPAs describing the same interaction with Arg346 are labelled L, C and P, respectively. IPAs are color-coded according the described interaction (*light gray*, apolar and aromatic interactions; *light red*, hydrogen bond (protein acceptor) and ionic bond (protein negatively charged), *light blue*, hydrogen bond (protein donor) and ionic bond (protein positively charged). **c** Graph-based alignment of the two sets of IPAs leading to an interaction-guided overlay of the two bound ligands. Please note that GRIM does not allow matching IPAs from different origin (e.g. l-type with p-types for example)

acceptor (0.930), h-bond donor (0.834), negative ionizable (0.993), positive ionizable (0.966), metal complexation (0.985). An edge is observed between two vertices of the product graph after computing distances between the two reference IPAs and the two target IPAs. If the difference is below a given threshold [37], an edge is created. The largest cliques are then detected using the Bron–Kerbosch algorithm [38] with pivoting and pruning improvements [39]. Each IPA of the target is matched with the corresponding reference IPA using a quaternion-based characteristic polynomial [40]. It returns both the translation vector and the rotation matrices to match target and reference graphs as well as a Graph-alignment score (GRIMscore). As the graph is specific of a given protein–ligand interaction pattern, two sets of protein–ligand coordinates can therefore be easily compared by aligning the corresponding graphs (Fig. 1c). According to a previous benchmark, a GRIMscore value above 0.70 is indicative of a statically significant similarity of interaction patterns.

When applied to rescore docking poses generated by Surflex-Dock [41], GRIM rescoring significantly outperformed the Surflex-Dock scoring function in a retrospective virtual screening exercise aiming at recovering true actives from DUD-E decoys [42] for 10 targets of pharmaceutical interest [36]. We herewith present a prospective application of the GRIM graph matching method to the problem of docking pose selection by predicting, prior to the release of the corresponding X-ray coordinates (D3R Grand Challenge 2015) [43], the binding modes of 36 inhibitors bound

to two different targets of pharmaceutical interest (HSP90α, MAP4K4). This manuscript will only highlight results obtained for Stage 1 (docking pose accuracy) of the D3R challenge.

Computational methods

HSP90α dataset

Four input protein structures (PDB ID: 2JJC, 2XDX, 4YKY, 4YKR) and 180 known HSP90α inhibitors (SMILES strings, Supplementary Table 1) were directly downloaded from the D3R Grand Challenge 2015 website [43] as a zipped archive file (279_data_303589.tar.gz).

HSP90α protein structures were prepared for docking as follows. First, existing hydrogen atoms were removed from the 4 input HSP90α structures and added back while optimizing both the protonation and tautomeric states using Protoss v.2.0 [44]. Two conserved waters mediating the interactions between protein and ligands (HOH2078 and HOH2166 for 2JJC; HOH2029 and HOH2054 for 2XDX; HOH6 and HOH233 for 4YKR; HOH5 and HOH198 for 4YKY) were kept, other water molecules were removed. Next, each protein structure was saved in 4 copies varying by the presence/absence of bound waters as follows: with both water molecules (wat2), with the first water molecule (wat1a), with the second water molecule (wat1b), without waters (dry). In total, 16 structures were therefore used as input for docking the 180 HSP90α inhibitors, which were provided by D3R Grand Challenge. Note, that binding mode should have been predicted only for 6 of those compounds.

Ligands from HSP90α crystal structures were checked manually (bond order, protonation and tautomeric states) and corrected whenever necessary. Protein and ligand structures were separately saved in MOL2 format using SYBYLX-2.1 [45]. In addition, 176 HSP90α-inhibitor complexes (Supplementary Table 2) were defined as templates for graph matching by searching the RCSB Protein Data Bank [46] for the P07900 UniProt [47] accession number and a known bound ligand. These 176 complexes were further processed as described above.

Starting from the provided input SMILES strings of 180 HSP90α inhibitors, hydrogen atoms were added and a 3D conformation was generated for every ligand using Corina v3.40 [48]. All Ligands were then saved in MOL2 format.

MAP4K4 dataset

Two input protein structures (PDB ID: 4OBO, 4U44) and 30 known MAP4K4 inhibitors (SMILES strings, Supplementary Table 3) were directly downloaded from the D3R Grand Challenge 2015 website [43] as a zipped archive file (280_data_473989.tar.gz). Furthermore, 6 additional MAP4K4-inhibitor complexes (Supplementary Table 4) were retrieved by searching the RCSB Protein Data Bank for the O95819 UniProt accession number and a known bound ligand. The 8 protein structures were prepared for docking using the protocol described for the HSP90α dataset. No bound water molecules were conserved in the present case.

Starting from the input SMILES strings of the 30 MAP4K4 inhibitors, hydrogen atoms were added and 3D conformations were generated using Corina v3.40 [48]. All Ligands were then saved in MOL2 format.

Docking

Ligands were docked to input protein structures using Surflex-Dock v.2745 [41]. For each protein input structure, a *protomol* was first generated using a list of binding site residues (including bound waters) for which at least one heavy atom was closer than 6.5 Å from at least one ligand heavy atom. The docking accuracy parameter set *-pgeom* was used. The -pgeom option starts each docking from 4 initial and different poses to ensure good search coverage, turns on ligand minimization prior to docking and after docking (in-pocket minimization), ensures that the returned poses are different from one another by at least 0.5 Å rmsd, and saves a total of 20 poses (ranked by Surflex-Dock energy score, from 000 to 019). In summary a total of 57,600 ($180 \times 16 \times 20$) and 4800 ($30 \times 8 \times 20$) poses were generated for the HSP90α, and MAP4K4 inhibitors respectively.

GRIM rescoring

Each Surflex-Dock pose was compared to the list of template complexes (176 for HSP90α inhibitors, 8 for MAP4K4 inhibitors). The interaction pattern of each docking pose was computed with IChem [36], aligned to that of the corresponding templates by graph-based alignment and ranked by GRIMscore. For every HSP90α ligand to dock, all poses were merged, regardless of the input protein structure and graph template used, and ranked by decreasing GRIMscore. For each MAP4K4 inhibitor, poses that do not exhibit at least one hydrogen-bond to the hinge region of the MAP4K4 kinase (residues E106, M107, C108) were discarded from further evaluation. Such poses were detected thanks to the protein–ligand interaction fingerprint generator [29] embedded in the IChem toolkit [36]. The 5 remaining poses with the highest GRIMscores (GRIM-1 to GRIM-5) were saved for every MAP4K4 ligand. For HSP90α inhibitors, a slightly different protocol was used to reflect the much higher number of templates and GRIM comparisons. To avoid retrieving too similar

solutions, all poses were then clustered using an agglomerative method and a complete linkage clustering, starting from the highest GRIMscore (seed) and using a 2 Å rmsd threshold from the seed pose, until five different clusters were defined for each ligand. A representative pose (highest GRIMscore) for each of the 5 clusters was finally saved and ranked from 1 (GRIM-1) to 5 (GRIM-5) by decreasing GRIMscore.

Results and discussion

Predicting the binding mode of 6 HSP90α inhibitors

The first part of the challenge consisted in predicting the bound conformation of 180 HSP90α inhibitors from three chemical series (benzimidazolones, aminopyridines, benzophenones), given four reference input protein structures co-crystallized with at least one inhibitor of the above-cited three chemical series. A particular emphasis was put on six inhibitors (Fig. 2) whose protein-bound X-ray structures had to be released just at the closure of the first step (pose prediction accuracy) of the D3R Grand Challenge 2015. Since HSP90α inhibitors notoriously use conserved water molecules [49] to recognize the ATP-binding site, we decided to generate four sets of protein coordinates for each of the provided 4 input structures that just differ in the number of bound waters (none, one or two; see "Computational methods"). To use knowledge about inhibitor binding to the HSP90α target, we further retrieved 176 additional protein–ligand X-ray structures from the Protein Data Bank and ensured that all these inhibitors were occupying the same binding site that the 4 ligands co-crystallized with the input reference structures. Docking of all inhibitors to the 16 input structures was completely unrestrained (beside defining the common binding-site) and led to a total of 57,600 poses which were all compared to the 176 template structures using our GRIM interaction pattern matching method. To ascertain the generation of a few representative but diverse poses, we decided to cluster docking solutions using a 2 Å-rmsd threshold and provided up to 5 poses for each of 180 ligands (Table 1). Analyzing the rmsd of predicted poses to the true X-ray solution (released just after closure of the challenge) shows that our interaction pattern rescoring strategy achieves an outstanding accuracy since top-1 GRIM poses are predicted with a mean rmsd of 1.06 Å (Table 1). The top-1 GRIM pose of only two compounds (hsp90_44, hsp90_175) is predicted with a rmsd higher than 1 Å. The larger value of 2.47 Å (hsp90_44) is mainly due to pose differences occurring at the accessible pyridine-3-sulfonamide that does not strongly interact with the binding site; the position of the buried benzimidazolone core being nicely predicted with a rmsd of 0.53 Å (Fig. 3a). For compound hsp90_175, the main difference (rmsd = 1.67 Å) lies in the rotation of a single dihedral angle that drifts a phenol ring from its X-ray pose.

Since we intentionally clustered poses to avoid generating too many redundant answers, the quality of GRIM poses logically deteriorates when other solutions are considered (Table 1). Four 4 out of the 6 ligands, the GRIM-1 pose is by far the closest to the true X-ray structure which greatly facilitates the analysis of our rescoring. In all cases, the top solution selected by GRIM is better than that

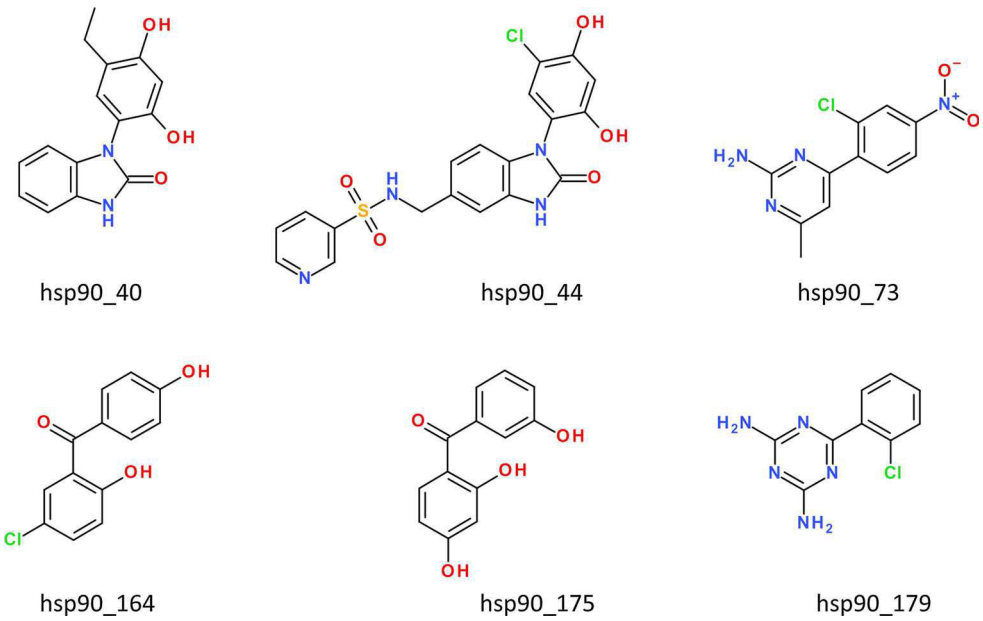


Fig. 2 Structure and name of six HSP90α inhibitors to dock and to determine binding mode

Table 1 Accuracy of pose selection (rmsd in Å to X-ray solution) for six HSP90α inhibitors

Compound	rmsd to X-ray, Å						
	GRIM-1[a]	GRIM-2[b]	GRIM-3[c]	GRIM-4[d]	GRIM-5[e]	Surflex-1[f]	Best[g]
hsp90_40	0.44	1.69	5.54	2.35	6.12	0.59	0.44
hsp90_44	2.48	4.19	2.79	1.49	2.78	4.17	1.36
hsp90_73	0.85	2.49	5.78	5.60	3.31	2.01	0.72
hsp90_164	0.37	5.56	5.87	3.53	5.42	0.85	0.37
hsp90_175	1.67	5.85	1.35	6.00	5.72	1.81	0.73
hsp90_179	0.54	2.12	4.53	3.63	3.44	2.00	0.27
Mean rmsd	1.06	3.65	4.31	3.77	4.47	1.91	0.65

[a] 1st pose according to GRIMscore

[b] 2nd pose according to GRIMscore

[c] 3rd pose according to GRIMscore

[d] 4th pose according to GRIMscore

[e] 5th pose according to GRIMscore

[f] 1st pose according to Surflex-Dock score

[g] Lowest rmsd pose

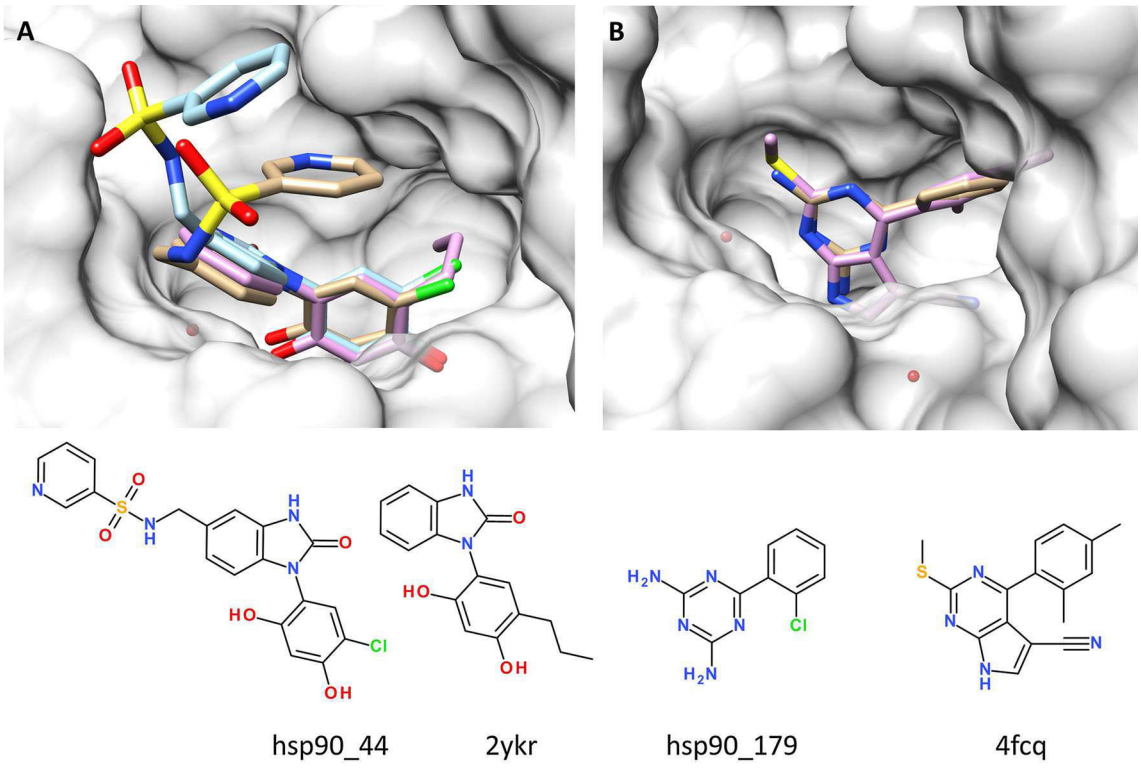


Fig. 3 Predicted versus X-ray pose of two HSP90α inhibitors. Heteroatoms are colored in *blue* (nitrogen), *red* (oxygen), *yellow* (sulfur), and *green* (chlorine). The chemical structures of the two inhibitors are displayed below the binding poses. **a** Predicted binding mode of hsp90_44 (*tan sticks*) to HSP90α ATP-binding site (*white surface*). The pose has been selected by interaction pattern similarity to that of another HSP90α inhibitor co-crystallized with PDB entry 2ykr (*plum sticks*). The true X-ray pose is indicated by cyan sticks. **b** Predicted binding mode of hsp90_179 (*green sticks*) to HSP90α ATP-binding site (*white surface*). The pose has been selected by interaction pattern similarity to that of another HSP90α inhibitor co-crystallized with PDB entry 4fcq (*plum sticks*)

predicted by the native scoring function embedded in Surflex-Dock (mean rmsd = 1.91 Å; Table 1). As observed for almost all docking engines, the top-ranked pose is rarely the absolute best solution (the closest to the true X-ray pose). Among the set of 320 poses generated for each ligand, the lowest-rmsd pose is indeed very close

(mean rmsd = 0.65 Å; Table 1) to the X-ray solution, thereby attesting the quality of Surflex-Dock as pose generator. For two out of the six ligands (hsp90_40, hsp90_164), the absolute best pose is ranked first by GRIM (Table 1). For two others (hsp90_73, hps90_179), the rmsd difference is so tiny (<0.3 Å) that these poses can be considered as almost equivalent.

We next looked for those cases where GRIM rescoring was able to rank at first position a near-native pose, and identified which protein input structures and which interaction pattern template had been used to select this particular pose (Table 2). In two cases (hsp90_40, hs90_175), the ligand to dock is a very close analog of the co-crystallized ligand from the input reference structure (2D similarity >0.90), it is therefore logical that the later protein structures and corresponding interaction patterns are used by GRIM to select the top pose. As a consequence, the interaction pattern of the predicted pose is very similar to that of the template (GRIMscore >0.85). However, the remaining ligands were posed by interaction pattern similarity to that of chemically different template ligands (2D similarity <0.60) thereby nicely illustrating the power of the knowledge-based rescoring method. A prototypical example is given by the 6-phenyl-1,3,5-triazine-2,4-diamine hsp90_179 (Table 2) whose correct pose (rmsd = 0.54 Å) has been deduced from that of a chemically unrelated 7H-pyrrolo[2,3-d]pyrimidine inhibitor (PDB entry 4fcq) that however exhibits a very similar binding mode (Fig. 3b) and interaction pattern (GRIMscore = 0.79).

Predicting the binding mode of 30 MAP4K4 inhibitors

The second challenge aims at predicting the bound conformation of MAP4K4 inhibitors, and is much more demanding than the previous one for many reasons: (1) the dataset to dock is larger (30 inhibitors; Fig. 4 and Supplementary Table 3) and much more chemically diverse (17 different chemotypes, 11 low molecular-weight fragments), (2) the number of templates (known MAP4K4-inhibitor X-ray structures) is lower with only 8 PDB complexes and three chemotypes (amino-quinazolines, amino-pyrrolotriazines, hydroxydihydropyridinone; Supplementary Table 4).

Although Surflex-Dock is able to propose at least one very reliable docking pose for 28 out of the 30 ligands (mean rmsd of the best possible pose = 0.94 Å; Table 3). The native Surflex-Dock scoring function and GRIMscore cannot detect near-native poses (<2 Å rmsd) as top-1 solution for 17 and 19 inhibitors, respectively. Rescoring by interaction pattern graph similarity (GRIM) provides overall better poses (mean rmsd of GRIM-1 pose = 3.18 Å) than Surflex-Dock (mean rmsd of Surflex-1 pose = 3.63 Å) but their accuracy remains lower than that observed for the previous HSP90α dataset. Despite their medium accuracy, it remains reassuring that the quality of the poses decreases with the GRIM rank (Table 3).

We next looked for the reasons explaining why it is so challenging to find near-native poses for MAP4K4 inhibitors. The first reason is that the MAP4K4 set of inhibitors contains a significant amount (11 out of 30 compounds) of low molecular weight fragments (MW < 250 and heavy atoms count <20). Out of these 11 fragments, only three of them (27 %) are well posed by GRIM (Fig. 5). Conversely, the success in predicting near-native poses for higher molecular weight ligands (heavy atom count ≥20) is significantly higher (8 out of 19, ratio = 42 %; Fig. 5).

Upon examining the GRIM docking poses of all 30 MAP4K4 inhibitors, we could identify three possible scenarios. The first one relates to 9 lead-like compounds

Table 2 Characteristics of GRIM top-ranked docking poses for six HSP90α inhibitors

Ligand	Protein[a]	Pose[b]	Template[c]	GRIMscore[d]	Tc[e]	rmsd[f]
hsp90_40	4ykr_wat1a[g]	000	4ykr_wat2	0.89	0.94	0.44
hsp90_44	4ykr_wat1a	004	4ykr_wat2	0.84	0.58	2.48
hsp90_73	2xdx_wat1a	000	2xdx_wat2	0.80	0.63	0.85
hsp90_164	4yky_wat2	005	4yky_wat2	0.89	0.77	0.37
hsp90_175	4yky_wat2	008	4yky_wat2	0.87	1.00	1.67
hsp90_179	4ykr_wat2	001	4fcq	0.79	0.52	0.54

[a] Set of protein coordinates used for docking

[b] Surflex–Dock pose number

[c] Set of protein–ligand coordinates used as template for graph matching; see "Computational methods" for the numbering of conserved water molecules in PDB input structures

[d] GRIMscore

[e] 2D chemical similarity (Tanimoto coefficient) between query and template ligands, calculated from 166-bit MDL public keys

[f] Root-mean square deviations (in Å) from X-ray pose

[g] See "Computational methods" section for the numbering of water molecules

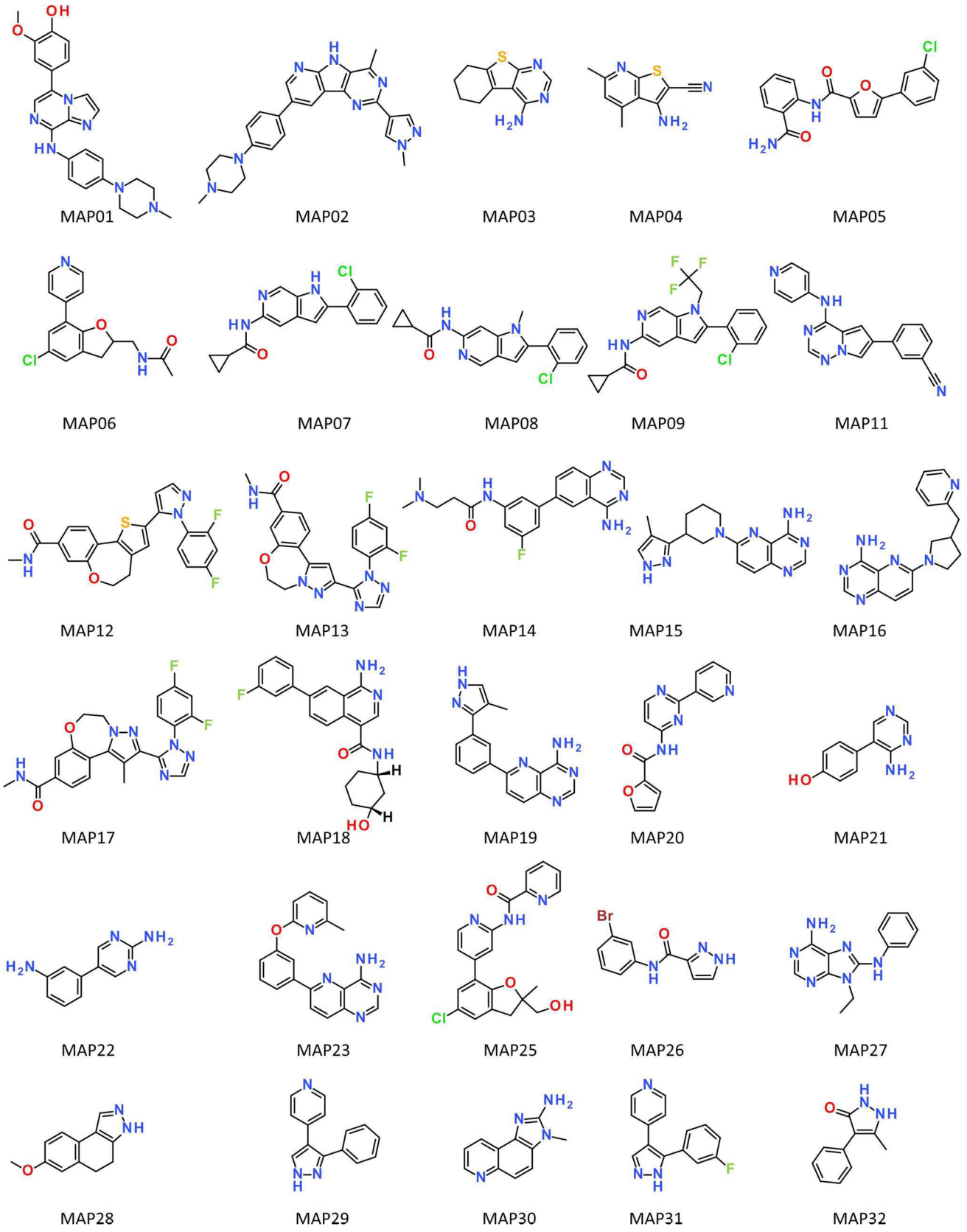


Fig. 4 Structure and name of 30 MAP4K4 inhibitors to dock

Table 3 Accuracy of pose selection (rmsd in Å to X-ray solution) for 30 MAP4K4 inhibitors

Compound	rmsd to X-ray, Å						
	GRIM-1[a]	GRIM-2[b]	GRIM-3[c]	GRIM-4[d]	GRIM-5[e]	Surflex-1[f]	Best[g]
MAP01	0.98	1.29	2.04	9.67	2.26	9.68	0.93
MAP02	4.35	4.37	3.25	3.45	3.39	1.47	1.08
MAP03	2.73	2.73	2.74	2.81	2.74	0.63	0.45
MAP04	5.12	5.21	5.21	5.21	1.95	5.46	1.54
MAP05	8.69	8.81	8.31	8.31	8.91	8.32	0.66
MAP06	5.19	5.10	4.92	2.05	2.27	2.32	1.10
MAP07	1.73	1.75	0.94	8.72	8.67	1.18	0.86
MAP08	0.71	1.69	1.79	0.83	1.79	0.88	0.49
MAP09	1.73	1.88	1.74	1.88	1.14	7.47	0.55
MAP11	2.26	8.30	8.39	1.36	0.63	0.80	0.76
MAP12	4.21	2.56	4.23	8.60	8.60	6.53	2.11
MAP13	6.11	6.13	6.21	6.18	6.11	7.22	2.25
MAP14	1.64	1.26	1.12	1.70	1.09	4.84	0.95
MAP15	3.49	3.46	3.69	3.51	2.32	2.33	0.39
MAP16	4.52	4.43	4.24	2.92	4.24	4.58	1.16
MAP17	3.09	4.74	9.36	4.77	3.02	2.39	1.24
MAP18	0.63	2.34	1.72	1.23	1.93	1.76	0.57
MAP19	1.33	1.34	3.66	7.51	2.66	0.62	0.62
MAP20	1.03	1.86	1.61	1.25	1.86	1.93	1.03
MAP21	0.51	0.55	0.50	0.46	5.21	0.51	0.35
MAP22	5.83	6.32	6.08	6.07	6.32	0.69	0.61
MAP23	2.29	4.59	4.89	4.89	5.06	2.25	1.27
MAP25	3.67	3.60	3.73	3.76	8.18	1.53	0.42
MAP26	2.74	3.17	2.78	2.78	2.78	6.60	0.90
MAP27	0.89	1.03	1.03	1.03	1.03	3.45	0.80
MAP28	3.00	4.15	3.24	4.24	4.15	2.70	1.85
MAP29	4.24	4.16	4.16	4.14	4.14	6.65	1.17
MAP30	4.12	4.04	4.06	3.38	3.59	3.93	0.63
MAP31	6.26	6.45	4.71	4.82	5.94	4.55	0.77
MAP32	0.98	2.20	2.20	2.20	2.20	5.21	0.73
Mean rmsd	3.18	3.65	3.75	3.99	3.81	3.63	0.94

[a] 1st pose according to GRIMscore

[b] 2nd pose according to GRIMscore

[c] 3rd pose according to GRIMscore

[d] 4th pose according to GRIMscore

[e] 5th pose according to GRIMscore

[f] 1st pose according to Surflex-Dock score

[g] Lowest rmsd pose

(MAP01, MAP07, MAP08, MAP09, MAP11, MAP14, MAP18, MAP19, and MAP23) that were successfully docked (rmsd < 2.5 Å) and scored by GRIM (GRIMscore > 0.68) because the corresponding interaction patterns are quite similar to that from one of the six templates exhibiting either the same or a bioisosteric scaffold (e.g. MAP18 binding pose; Fig. 6). Importantly, polar interactions are those that contribute the most to the GRIMscore, thereby ensuring that both key hydrogen bonds to the kinase hinge region and overall shape of the bound inhibitors are shared between docked compounds and templates. The second scenario applies to the three fragments (MAP21, MAP27, MAP32) whose poses were also precisely recovered with GRIM. In all cases, the good pose was inferred by bioisosterism (same interaction pattern but different chemical structure) to a larger template ligand (Table 3). For example, the hydroxyphenyl-aminopyrimidine scaffold of MAP21 is perfectly docked (rmsd to

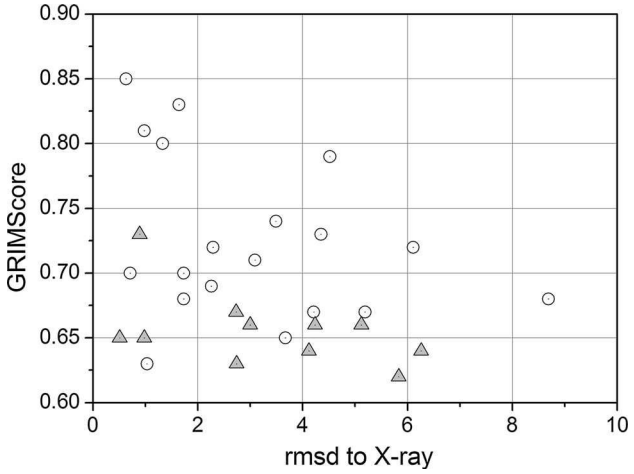


Fig. 5 Plotting the GRIMscore versus the rmsd to the X-ray pose for 30 MAP4K4 inhibitors. Lead-like and fragment-like inhibitors are represented by *white circles* and *gray triangles*, respectively

X-ray structure = 0.51 Å) to MAP4K4 because of the high similarity of its interaction pattern graph to that observed in the 4rvt PDB template, although the latter ligand exhibits a chemically different but bioisosteric scaffold (Fig. 7). It is interesting to notice that GRIM did not select an interaction pattern graph generated by a much chemically closer template ligand (e.g. 4obo, 4obq) therefore demonstrating that our pose selection protocol is really biased by protein–ligand interactions and not dominated by simple ligand chemical neighborhood. Since Surflex-Dock was able to generate at least one reliable pose for all these ligands, the reason for GRIM failure to detect it (third scenario) usually lies in wrong graph alignments dominated by hydrophobic interactions. A prototypical example is illustrated with the incorrect pose of MAP06 (rmsd to X-ray pose = 5.19 Å) by analogy to that of the 4obp template (Fig. 8) where GRIM optimizes the shape overlap between the two interaction patterns without a single shared hydrogen-bond. The overlay of the GRIM pose to that of the 4obp template notably highlights a very good match of both pyridine rings which serve as pure hydrophobic anchors to the MAP4K4 binding site. In fact, MAP06 H-bonds to the kinase hinge by its pyridine nitrogen atom (Fig. 8). This interaction is indeed found in some poses which were not rewarded by GRIM because of a lower overall GRIMscore. Since fragments with a dominant hydrophobic character exhibit simpler interaction patterns, the risk of misaligning the corresponding graphs to that of larger templates is relatively high, therefore explaining many of the herein observed failures (Table 4).

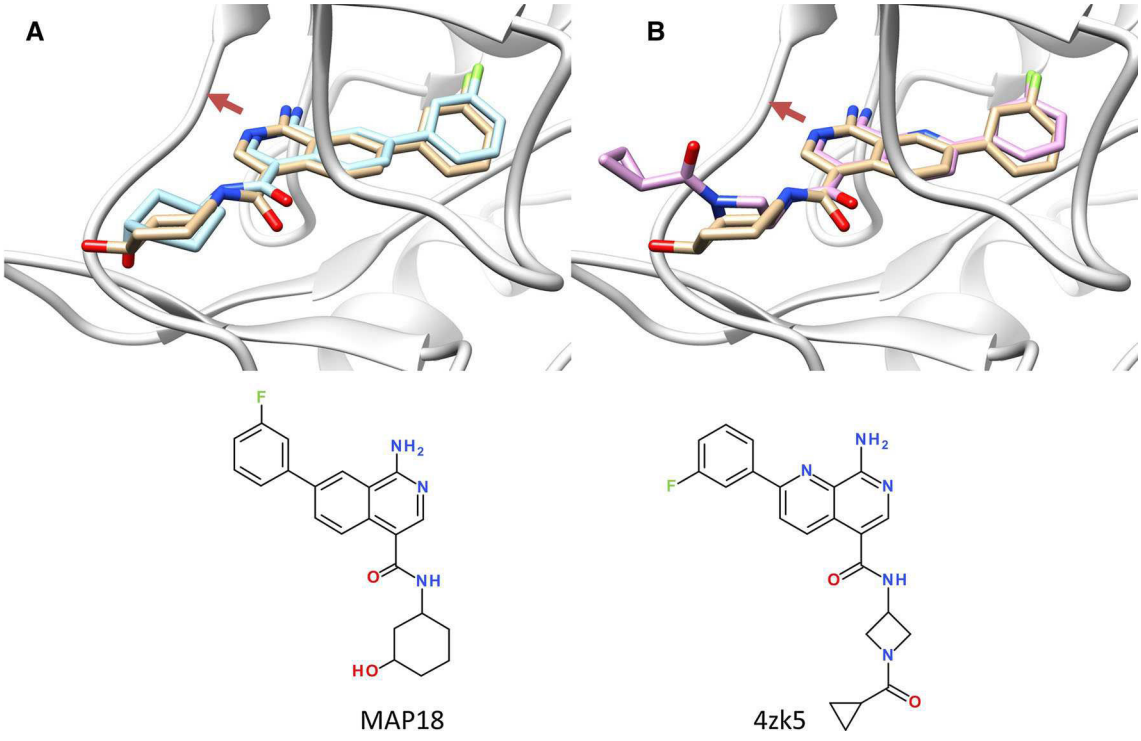


Fig. 6 Predicted versus X-ray pose of the MAP4K4 inhibitor MAP18. Heteroatoms are colored in *blue* (nitrogen), *red* (oxygen), and *green* (fluorine). A *red arrow* indicates the location of the hinge region (E106, M107, C108) of the kinase. The chemical structures of inhibitor and template are displayed below the binding poses. **a** Predicted (*tan sticks*) and X-ray poses (*cyan sticks*) of MAP18 bound to MAP4K4 (*white ribbons*). **b** The GRIM pose (*tan sticks*) has been selected by interaction pattern similarity to that of PDB template 4zk5 (*plum sticks*)

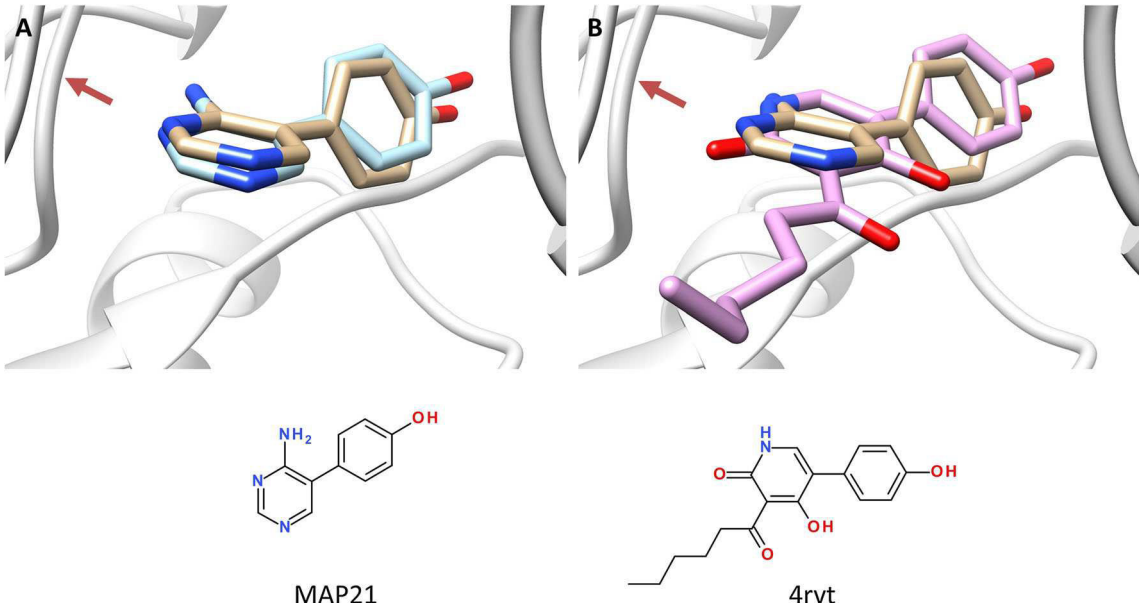


Fig. 7 Predicted versus X-ray pose of the MAP4K4 inhibitor MAP21. Heteroatoms are colored in *blue* (nitrogen), and *red* (oxygen). A *red arrow* indicates the location of the hinge region (E106, M107, C108) of the kinase. The chemical structures of inhibitor and template are displayed below the binding poses. **a** Predicted (*tan sticks*) and X-ray poses (*cyan sticks*) of MAP21 bound to MAP4K4 (*white ribbons*). **b** The GRIM pose (*tan sticks*) has been selected by interaction pattern similarity to that of PDB template 4rvt (*plum sticks*)

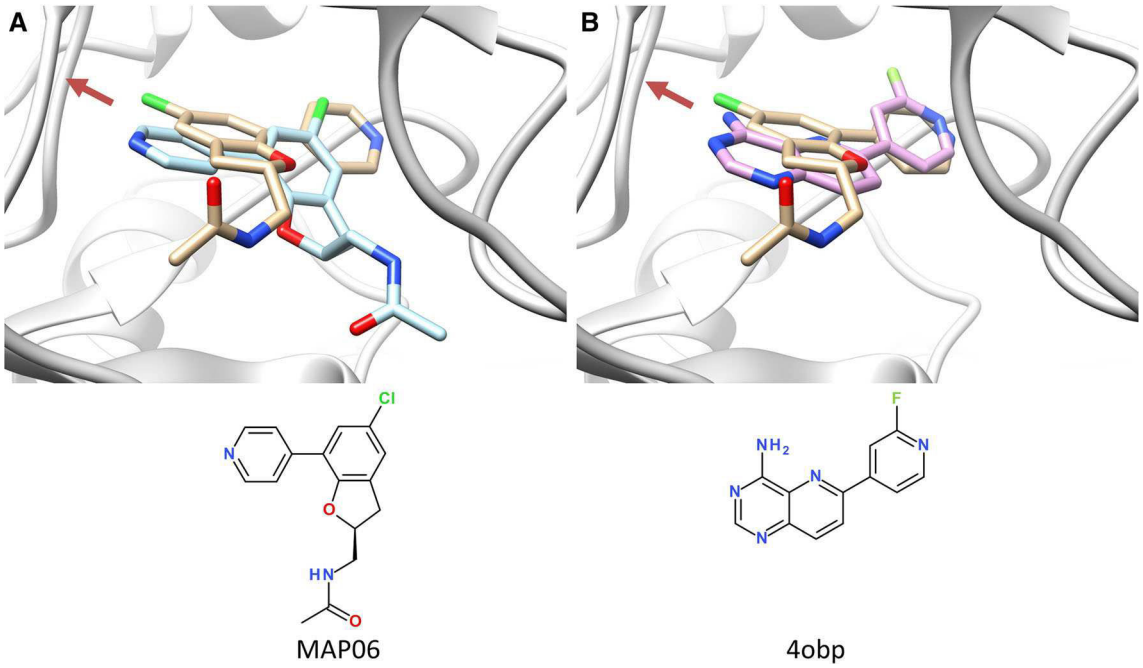


Fig. 8 Predicted versus X-ray pose of the MAP4K4 inhibitor MAP06. Heteroatoms are colored in *blue* (nitrogen), *red* (oxygen), and *yellow* (sulfur). A *red arrow* indicates the location of the hinge region (E106, M107, C108) of the kinase. The chemical structures of inhibitor and template are displayed below the binding poses. **a** Predicted (*tan sticks*) and X-ray poses (*cyan sticks*) of MAP06 bound to MAP4K4 (*white ribbons*). **b** The GRIM pose (*tan sticks*) has been selected by interaction pattern similarity to that of PDB template 4obp (*plum sticks*)

We recall here that all poses were prefiltered before GRIM scoring for hydrogen-bonding to the hinge region of the kinase. This filtering step improved the mean rmsd of the 30 MAP4K4 inhibitors from 3.62 to 3.18 Å. For 21 out of 30 inhibitors, the filter has no effect since exactly the same pose was selected by GRIM with or without filtering.

For 4 inhibitors (MAP01, MAP004, MAP12, MAP32), the filter positively contributes to the selection of better poses. Notably, the mean rmsd of compound MAP01 could be decreased from 10.78 to 0.98 Å. Conversely, the filtering was detrimental in 5 cases, most of the time with a very marginal rmsd increase. All rmsd values with and without the hydrogen bond filter are described in Supplementary Table 5.

Table 4 Characteristics of GRIM top-ranked docking poses for 30 MAP4K4 inhibitors

Ligand	Protein[a]	Pose[b]	Template[c]	GRI Mscore[d]	Tc[e]	rmsd[f]
MAP01	4obp	015	4obq	0.81	0.54	0.98
MAP02	4obq	000	4obq	0.73	0.61	4.35
MAP03	4zk5	010	4obq	0.67	0.51	2.73
MAP04	4u43	005	4obq	0.66	0.35	5.12
MAP05	4obp	003	4obq	0.68	0.46	8.69
MAP06	4obp	008	4obp	0.67	0.39	5.19
MAP07	4zk5	015	4zk5	0.70	0.62	1.73
MAP08	4zk5	000	4zk5	0.70	0.64	0.71
MAP09	4rvt	007	4rvt	0.68	0.30	1.73
MAP11	4obp	007	4obp	0.69	0.59	2.26
MAP12	4u43	008	4rvt	0.67	0.34	4.21
MAP13	4zk5	008	4zk5	0.72	0.60	6.11
MAP14	4obp	007	4obp	0.83	0.50	1.64
MAP15	4obp	001	4obp	0.74	0.47	3.49
MAP16	4zk5	005	4zk5	0.79	0.63	4.52
MAP17	4rvt	014	4rvt	0.71	0.36	3.09
MAP18	4zk5	007	4zk5	0.85	0.64	0.63
MAP19	4zk5	011	4zk5	0.80	0.45	1.33
MAP20	4u43	004	4obq	0.63	0.50	1.03
MAP21	4rvt	006	4rvt	0.65	0.36	0.51
MAP22	4zk5	017	4rvt	0.62	0.24	5.83
MAP23	4zk5	008	4zk5	0.72	0.52	2.29
MAP25	4u43	014	4obq	0.65	0.54	3.67
MAP26	4rvt	001	4rvt	0.63	0.28	2.74
MAP27	4zk5	002	4zk5	0.73	0.51	0.89
MAP28	4u43	003	4obq	0.66	0.32	3.00
MAP29	4zk5	018	4obq	0.66	0.31	4.24
MAP30	4zk5	008	4obp	0.64	0.58	4.12
MAP31	4zk5	014	4zk5	0.64	0.35	6.26
MAP32	4zk5	019	4obq	0.65	0.35	0.98

[a] Set of protein coordinates used for docking

[b] Surflex-Dock pose number

[c] Set of protein–ligand coordinates used as template for graph matching

[d] GRIMscore

[e] 2D chemical similarity (Tanimoto coefficient) between query and template ligands, calculated from from 166-bit MDL public keys

[f] Root-mean square deviations (in Å) from the a posteriori released X-ray pose

Comparative evaluation of GRIM rescoring

The release, by the D3R Grand challenge 2015 organizers, of results from all contributions permits a comparative evaluation of our pose selection method with respect to many others (Fig. 9). Two criteria have been retained to estimate the accuracy of every method. First, the mean rmsd of the best possible pose (lowest rmsd to the X-ray structure) was selected to reflect the overall quality of the posing algorithm. Second, the mean rmsd of the top-ranked pose illustrates the capacity of a scoring function to reward docking solutions that are very close to the true X-ray pose.

Among the 42 contributions to predict the binding pose of HSP90α inhibitors, GRIM is ranked 3rd when considering the average rmsd of the top-ranked pose (Fig. 9a). Seven methods deliver quite accurate answers with a mean rmsd of the top-ranked pose below 1.5 Å, one method being slightly better than GRIM (rmsd of 0.85 Å). Since contributions are anonymous, we are not aware, at the time this manuscript was written, of the corresponding method and its sophistication level.

The much more challenging MAP4K4 dataset drew less attention with 27 answers. The quality of the corresponding predictions is significantly lower than for the HSP90α dataset (Fig. 9b). Only 3 contributors predicted the pose of the 30 MAP4K4 inhibitors with an accuracy below 3.5 Å when considering the top-ranked pose. GRIM is one of these 3 methods being ranked second in this challenge (Fig. 9b). Looking at the accuracy of the best possible pose clearly highlights a docking problem since deviations to the X-ray pose remains between 2 and 3 Å for the best methods (Fig. 9b). Reasons for failures have already been discussed in the previous section of this manuscript and therefore do not only concern our docking engine (Surflex-Dock) but also all other dockers used in this competition.

We do not know whether it is the same method that slightly outperform GRIM in predicting the pose of both HSP90α and MAP4K4 inhibitors. Our interaction pattern-guided pose selection strategy is anyhow quite robust and accurate, with respect to competitor methods as it ranks 3rd and 2nd, respectively for the two sets of predictions. As to be expected, the quality of the results depends on the preexisting knowledge. When numerous and diverse interaction patterns are available for a particular target (e.g. HSP90α dataset), GRIM docking poses are very accurate. If less information is known (e.g. MAP4K4 dataset), the quality of the GRIM poses logically deteriorates but still remains better than that obtained without GRIM rescoring for the same set of poses. We have not investigated here the possibility to select more interaction pattern templates (e.g. from other protein kinases) for posing MAP4K4 inhibitors, as preliminary GRIM pairwise comparisons between the eight available MAP4K4-inhibitor complexes and 1548

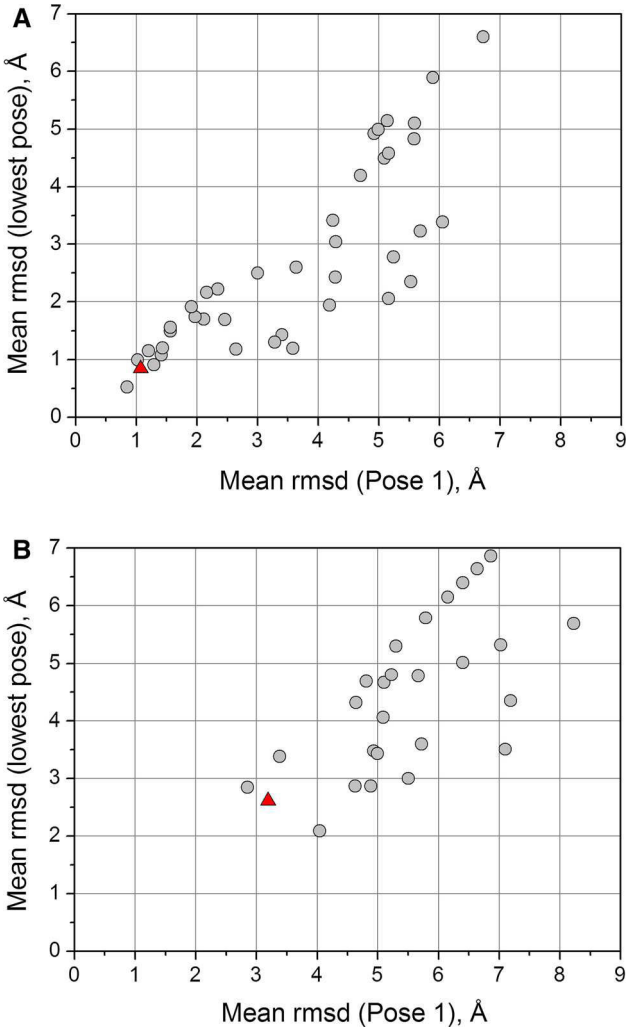


Fig. 9 Comparative evaluation of GRIM (*red triangles*) with other approaches (*gray dots*) in predicting the binding mode of 36 inhibitors (6 HSP90α inhibitors, 30 MAP4K4 inhibitors) prior to the release of their protein-bound X-ray coordinates. Posing accuracy is evaluated by the rmsd of predicted poses to the X-ray solution. **a** Plotting the mean rmsd of the top-ranked pose versus the mean rmsd of the absolute best (lowest rmsd) pose for six HSP90α inhibitors; **b** Plotting the mean rmsd of the top-ranked pose versus the mean rmsd of the absolute best (lowest rmsd) pose for 30 MAP4K4 inhibitors

protein kinase-inhibitor complexes (sc-PDB dataset) [37] were not particularly promising (GRIMscore <0.70). In some cases however, a protein family-based GRIM scoring strategy has been shown to be useful [36] and should not be forgotten. At this point, it should be recalled that selecting a near-native pose by GRIM matching does not mean that predicting the binding free energy of that pose with state-of-the art scoring functions will deliver good results. Indeed, we could not find any correlation (Pearson $R = 0.19$) between the Surflex-Dock score of GRIM-1 poses and experimental binding affinities of 180 HPS90α inhibitors (stage 2 of the challenge).

Conclusions

The herein presented GRIM method rescores docking poses by interaction pattern graph similarity to known protein–ligand X-ray structures. The methodology is both very simple and intuitive. Basically, the method automatizes the reasoning of a molecular modeler: *Does this pose remind me the binding mode of known ligands for this protein or its close homologues?*

Conceptually, it is different from many shape or template-matching docking methods [23–26] recently reported to outperform free docking in generating reliable poses. GRIM operates on freely generated docking poses but will just reward that poses which lead to interaction patterns similar to known ligands of the same or related target protein. In ca. 80–90 % of test cases, state-of-the-art docking engines propose a set of poses out of which at least one is close to the X-ray solution [12]. GRIM can therefore be used in addition to any of these dockers to prioritize the most relevant ones. The method is fast (20 ms/pose on average) and independent on the docking engine, however protein–ligand coordinates should be provided in a standard MOL2 format.

When applied to the a priori prediction of binding poses for 36 new inhibitors of two different targets, GRIM compares very favorably with competing methods as it ranked 3rd and 2nd, for HSP90α and MAP4K4 dataset respectively, in predicting near native poses. In most cases, the top-ranked pose as predicted by GRIM is the one that is the closest to the true solution. As any knowledge-based method, the accuracy of GRIM depends on existing experimental data. Depending on the target, the number and chemical diversity of co-crystallized ligands may vary quite significantly. GRIM rewards the pose with the closest interaction pattern to that seen in any other crystal of the same target, independently of how frequently this pose has already been obtained experimentally. The more chemically-diverse ligands co-crystallized with the target (or close homologues) are available, the higher the probability of the first GRIM pose being near native. The user should therefore be aware of the target-dependent applicability domain of the method, before using it blindly. The corresponding executable (IChem) is available for non-profit academic research at http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html.

Supporting information

List of 180 HSP90α inhibitors to dock, list of 176 PDB templates for HSP90α-inhibitor complexes, list of 30 MAP4K4 inhibitors to dock, list of 8 PDB templates for

MAP4K4-inhibitor complexes, effect of hydrogen-bond filtering on the quality of GRIM top-ranked poses.

Acknowledgments We thank the LABEX ANR-10-LABX-0034 Medalis for a post-doctoral fellowship to I.S. We also acknowledge the National Center for Scientific Research (CNRS, Institut de Chimie) and the Alsace Region for a doctoral fellowship to FDS. The High-performance Computing Center (University of Strasbourg, France) and the Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) are acknowledged for allocation of computing time and excellent support.

References

1. Chen YC (2015) Beware of docking! Trends Pharmacol Sci 36:78–95
2. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule–ligand interactions. J Mol Biol 161:269–288
3. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 153(Suppl 1):S7–26
4. Yuriev E, Holien J, Ramsland PA (2015) Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. J Mol Recognit 28:581–604
5. Sousa SF, Ribeiro AJ, Coimbra JT, Neves RP, Martins SA, Moorthy NS, Fernandes PA, Ramos MJ (2013) Protein–ligand docking in the new millennium—a retrospective of 10 years in the field. Curr Med Chem 20:2296–2314
6. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373
7. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57:225–242
8. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
9. Smith RD, Damm-Ganamet KL, Dunbar JB Jr, Ahmed A, Chinnaswamy K, Delproposto JE, Kubish GM, Tinberg CE, Khare SD, Dou J, Doyle L, Stuckey JA, Baker D, Carlson HA (2016) CSAR benchmark exercise 2013: evaluation of results from a combined computational protein design, docking, and scoring/ranking challenge. J Chem Inf Model 56:1022–1031
10. Damm-Ganamet KL, Smith RD, Dunbar JB Jr, Stuckey JA, Carlson HA (2013) CSAR benchmark exercise 2011–2012: evaluation of results from docking and relative ranking of blinded congeneric series. J Chem Inf Model 53:1853–1870
11. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32:742–755
12. Li Y, Han L, Liu Z, Wang R (2014) Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. J Chem Inf Model 54:1717–1736
13. Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, Chilov GG (2011) CSAR scoring challenge reveals the need for new concepts in estimating protein–ligand binding affinity. J Chem Inf Model 51:2090–2096
14. Wang JC, Lin JH (2013) Scoring functions for prediction of protein–ligand interactions. Curr Pharm Des 19:2174–2182
15. Virtanen SI, Niinivehmas SP, Pentikainen OT (2015) Case-specific performance of MM-PBSA, MM-GBSA, and SIE in virtual screening. J Mol Graph Model 62:303–318
16. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M (2005) Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 48:4040–4048
17. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model 51:69–82
18. Li L, Wang B, Meroueh SO (2011) Support vector regression scoring of receptor–ligand complexes for rank-ordering and virtual screening of chemical libraries. J Chem Inf Model 51:2132–2138
19. Zilian D, Sotriffer CA (2013) SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. J Chem Inf Model 53:1923–1933
20. Ballester PJ, Schreyer A, Blundell TL (2014) Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? J Chem Inf Model 54:944–955
21. Khamis MA, Gomaa W, Ahmed WF (2015) Machine learning in computational docking. Artif Intell Med 63:135–152
22. Gabel J, Desaphy J, Rognan D (2014) Beware of machine learning-based scoring functions-on the danger of developing black boxes. J Chem Inf Model 54:2807–2815
23. Hindle SA, Rarey M, Buning C, Lengauer T (2002) Flexible docking under pharmacophore type constraints. J Comput Aided Mol Des 16:129–149
24. Kelley BP, Brown SP, Warren GL, Muchmore SW (2015) POSIT: flexible shape-guided docking for pose prediction. J Chem Inf Model 55:1771–1780
25. Kumar A, Zhang KY (2016) Application of shape similarity in pose selection and virtual screening in CSARdock2014 exercise. J Chem Inf Model 56:965–973
26. Gao C, Thorsteinson N, Watson I, Wang J, Vieth M (2015) Knowledge-based strategy to improve ligand pose prediction accuracy for lead optimization. J Chem Inf Model 55:1460–1468
27. Deng Z, Chuaqui C, Singh J (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein–ligand binding interactions. J Med Chem 47:337–344
28. Anighoro A, Bajorath J (2016) Three-dimensional similarity in molecular docking: prioritizing ligand poses on the basis of experimental binding modes. J Chem Inf Model 56:580–587
29. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 47:195–207
30. Kelly MD, Mancera RL (2004) Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. J Chem Inf Comput Sci 44:1942–1951
31. Mpamhanga CP, Chen B, McLay IM, Willett P (2006) Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. J Chem Inf Model 46:686–698
32. Chalopin M, Tesse A, Martinez MC, Rognan D, Arnal JF, Andriantsitohaina R (2010) Estrogen receptor alpha as a key target of red wine polyphenols action on the endothelium. PLoS ONE 5:e8554
33. Venhorst J, Nunez S, Terpstra JW, Kruse CG (2008) Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. J Med Chem 51:3222–3229
34. de Graaf C, Rein C, Piwnica D, Giordanetto F, Rognan D (2011) Structure-based discovery of allosteric modulators of two related

class B G-protein-coupled receptors. ChemMedChem 6:2159–2169
35. de Graaf C, Kooistra AJ, Vischer HF, Katritch V, Kuijer M, Shiroishi M, Iwata S, Shimamura T, Stevens RC, de Esch IJ, Leurs R (2011) Crystal structure-based virtual screening for fragment-like ligands of the human histamine H(1) receptor. J Med Chem 54:8195–8206
36. Desaphy J, Raimbaud E, Ducrot P, Rognan D (2013) Encoding protein–ligand interaction patterns in fingerprints and graphs. J Chem Inf Model 53:623–637
37. Desaphy J, Bret G, Rognan D, Kellenberger E (2015) sc-PDB: a 3D-database of ligandable binding sites—10 years on. Nucleic Acids Res 43:D399–D404
38. Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. Commun ACM 16:575–577
39. Johnston HC (1976) Cliques of a graph—variations on the Bron–Kerbosch algorithm. Int J Parallel Prog 5:209–238
40. Theobald DL (2005) Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. Acta Crystallogr A 61:478–480
41. Jain AN (2007) Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J Comput Aided Mol Des 21:281–306
42. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55:6582–6594
43. Drug Design Data Resource. https://drugdesigndata.org/about/grand-challenge-2015
44. Bietz S, Urbaczek S, Schulz B, Rarey M (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. J Cheminform 6:12
45. Tripos International, St. Louis, MO 63144–2319, USA
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
47. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212
48. Molecular Networks GmbH, Erlangen, Germany
49. Kung PP, Sinnema PJ, Richardson P, Hickey MJ, Gajiwala KS, Wang F, Huang B, McClellan G, Wang J, Maegley K, Bergqvist S, Mehta PP, Kania R (2011) Design strategies to target crystallographic waters applied to the Hsp90 molecular chaperone. Bioorg Med Chem Lett 21:3557–3562

Franck Da Silva

Cartographie des interfaces protéine-protéine et recherche de cavités droguables

Résumé

Les interfaces protéine-protéine sont au cœur de nombreux mécanismes physiologiques du vivant. Les caractériser au niveau moléculaire est un donc enjeu crucial pour la recherche de nouveaux candidat-médicaments.

Nous proposons ici de nouvelles méthodes d'analyse des interfaces protéine-protéine à visée pharmaceutique. Notre protocole automatisé détecte les interfaces au sein des structures de la Protein Data Bank afin de définir les zones d'interactions à potentiel pharmacologique, les cavités droguables, les ligands présents à l'interface ainsi que les pharmacophores directement déduits à partir des cavités. Notre méthode permet de réaliser un état de l'art des informations disponibles autour des interfaces protéine-protéine ainsi que de prédire de nouvelles cibles potentielles pour des molécules candidats médicaments.

Mots clés : base de données, bioinformatique, cavité, chémoinformatique, classifieurs, interface protéine-protéine, pharmacophore, protéine, site de liaison.

Résumé en anglais

Protein-protein interfaces are involved in many physiological mechanisms of living cells. Their characterization at the molecular level is therefore crucial in drug discovery.

We propose here new methods for the analysis protein-protein interfaces of pharmaceutical interest. Our automated protocol detects the biologicaly relevant interfaces within the Protein Data Bank structures, droguables cavities, ligands present at the interface and pharmacophores derived directly from the cavities. Our method enables a state-of- the-art of all available structural information about protein-protein interfaces and predicts potential new targets for drug candidates.

Keywords: binding site, cavities, classifiers, computational biology, computational chemistry, database, pharmacophore, protein, protein-protein interface.