Détection, caractérisation et comparaison des interactions protéine - ligand

Nicolas Shinada

Université Sorbonne Paris Cité

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale Hématologie, Oncogenèse et Biothérapies ED 561

INSERM UMR_S1134 / Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB)

Discngine S.A.S

Détection, caractérisation et comparaison des interactions protéine - ligand

Par Nicolas Shinada

Thèse de doctorat de Biothérapies

Dirigée par Alexandre de Brevern

Présentée et soutenue publiquement à l'Institut National de la Transfusion Sanguine le 28 Janvier 2019

Président du jury : Rodrigues-Lima, Fernando / PR / Université Paris 7 Diderot

Rapportrice : Douguet, Dominique / CR / Université de Nice Sophia – Antipolis

Rapporteur : Bonnet, Pascal / PR / Université d'Orléans

Examinatrice : Acher, Francine / DR / Université Paris Descartes

Examinateur : Sperandio, Olivier / CR / Institut Pasteur

Directeur de thèse : de Brevern, Alexandre / DR / INTS

Co-directeur de thèse : Schmidtke, Peter / Chercheur / Discngine

Détection, caractérisation et comparaison des interactions protéine - ligand

Appréhender le mécanisme fin de la liaison d'un ligand sur sa protéine est un enjeu majeur de l'industrie pharmaceutique. L'objectif principale de cette thèse est d'améliorer la compréhension de ce mécanisme par l'étude des interactions moléculaires. A cette fin, une méthode de détection de contacts à grande échelle fut mise en place. Les premiers chapitres illustrent les différents types d'interaction recensés dans la littérature ainsi que les outils qui ont été développés au cours de cette thèse pour en permettre l'analyse. A l'aide des contacts générés par notre méthode, une analyse exhaustive a été réalisée sur un ensemble de complexe protéine - ligand caractérisant l'environnement d'interaction des atomes halogènes. Les résultats obtenus mettent en avant la complexité de la description des interactions dans un contexte protéine - ligand. Une analyse structurale des modes de liaison des ligands issus de la PDB a été effectuée pour évaluer de manière quantitative cette diversité. Un travail spécifique sur ka redondance importante des représentations de complexes protéine - ligand a été effectué amenant une stratégie innovante dédiée aux complexes protéine - ligand. L'interaction et la fonction d'une protéine étant très étroitement liée à sa structure, le comportement dynamique des structures hélicoïdales dans les protéines a été étudiée. L'utilisation d'une méthode de regroupement met en évidence une stabilité jusque-là peu soupçonnée des hélices les plus rares.

Mots clefs : chémoinformatique, bioinformatique, dynamique moléculaire, conception médicamenteuse, protéine - ligand, récepteur, mode de liaison, interaction moléculaire, contact, fragment moléculaire, halogène, structure des protéines

Detection, characterization and comparison of molecular interactions in protein - ligand

Apprehending the binding mechanism in a protein – ligand complex is a major goal in pharmaceutical industry. The objective of this thesis was to improve the understanding of this mechanism through molecular interactions study. Consequently, a large-scale contact detection protocol was designed to achieve this goal. The first chapters highlight known interaction types observed in the literature and the resulting tools that were developed during this thesis. Using our dataset of intermolecular contacts, a comprehensive analysis underlines the intricacy of describing interaction patterns of halogen atoms in the protein-ligand context. Then, a structural comparison of ligand binding modes quantitatively assesses its diversity on the entire PDB dataset. Finally, protein function and interaction mechanism are strongly related to its structure. Using a clustering approach, dynamic behavior of helix structures was highlighted through transitional patterns and unsuspected stable conformations for rare helices.

Keywords: chemoinformatics, bioinformatics, molecular dynamics, drug design, protein-ligand, receptor, binding mode, molecular interaction, contact, molecular fragment, halogen, protein structure

Table des matières

Liste des illustrations

Liste des tableaux

Première partie : Structure des protéines et mécanisme de liaison

Chapitre 1 : Structure des protéines

A. Les molécules de la vie

La définition du vivant est complexe. Une cellule est composée au moins d'une membrane composée de lipides, de protéines qui assurent les fonctions biologiques majeures, d'Acide Désoxyribonucléique Nucléotide (ADN) qui stocke l'information et d'Acide Ribonucléique Nucléotide (ARN). Son ADN définit chaque être vivant et joue le rôle de support de l'information définissant un être vivant spécifique. Sa composition est basée sur 4 bases azotés : adénine, guanine, thymine et cytosine. Les molécules d'ADN vont être transcrites en molécule ARNm par le biais d'un mécanisme complexe (pré-ARNm, épissage, ...). Ce brin d'ARNm permettra ensuite la création des protéines lors du processus de traduction.

Dans ce processus, le ribosome est l'élément biologique responsable de la traduction. En fonction du triplet de bases nucléiques (ou codon) reconnus le long de l'ARNm, il va ajouter un acide aminé particulier à la chaine protéique naissante. Vingt acides aminés naturels différents sont utilisés. L'ajout de ces acides aminés est séquentiel, il se fait itérativement suivant l'enchainement des codons pour former une chaine peptidique. La formation de celle-ci est assurée par des liaisons covalentes entre les acides aminés, ce sont les liaisons peptidiques. Un acide aminé lié par une liaison peptidique est appelé plus communément résidu.

B. Composition des protéines

Ces acides aminés ont tous une structure commune composé d'une chaîne principale formée d'un acide carboxylique et d'un groupement fonctionnel amine, et d'une chaine latérale qui leur est propre. Ces deux premières fonctions chimiques permettent la formation de liaison

peptidique par l'intermédiaire du groupement carboxylique du résidu n-1 et du groupement amine du résidu n.

La chaine latérale, distincte pour chaque acide aminé, lui confère des propriétés spécifiques. Il existe une catégorisation de ces acides aminés selon leurs propriétés électrochimiques. Les acides aminés disposant d'un cycle aromatique au sein de leur chaine latérale (tryptophane, tyrosine, phénylalanine et l'histidine selon sa protonation) sont généralement regroupés ensemble. Les acides aminés dont la chaîne latérale est chargée à pH biologique (7,4) sont catégorisés en deux sous-groupes : les acides aminés chargés positivement (lysine, arginine et dans certains cas l'histidine) et négativement (acide aspartique et acide glutamique). La thréonine, cystéine, sérine, asparagine et glutamine sont des acides aminés considérés comme polaires. Ceux disposant d'une chaine latérale hydrophobe forment également un groupe : alanine, valine, isoleucine, valine, glycine, méthionine. Enfin, la proline se distingue puisque sa chaine latérale est une partie intégrante du squelette polypeptidique et est donc souvent considéré comme un cas à part entier (acide iminé) (voir Figure 1).

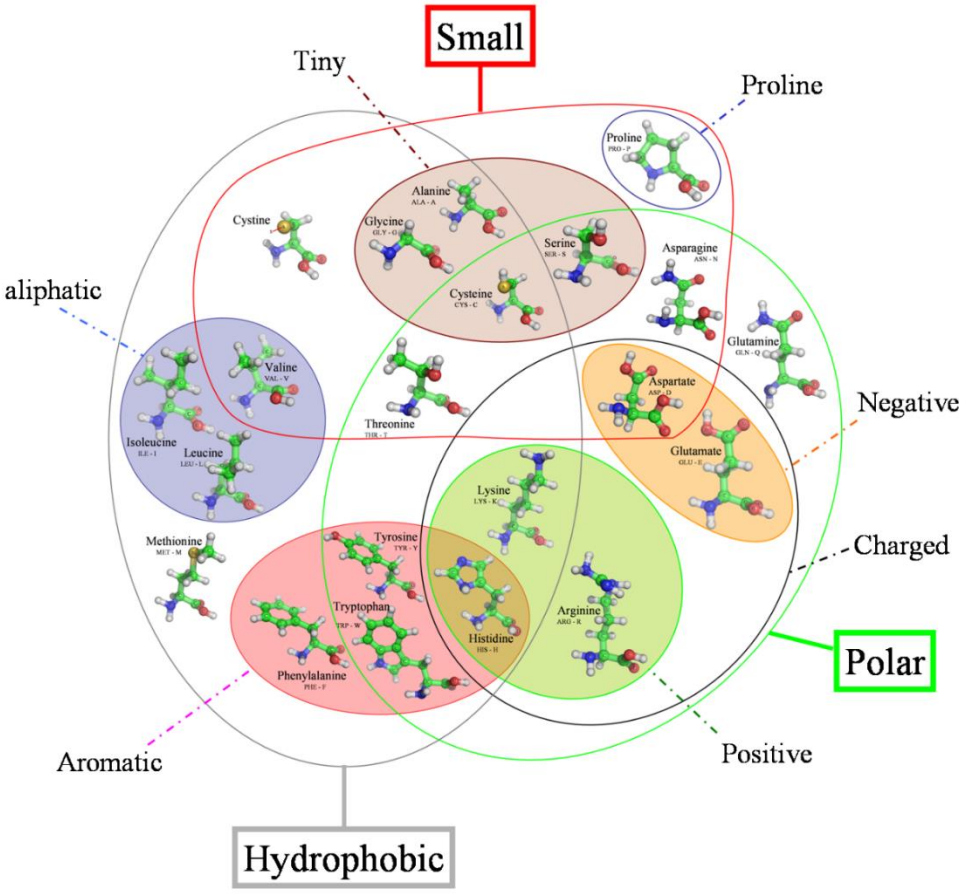


Figure 1 Structure et classification des 20 acides aminés naturels présents dans l'organisme humain selon les propriétés chimiques [1].

C. Structure protéique

L'association d'acides aminées forme une chaine polypeptidique. Cette dernière, seule ou complexée à d'autres chaines polypeptidiques, constitue la macromolécule biologique active, la protéine. La structure dite primaire représente l'enchainement successif des résidus le long de cette chaîne. Les différentes protéines vont se distinguer par leur composition en acides aminés mais aussi par le nombre de chaines polypeptidiques les composant. En effet, la composition spécifique en acide aminé va déterminer notamment sa structure tridimensionnelle, comme illustré sur la Figure 2 [2].

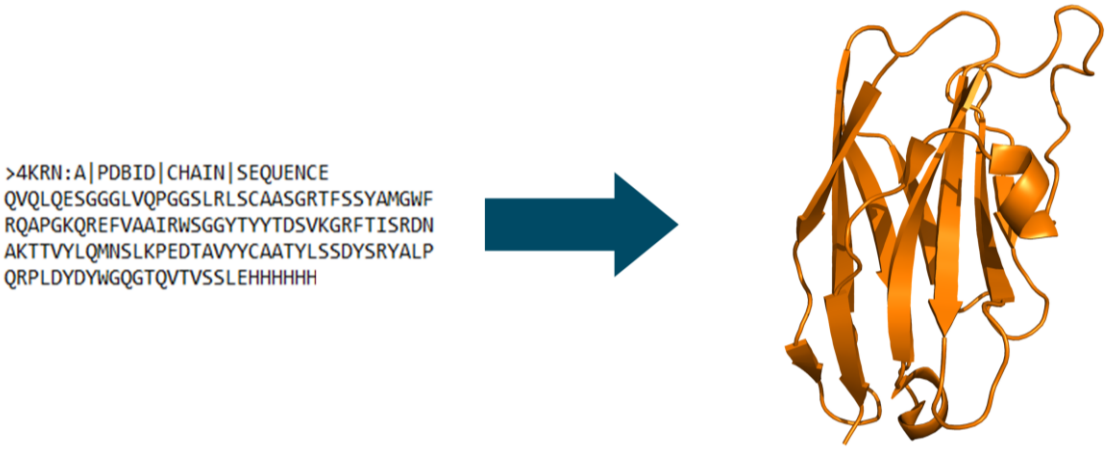


Figure 2 Séquence en acides aminés et sa structure tridimenssionnelle correspondante du domaine VHH d'un anticorps inhibiteur du récepteur de l'EGF résolue par cristallographie par rayons X (code PDB 4krn, image générée par PyMOL).

En 1951, Pauling & Corey [3] proposent à partir de modèles chimiques l'existence de motifs peptidiques répétés, et mettent en évidence des motifs structuraux fréquents nommées structures secondaires. En 1963, Ramachandran [4] met en évidence la redondance de conformations angulaires préférentielles entre deux résidus successifs observés dans les protéines. La répartition des angles dièdres ψ et ϕ décrivant la torsion du résidu n par rapport au résidu n-1 a été étudiée (voir Figure 3A). La répétition de ces angles, i.e. de conformations 3D préférentielles entre deux résidus, est la conséquence d'un potentiel énergétique favorisant ces arrangements conformationnels (voir Figure 3B).

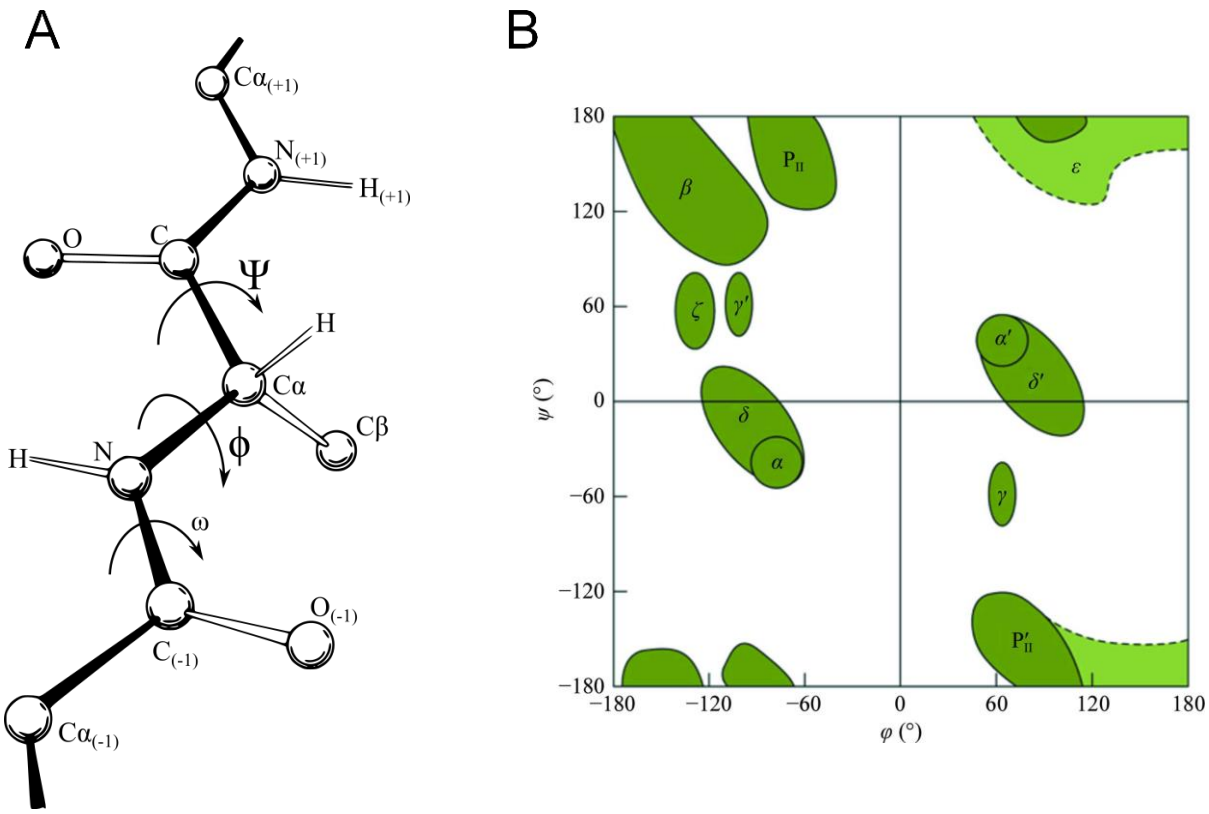


Figure 3 Description schématique des angles φ et ψ entre les résidus N-1, N et N+1 et B. Distribution bidimensionnelle, ou diagramme de Ramachandran, de l'ensemble des angles φ et ψ observées dans les structures protéiques [5, 6].

La succession spécifique d'angles dièdres ϕ et ψ ainsi que la présence de réseaux spécifiques de liaisons hydrogènes au niveau du squelette polypeptidique a mis en exergue l'existence de motifs tridimensionnels spécifiques.

Parmi les conformations les plus fréquentes, les motifs hélicoïdaux sont largement représentés : il est estimé qu'un résidu sur 3 est engagé dans ce type de conformation.

L'hélice α, la forme hélicoïdale la plus fréquemment observée, a un motif répété, *pattern* en anglais, de liaisons hydrogènes entre les résidus i et i+4, stabilisant ainsi leur forme spécifique (voir Figure 4A). Les valeurs moyennes d'angles ϕ et ψ adoptés par les hélices α se situent autour de -60° et -45°. Différents motifs hélicoïdaux, plus rares, ont été recensés comme les hélices 310 et hélices π présentant un réseau de liaisons hydrogènes tous les 3 et 5 acides aminés respectivement.

Un autre motif structural courant est le feuillet β composé de brins β. Un brin β est une partie de la chaine polypeptidique dont le squelette forme des brins étendus, relativement linéaires.

Composé de 3 à 10 acides aminés, plusieurs brins vont former un feuillet β par la présence d'un réseau de liaisons hydrogènes latérales entre leurs squelettes (voir Figure 4B). En conséquence, les angles ϕ et ψ observés le long d'un brin β se situent en moyenne aux alentours de -130° et 125° (voir Figure 3B).

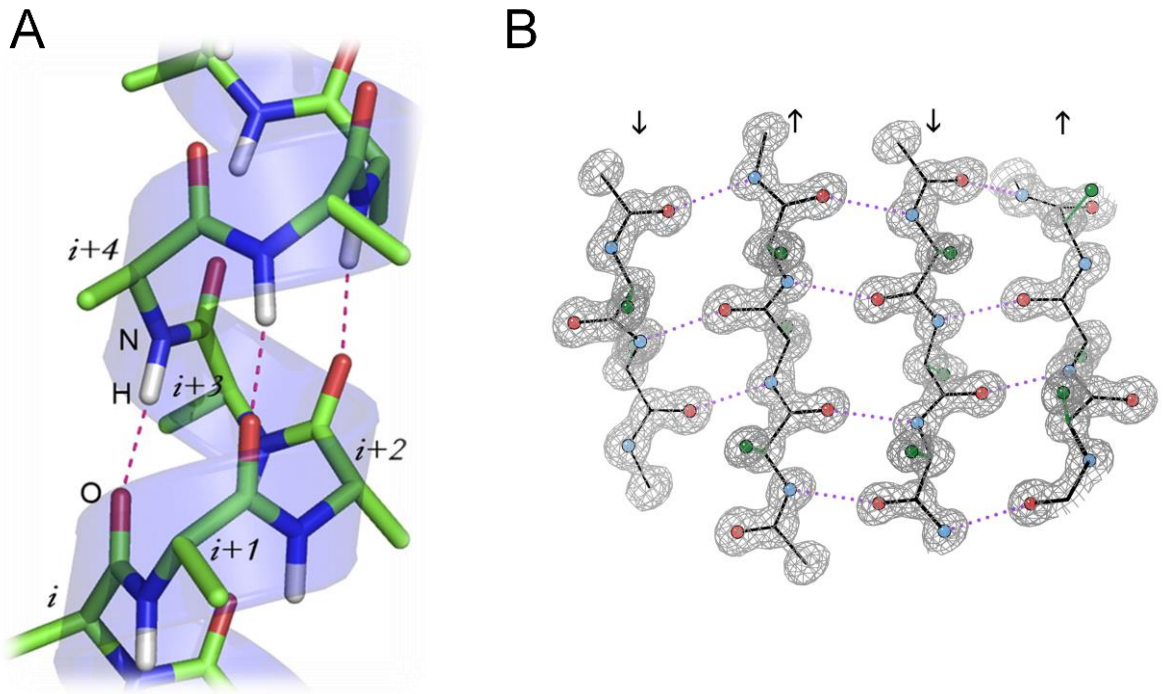


Figure 4 Représentations 3D de structures secondaires A. Motif de liaisons hydrogènes présent dans une hélice α B. Motif de liaisons hydrogènes dans un feuillet β antiparallèle [5, 7].

Il existe d'autres structures secondaires beaucoup moins étudiées telles que les hélices polyproline de type II et les coudes β.

La disposition de ces structures secondaires au sein d'une protéine va amener au repliement général de la protéine, sa structure tertiaire. Cette forme spécifique permettra la réalisation de sa(ses) fonction(s) biologique(s).

Certaines protéines existent sous la forme d'un ensemble de structures tertiaires, parfois identiques (homomultimères) ou différentes (hétéromultimères). L'agencement de structures tertiaires entre elles constitue la structure quaternaire. Certains transporteurs transmembranaires par exemple, tel que la protéine CFTR (*Cystic fibrosis transmembrane conductance regulator*), sont constitués de plusieurs sous-unités et ne sont fonctionnels que sous forme de complexe (voir Figure 5).

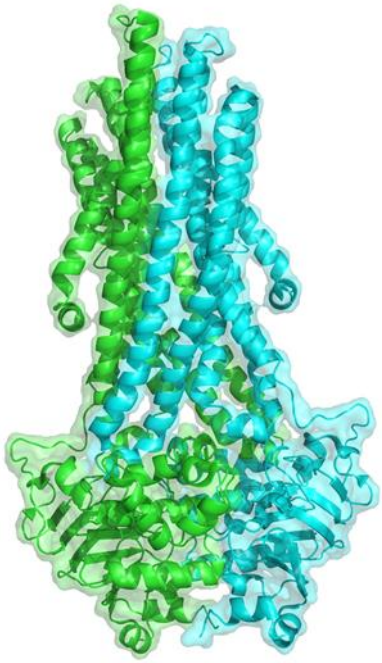


Figure 5 Représentation 3D de la structure quaternaire de domaines transmembranaires du transporteur CFTR (MSDs pour Membrane-Spanning Domains, code PDB 4a82). Chaque chaine de l'homodimère est colorée individuellement.

D. Résolutions et qualité des structures protéiques

a. Techniques de résolution

L'obtention de structure 3D protéique a été réalisé pour la première fois en 1958 par la résolution du modèle tridimensionnelle de la myoglobine. J.C. Kendrew et collaborateurs établissent ainsi le premier modèle par cristallographie aux rayons X, dont l'élucidation permit à J.C. Kendrew d'obtenir le prix Nobel de Chimie en 1962 [8]. 60 ans plus tard, malgré de nombreuses avancées dans la physique et de la chimie, la cristallographie reste la technique la plus utilisée pour obtenir une structure protéique tridimensionnelle car les structures résultantes sont notamment jugées plus stables et elle peut s'appliquer à une grande variété de tailles macromoléculaires [9].

Cette approche nécessite l'obtention des macromolécules sous formes de cristaux dans des conditions physico-chimiques très spécifiques. Une source de rayonnement, ici les rayons X, est ensuite projetée sur le cristal. Grâce à la longueur du faisceau de rayons X, les électrons (autour de chaque atome) vont diffracter les faisceaux de rayons X, leur disposition spatiale se traduisant ainsi par des perturbations du faisceau sur le détecteur, illustré sur la Figure 6A.

L'opération est répétée sur les différentes faces du cristal et la position des atomes ainsi que leur arrangement dans l'espace est ensuite approximée en utilisant la transformation de Fourrier, correspondant à la densité électronique (voir Figure 6B). La densité électronique obtenue est une vue statistique de la structure sur 2 grandeurs : le temps de la prise, et le désordre au sein du cristal. Le modèle final raffiné au cours d'étapes itératives est une structure représentative correspondant au mieux au nuage de densité observé.

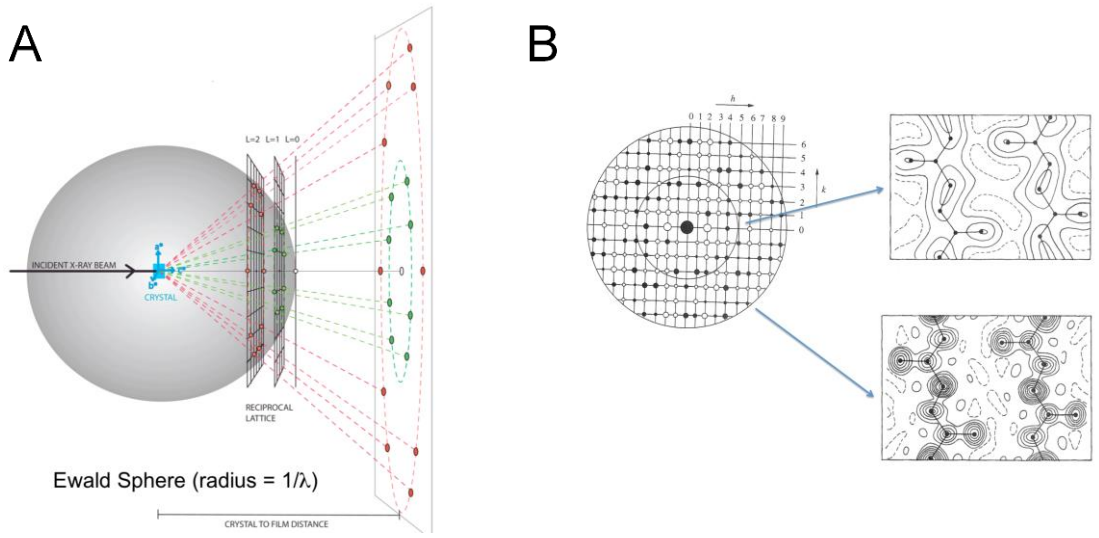


Figure 6 Représentation schématique de la méthode de résolution par cristallographie à rayons X A. Projection d'un rayon X sur le cristal ainsi que la diffraction obtenue B. A partir de la diffraction, reconstruction des phases et application de la transformation de Fourier pour l'obtention des densités électroniques.

La seconde technique fréquemment utilisée dans la résolution de structure protéique est la spectroscopie par Résonance Magnétique Nucléaire, plus communément appelée RMN. En 1985, la première structure d'une protéine fut obtenue par RMN par K. Wuthrich, prix Nobel de chimie en 2003 [10]. Un des principaux avantages de cette technique réside dans le maintien de la protéine dans une solution sans avoir à passer par une phase de transformation de type cristallisation, permettant notamment d'étudier le comportement du solvant. Cette solution permet également d'étudier la structure de protéines non cristallisables. Ce procédé repose sur le spin nucléaire que peuvent avoir certains atomes (nombres de protons et neutrons impairs) comme l'hydrogène ou l'azote. La molécule, placée dans un champ magnétique, va voir ses noyaux s'exciter et émettre un rayonnement sous forme de radiofréquence propre à chaque atome. La fréquence de résonance émise par chaque atome est spécifique à sa nature, au nombre d'atomes l'environnant, à la nature de ses liaisons

covalentes…. Ce phénomène est plus communément appelé déplacement chimique, exprimé en ppm (partie par millions). Un récepteur va capter ces résonances et les retranscrire sous forme de spectres, chaque spectre représentant un atome spécifique. L'interprétation du spectre de RMN permet d'obtenir des contraintes de distance entre atomes, et l'application de ces contraintes permet ensuite d'élaborer le modèle 3D de la protéine. La limitation principale de cette technique concerne sur la taille maximale des protéines pouvant être résolue, n'excédant que difficilement 30kDa.

D'usage moins fréquent, la cryo-microscopie électronique permet de visualiser de gros complexes macromoléculaires à une échelle de taille de l'ordre du nanomètre telles que les virus par exemple. Des exemples récents montrent de belles réussites [11]. Enfin, d'autres techniques émergent dans le domaine de la résolution de structures protéique telles que XFEL (laser à électrons libres et rayons X) [12] ou la diffraction par fibres [13].

Ces structures protéiques, au-delà de contenir la protéine en elle-même, peuvent aussi être associées à des molécules d'eau, des ions, des cofacteurs, et aussi de ligands, qui nous intéressent tout particulièrement ici.

b. *Protein DataBank* (PDB)

L'ensemble de ces structures est stocké sous forme de fichiers textes formatés. Ces fichiers ont été nommés selon la plateforme sur laquelle la plus grande partie est hébergée et qui en définit les conventions : la Protein Data Bank (PDB, https://www.rcsb.org/) [14]. Cette base de données, crée en 1971 par Hamilton et collaborateurs [15], recense et stocke la majorité des structures protéiques. Elle est publique et accessible gratuitement, chaque soumission faisant l'objet d'une évaluation spécifique. Elle était composée de 23 553 structures en 2003, mais ce nombre a depuis augmenté pour contenir plus de 146 000 structures macromoléculaires en Novembre 2018. Cette amélioration est notamment due à l'évolution des techniques et la réduction du coût de résolution des protéines atteignant aujourd'hui plus de 10 000 structures ajoutées chaque année.

Cette plateforme publique, en plus de conserver les fichiers pdb, donne accès à un grand nombre d'informations grâce à son interface web et ses connexions à d'autres bases de données. Il est ainsi possible de retrouver certaines annotations issues de UniProt (https://www.uniprot.org/) ou SCOP (http://scop.berkeley.edu/) par exemple. UniProt

(Universal Protein Resource) est une base de données recensant les séquences protéiques, ainsi que des annotations fonctionnelles provenant de différents organismes [16]. SCOP (pour Structural Classification of Proteins), est une base de données répertoriant la classification des domaines protéiques à partir de leur structure, en incorporant l'information de séquence [17].

c. Qualité de la structure

Toutefois, ces structures protéiques ne sont pas qualitativement équivalentes. La majorité des structures présentes dans la PDB provenant de cristallographies à rayons X, la qualité d'une structure est le plus fréquemment jugé par sa résolution. Ce critère quantitatif représente les distances des atomes positionnés dans la carte de densité électronique. Plus la résolution est basse, plus les atomes peuvent être distingués les uns des autres. La résolution est donc dépendant de la flexibilité d'une molécule, ainsi des régions plus stables auront une meilleure résolution que des structures locales flexibles. Ce critère est généralement moyenné, l'ensemble des valeurs sont accessibles dans la PDB. Aussi, la présence de détails à petite échelle sera plus importante dans des meilleures résolutions, la résolution est dite faible (voir Figure 7). Il est généralement accepté par la communauté scientifique qu'une résolution inférieure à 2.5Å corresponde à une structure correcte, relativement fiable (voir Figure 7B). Une résolution autour de 1Å permet même de positionner de manière relativement précise une grande partie des atomes d'hydrogènes (voir Figure 7A). En relation avec les techniques développées auparavant, la grande majorité des structures cristallographique ont une résolution autour de 2Å dans la PDB. Les structures cryo-MET se situent plutôt aux alentours de plus de 10Å, mais de récentes avancées ont permis d'atteindre des résolutions de quelques angströms [11].

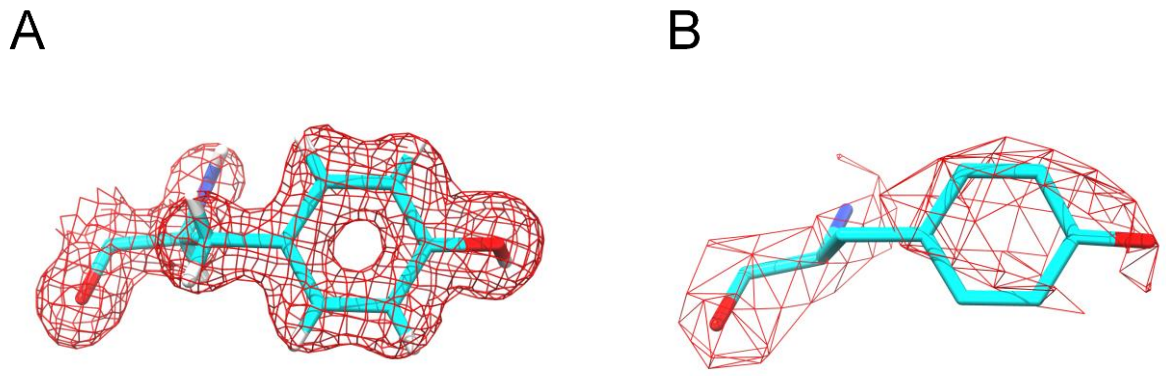


Figure 7 Représentation 3D de la densité électronique d'une tyrosine issue de la protéine HSP90. A. Résolution de 1,2Å (code PDB 3t0h) B. Résolution de 2,8Å (code PDB 2xhx).

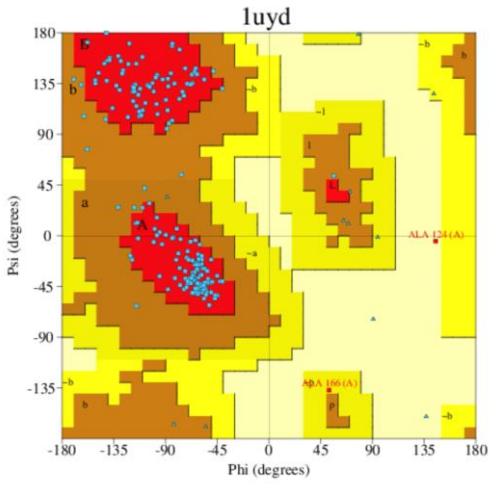
Le R-facteur, ou facteur cristallographique est une mesure quantifiant la qualité du modèle atomique généré par rapport aux données cristallographiques. La position des atomes étant obtenue par la diffraction de rayons X sur un cristal, le cristallographe va générer un modèle 3D sur lequel sera calculé une diffraction théorique. La différence entre la diffraction expérimentale et la prédiction du modèle constitue ce R-facteur. Une valeur proche de 0 signifie un modèle identique aux données cristallographiques tandis qu'une valeur de 0,6 est assimilée à de l'aléatoire. Une phase itérative d'affinage du modèle est effectuée afin qu'il concorde au mieux aux données expérimentales, diminuant ainsi le R-facteur. Cet affinage est réalisé sur 90% des données expérimentales d'une structure. Le R-facteur libre sera calculé à la fin de l'affinage sur les 10% restant afin d'évaluer le biais introduit lors de la phase d'optimisation du modèle.

La teneur en eau du cristal a un rôle important dans la qualité de la structure. Plus un cristal est riche en eau, plus faible sera la diffraction et plus grande sera la résolution [18]. Au-delà de l'aspect quantitatif, la comparaison visuelle de la carte de densité électronique par rapport à la disposition de chaque atome permet aussi de juger de la qualité.

Des analyses structurales peuvent aussi être effectuées afin d'évaluer une structure. La longueur des liaisons covalentes est généralement considérée pour éviter toutes aberrations moléculaires. De même, les déviations angulaires ϕ et ψ au niveau des liaisons peptidiques sont évaluées et comparées au diagramme de Ramachandran. L'absence de clashs stériques

détectables par les distances entre atomes permet aussi de s'assurer du bon positionnement des atomes. Des outils dédiés tel que ProCheck [19] ou MolProbity [20] ont été conçus pour répondre à ces questions, illustré sur la Figure 8. L'évaluation de la qualité des structures cristallographiques est en constante évolution et de nouvelles mesures et outils sont fréquemment mis à disposition, notamment par la X-Ray Validation Task Force de la PDB [21].

A



B

		No. of residues	%-tage
Most favoured regions	[A,B,L]	170	90.4%
Additional allowed regions	[a,b,l,p]	16	8.5%
Generously allowed regions	[~a,~b,~l,~p]	1	0.5%
Disallowed regions	[XX]	1	0.5%*
Non-glycine and non-proline residues		188	100.0%
End-residues (excl. Gly and Pro)		2	
Glycine residues		15	
Proline residues		4	
Total number of residues		209	

Parameter	Score	Average Score
Dihedral angles:-		
Phi-psi distribution	-0.18	
Chi1-chi2 distribution	-0.07	
Chi1 only	0.18	
Chi3 & chi4	0.29	
Omega	-0.90*	
		-0.31
Main-chain covalent forces:-		
Main-chain bond lengths	0.14	
Main-chain bond angles	0.04	
		0.09
OVERALL AVERAGE		-0.15

C

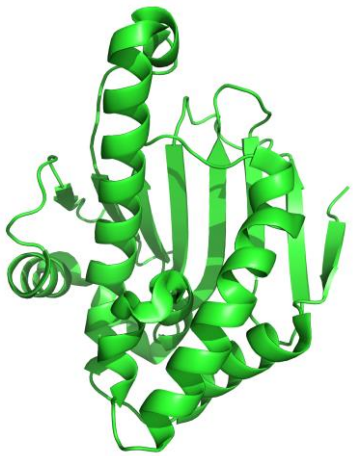


Figure 8 Analyse de la qualité d'une structure issus de ProCheck pour la protéine HSP90α (code PDB 1uyd) A. Diagramme de comparaison des angles φ et ψ mesurés par rapport au diagramme de Ramachandran. B. Analyse statistique du diagramme de Ramachandran et scores correspondants. C. Représentation 3D de la protéine.

Chaque fichier pdb comporte aussi une information sur des facteurs de température pour chaque atome présent dans la structure, nommés B-facteurs. L'agitation des atomes, provoquée par la température, induit une distribution élargie de la densité d'électrons qui

sera quantifiée par cette valeur. La variabilité d'une structure au sein des mailles d'un cristal participe également à cette baisse de précision de la densité électronique. Ainsi, le B-facteur reflète la confiance avec laquelle la position de l'atome est approximée dans la structure. Une valeur en dessous de 10 permet un positionnement relativement exact alors qu'une valeur supérieure à 50 signifie grossièrement que l'atome en question est très mobile et sa position est donc très vague.

Toutefois, chaque structure est résolue dans des conditions différentes. Ainsi, les propriétés physico-chimiques et de température vont varier en fonction des équipes de cristallographes, de la protéine à résoudre etc. etc. Les B-facteurs, spécifiques de chaque cristal, doivent généralement être normalisé en fonction de la moyenne des B-facteurs du cristal (voir Figure 9B). Une même protéine peut tout à fait avoir deux conformations différentes pour plusieurs raisons (i) différents états d'équilibre atteints dans le cristal, (ii) des conditions de cristallographie différentes, et (iii) des régions à fort désordre structural.

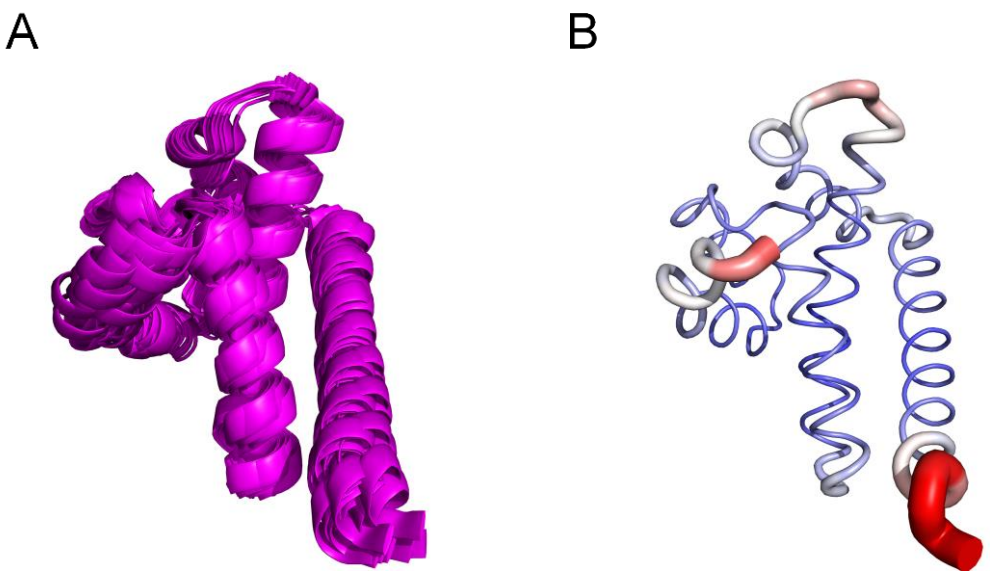


Figure 9 Illustrations 3D de la protéine co-chaperone CDC37 résolu par A. RMN (code PDB 2k5b) et par B. cristallographie à rayons X (coloration par facteurs B, code PDB 2w0g, images générées par PyMOL). Les zones caractérisées comme flexibles par les valeurs élevées de facteurs B (en rouge) sont aussi observables dans la structure RMN.

d. Aspect dynamique

Un aspect important à considérer dans l'étude des structures protéiques repose sur leur caractéristique dynamique. En effet, la forme d'une protéine n'est pas rigide et déterminée

dans le temps. Elles vont adopter un nombre de conformations préférentielles, stables énergétiquement, sous de multiples contraintes tels que le solvant, des interactions moléculaires et des contraintes intra-protéines. Les structures 3D obtenues sont donc comparables à des photographies à un instant T dans un état d'équilibre.

Chapitre 2 : Interactions entre atomes

Toute fonction biologique est réalisée par l'intermédiaire d'interactions entre deux voire plusieurs éléments. La traduction d'un ARN messager en acides aminés se fait par l'intermédiaire d'une interaction ribosome – ARNm par exemple. L'apparition d'un trouble biologique tel qu'une maladie est notamment due à une perturbation dans l'interaction entre deux éléments. Ce trouble peut être induit par une absence, une insuffisance voire une interaction trop forte entre deux molécules. La conception médicamenteuse, *drug design* en anglais, va permettre l'élaboration de molécules permettant de réguler les mécanismes dysfonctionnels en interagissant directement avec la cible désirée ou via des cibles en relation avec la voie métabolique. Néanmoins, la compréhension des mécanismes régissant l'interaction entre deux composés reste encore parcellaire aujourd'hui. Elle continue de faire l'objet de nombreuses études jusqu'à aujourd'hui [22, 23].

A. Principes de bases

Un atome est un élément constituant de toutes substances. Il correspond à un ensemble de particules subatomiques. Un atome est composé de deux éléments principaux : un noyau et un nuage électronique. Le noyau est lui-même formé de protons et de neutrons, respectivement porteurs de charges positives et neutres. Ils vont constituer 99% de la masse de l'atome autour duquel les électrons, particules de nature électronégative, vont graviter dans des orbitales dites électroniques. Schématiquement, un atome est constamment représenté de manière sphérique. Mais la forme du nuage électronique peut fortement varier en fonction de divers éléments tels que la nature de l'atome, le nombre et la nature des liaisons covalentes qu'il effectue, la nature des atomes liés de manière covalente ou encore de la polarisation de l'environnement. Les propriétés du nuage électronique et l'équilibre

spatial des charges protons/électrons autour du noyau sont essentiellement responsables des interactions chimiques.

Deux grands types d'interactions chimiques sont distinguables : la liaison covalente et les interactions moléculaires.

a. Liaisons covalentes

La première repose sur le partage d'électrons entre deux atomes. Ainsi, deux atomes de carbone à une distance relativement proche (inférieure à 2Å) peuvent chacun mettre en commun un électron issu de leur valence externe et former une liaison dite covalente (voir Figure 10). Ce type de liaison permet de maintenir l'intégrité moléculaire en liant « physiquement » les atomes entre eux. Elle se produit essentiellement sur des atomes d'électronégativité relativement proche. L'énergie estimée de ce type d'interaction dépend des éléments et du nombre d'électrons impliqués, plusieurs paires d'électrons peuvent être partagées donnant lieu à des ordres de liaisons supérieurs (doubles ou triples liaisons). L'estimation énergétique d'une liaison simple entre deux atomes de carbones est de l'ordre de 80 kcal/mol tandis qu'une liaison double est proche de 140 kcal/mol.

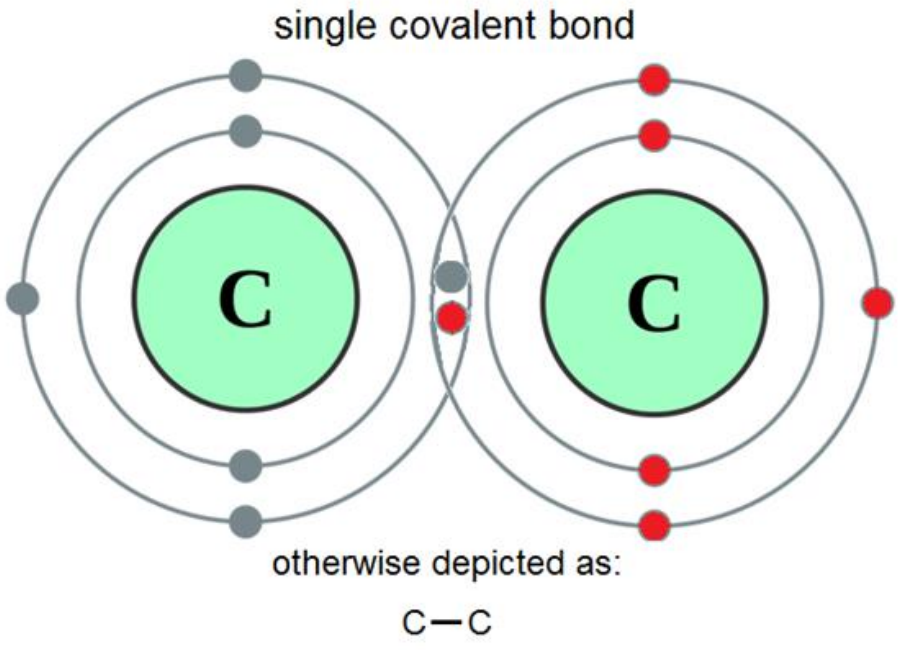


Figure 10 Description schématique d'une liaison simple covalente entre deux atomes de carbone partageant un électron de leur couche externe respective (image provenant de https://antisensescienceblog.wordpress.com).

b. Interactions moléculaires

Les travaux ici présentés portent essentiellement sur ce second grand type d'interaction, les interactions moléculaires ou interactions non covalentes. Elles se font généralement par l'intermédiaire de mécanismes électrostatiques ; elles impliquent des forces attractives et répulsives induites par les charges portées par les électrons et protons. Elles engagent communément deux atomes (ou deux groupements chimiques) présentant des propriétés électrostatiques concordantes, par exemple, un moment polaire positif et un moment polaire négatif. Ces interactions intermoléculaires sont beaucoup plus faibles énergétiquement que les liaisons covalentes et sont transitoires dans le temps car facilement dissociables.

Traditionnellement, seules les interactions attractives sont considérées dans la grande majorité des études. Ces interactions nécessitent deux éléments dont les charges par nature opposées, classiquement un électron et un proton, vont se faire face créant ainsi cette force attractive.

Elles sont impliquées dans de nombreux mécanismes biologiques. En effet, la formation des structures secondaires et de manière plus globale le repliement tertiaire est largement influencé par la présence d'interactions moléculaires intra-protéine. De plus, ces interactions vont régir les différentes interactions macromoléculaires entre les divers éléments retrouvés dans une cellule, telles les interactions protéine - ligand, protéine-protéine, protéine-ADN.

Cependant, les interactions impliquant deux moments polaires de nature identiques, par exemple des moments positifs, sont régulièrement négligées dans la littérature. Ces interactions, dites défavorables, sont pourtant présentes dans le contexte biologique mais peu étudiées.

c. Approximation énergétique

La liaison entre deux macroéléments tels une protéine et son ligand peut être caractérisée par son énergie thermodynamique afin de quantifier sa force générale. Cette énergie dite libre, correspond à la quantité d'énergie « produite » lors de l'interaction protéine - ligand, ou d'une autre manière comme la quantité d'énergie nécessaire pour rompre l'interaction. Cette énergie libre dépend de deux composantes principales, l'enthalpie et l'entropie, selon la formule :

$$\Delta G = \Delta H - T\Delta S$$

avec ΔG l'énergie libre de la réaction, ΔH la composante enthalpique du système, T la température et ΔS la composante entropique du système.

La composante entropique désigne l'impact des forces extérieures du système ainsi que son degré de désorganisation, tandis que la composante enthalpique correspond dans notre cas, à l'ensemble des forces attractives aboutissant à l'interaction protéine - ligand. Différents types de profils thermodynamiques peuvent être obtenus en fonction de la nature et de l'impact de chaque composante. Ainsi, des complexes protéine - ligand seront engendrés dans certains cas par une dominante enthalpique tandis que dans d'autres cas, la composante entropique jouera un rôle prépondérant.

Bien qu'à première vue, cette équation parait relativement simple, la réalité est tout autre. Le calcul de ces deux composantes nécessite, entre autres, la contribution du solvant et du soluté et sont difficilement approximables dans des structures cristallographiques. De plus, le phénomène de coopérativité rend la composante enthalpique non-linéaire : l'énergie libre d'un complexe protéine - ligand est plus importante que la somme des énergies attractives de ses interactions.

De nombreuses méthodes permettent l'approximation de l'énergie libre d'interaction. Les approches les plus précises, faisant appel à l'utilisation de champ de force (voir Annexe 2), telles que FEP (*free energy perturbation*) donnent des résultats convaincants dans la comparaison aux données expérimentales [24]. D'autres approches utilisant mécaniques quantiques et moléculaires (QM/MM) ont aussi été utilisés [25]. Cependant, leur utilisation nécessite des ressources computationnelles importantes et les valeurs prédites restent approximatives [26].

Des méthodes plus simples sont employées pour des contraintes de rapidité dans les logiciels d'assemblage (*docking* en anglais). Un exemple d'intérêt est *GoldScore* développé par Jones et collaborateurs [27], dont l'approximation est effectuée par la somme de l'ensemble des énergies de chaque interaction moléculaire dans le complexe protéine - ligand. *ChemScore* [27] implémenté dans le logiciel GOLD utilise une méthode empirique en additionnant la contribution de chaque élément thermodynamique comme une composante positive (liaisons hydrogènes, lipophilique, …) ou pénalisante (clashs, liaisons covalentes, …).

Cependant, ce type d'approches repose sur des observations parcellaires favorisant certains types d'interactions. Ainsi, chaque méthode est généralement performante sur une famille

protéique spécifique mais difficilement extrapolable sur l'ensemble du protéome. De plus, la relation liant les différentes composantes dans le calcul énergétique n'est pas linéaire mais bien plus complexe et ne peut être résumée à une somme d'interactions. Les approximations effectuées par ce type d'approche sont donc assez loin de la réalité.

Afin de faciliter la visualisation et compréhension de ce mécanisme complexe, les interactions moléculaires sont généralement classées en fonction de différents critères comme la nature des éléments impliqués ainsi que leurs arrangements spatiaux géométriques. Les différents types d'interaction vont être décrits par la suite telles que les liaisons hydrogènes, les interactions aromatiques, les liaisons halogènes, les contacts hydrophobes ainsi que les interactions *lone pair* – π.

B. Liaisons hydrogènes

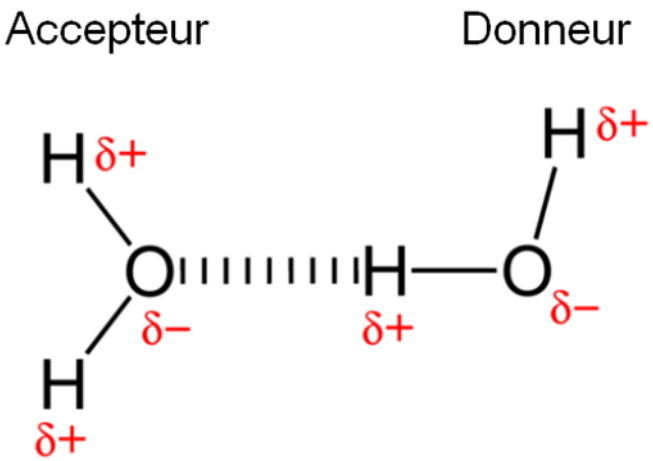


Figure 11 Schéma d'une liaison hydrogène entre deux molécules d'eau.

L'interaction électrostatique la plus fréquemment observée est la liaison hydrogène. Deux atomes spécifiques sont requis pour générer ce type d'interaction moléculaire : un accepteur et un donneur de liaison hydrogène, la Figure 11 décrivant ces deux acteurs. L'accepteur est un atome disposant d'une charge partielle électronégative, ici le doublet non liant de l'oxygène sur une molécule d'eau. Le donneur de liaison hydrogène est communément un hydrogène polaire lié de manière covalent à un atome électronégatif, souvent un atome d'azote ou d'oxygène, créant ainsi un moment dipolaire important, ici l'hydrogène de la molécule d'eau.

L'énergie engendrée par ce type d'interaction varie entre 1 et 10 kcal/mol dans le contexte biologique (milieu aqueux) [28]. Cette valeur dépend de très nombreux paramètres comme la géométrie de l'interaction, de la nature de l'accepteur de liaison hydrogène, de l'environnement ...

A titre d'exemple, l'énergie d'interaction de deux molécules d'eau interagissant de manière optimale dans un solvant aqueux est estimée à 5 kcal/mol [29]. Les conditions optimales évoquées ici correspondent essentiellement à la géométrie observée entre les deux molécules d'eau. Ainsi, la distance optimale séparant l'hydrogène de la première molécule et l'oxygène de la seconde est de 1.9Å [30]. De même, l'hydrogène doit être dirigé spécifiquement vers le doublet non liant de l'oxygène qui doit réciproquement être dirigé vers la région la plus polarisée de l'hydrogène.

Des études de mécaniques quantiques simulant les différentes géométries d'une liaison hydrogène ont mises en évidence l'impact des variations géométriques sur l'énergie libre d'interactions. Une variation d'angle entre le donneur de liaison hydrogène et l'accepteur de 20° peut aboutir à une variation énergétique de l'ordre de 0,5 kcal/mol [31, 32].

Toutefois, comme évoqué plus haut, l'énergie de cette interaction est variable. De nombreuses études évoquent les interactions hydrogènes comme étant soit faibles, soit fortes en fonction de leur estimation énergétique [33, 34]. Les liaisons hydrogènes fortes correspondent à des interactions dont l'énergie est proche de la valeur optimale. Les liaisons hydrogènes faibles sont la résultante d'une géométrie imparfaite, d'un accepteur de liaison faible comme le soufre, un cycle aromatique ou de la polarisation modérée du donneur de liaison hydrogène.

Bien que son énergie d'interaction soit relativement faible, la liaison hydrogène est décisive dans de nombreux processus biologiques. Elle joue un rôle essentiel dans la reconnaissance de ligand par son récepteur par exemple. Elle est aussi largement impliquée dans la formation des structures secondaires, par exemple les hélices α nécessitent la présence de liaisons hydrogènes au sein de la chaine principale, entre les résidus *i* et *i*+4.

Ce type d'interaction ne doit pas être vue au cas par cas, mais dans son ensemble, ces liaisons peuvent former des réseaux complexes garantissant la stabilité de certains éléments. L'existence de l'eau sous forme de continuum est par exemple du aux interactions multiples réalisés par chaque molécule d'eau. Une molécule d'eau peut être impliquée dans 4 liaisons

hydrogènes à la fois : deux en tant qu'accepteur pour l'oxygène et deux en tant que donneur pour chaque hydrogène, représentés sur la Figure 12. Ces 4 interactions vont former un réseau aqueux.

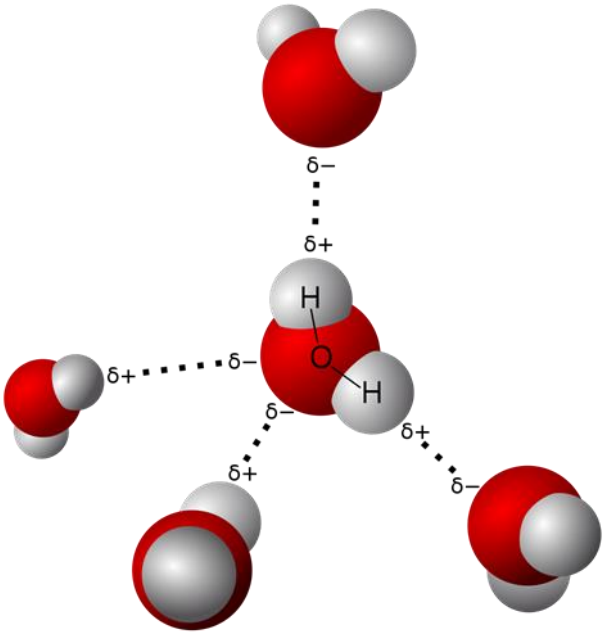


Figure 12 Représentation 3D d'un réseau de liaisons hydrogènes autour d'une molécule d'eau dans un milieu aqueux.

C. Interactions aromatiques

Les interactions aromatiques sont des interactions impliquant deux cycles décrits comme aromatiques. Un groupement aromatique, selon la définition d'Hückel [35], correspond à un composé cyclique plan possédant $4n+2$ électrons délocalisés, n représentant n'importe quel nombre entier. Les acides aminés tryptophane, phénylalanine et tyrosine disposent par exemple d'un groupement aromatique sur leur chaine latérale. L'implication de l'histidine est, quant à elle, plus controversée. La particularité de ces sous-structures repose sur l'agencement particulier des électrons délocalisés, équitablement répartis le long de l'alternance des liaisons simples et doubles, illustré sur la Figure 13. Ce phénomène a pour conséquence la création d'un dipôle spécifique autour de l'aromatique. Deux régions électronégatives, appelées régions π, sont situées de part et d'autre du plan de l'aromatique et sont centrées sur son centre de masse. Le pourtour de l'aromatique sera essentiellement chargé positivement par la faible densité du nuage électronique.

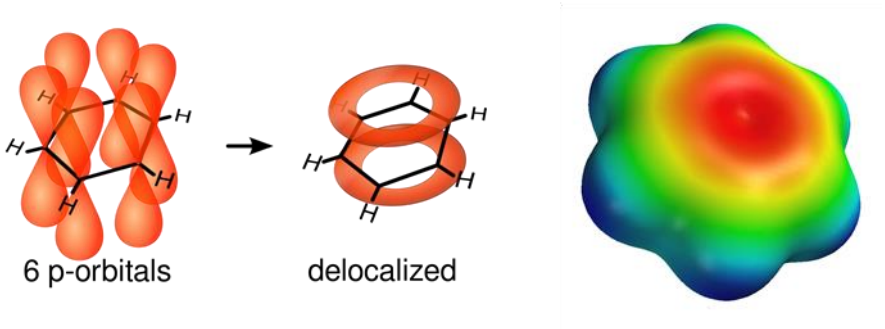


Figure 13 Schéma et illustration 3D de la distribution des électrons sur un groupement benzène.

L'intensité électrostatique de cette région π dépend de la nature et du nombre d'éléments composant l'aromatique ainsi que des substituants attachés à cette aromatique. Des groupements électroattracteurs tels que les halogènes fluor ou un groupement de type nitro -NO_2 inversent la polarité d'un cycle aromatique [36]. Deux régions déficientes en électrons se retrouvent ainsi situées de part et d'autre du plan de l'aromatique, nommées *π-hole*, tandis que les substituants seront fortement riches en électrons (voir Figure 14).

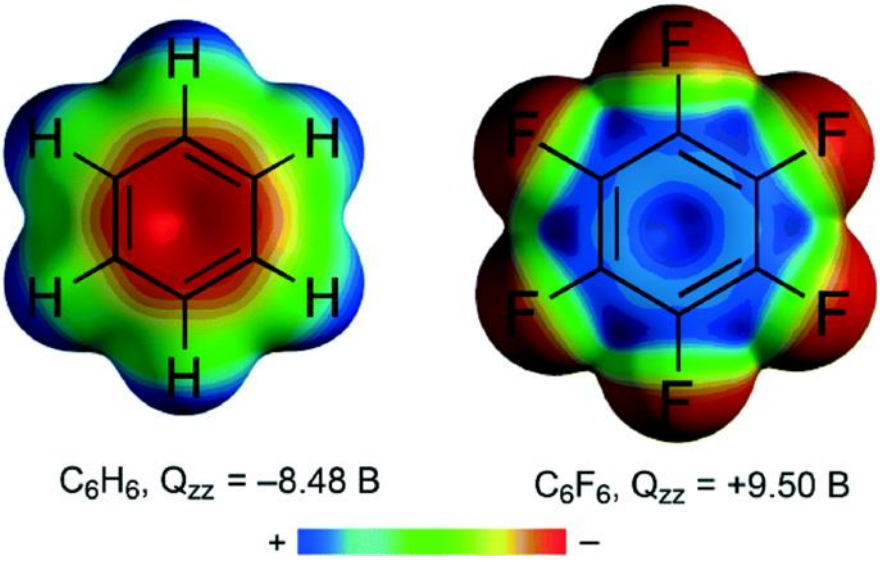


Figure 14 Surface de potentielle électrostatique du benzene et de l'hexafluorobenzene mettant en avant la distribution distincte des charges sur l'aromatique [37].

Plusieurs groupements fonctionnels peuvent ainsi interagir avec les aromatiques à commencer par un autre cycle aromatique. Ce phénomène est plus communément appelé *π-stacking* ou interaction aromatique-aromatique. La Figure 15 illustre certains arrangements spatiaux entre deux aromatiques, les plus favorables étant les conformations *T-shape* et

parallèle-décalé par la correspondance des régions électrostatiques se faisant face. Les critères géométriques utilisés pour ce type d'interactions sont généralement la distance séparant les deux centres de masses ainsi que l'orientation et le décalage entre ces deux centres. L'énergie observée pour ces deux conformations est généralement équivalente à 2 à 3 kcal/mol. Certains arrangements sont moins fréquemment observés mais soulèvent des questions quant à leur existence en tant qu'interaction. Deux aromatiques parfaitement parallèles sans décalage, conformation dite *sandwich*, devrait être théoriquement répulsif par le biais des deux régions riches en électrons se faisant face par exemple.

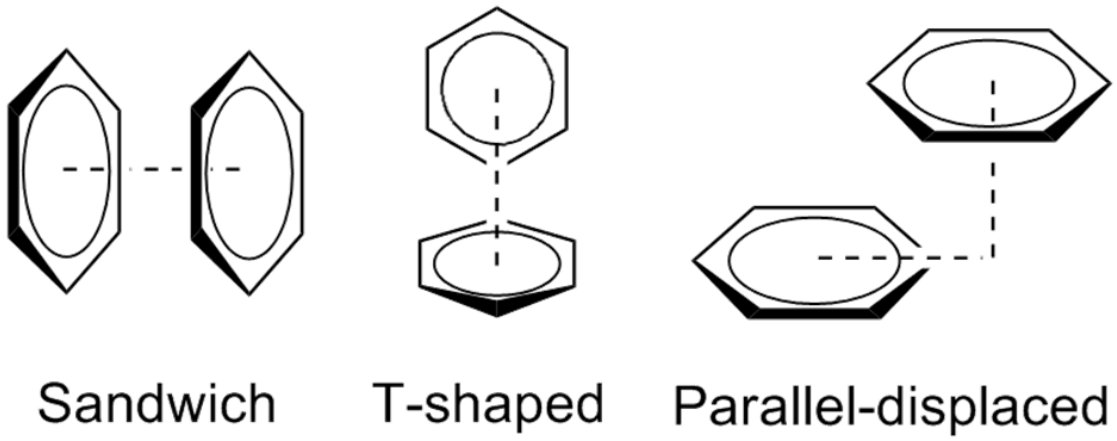


Figure 15 Illustration des arrangements conformationnels entre deux aromatiques.

D'autres éléments peuvent ainsi interagir avec un composé aromatique. En réalité, tout élément possédant une charge, partielle ou non, positive peut ainsi interagir avec une des régions π, par exemple un radical -OH formant une liaison hydrogène faible, ou un cation formant une interaction polaire. De même, l'absence d'électrons sur les parties latérales des aromatiques permet l'interaction avec des éléments électronégatifs tel l'azote sur le groupement indole du tryptophane.

D. Liaisons halogènes

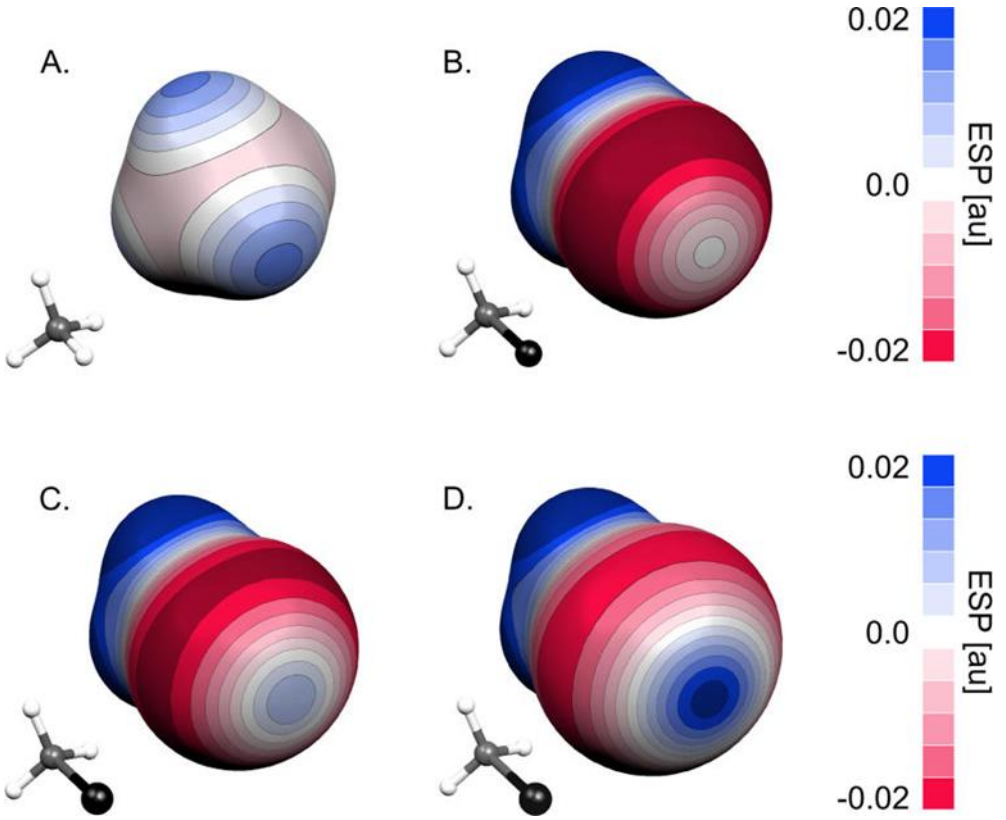


Figure 16 Représentations 3D de la surface potentiel électrostatique de différents groupements (ESP) du A. Méthane B. Chlorométhyle C. Bromométhyle D. Iodométhyle. Figure issue de Kolar et collaborateurs [38].

Les liaisons halogènes sont des interactions impliquant des atomes d'halogènes dit lourds. Les halogènes représentent les éléments appartenant au 17ème groupe de la table périodique que sont le fluor (F), le chlore (Cl), le brome (Br) et l'iode (I). Les interactions impliquant les halogènes ont fait récemment l'objet de réflexions avancées notamment par le biais d'études quantiques [38]. Mises en lumière pour la première fois en 1954 par Mulliken [39], leur intérêt en matière de *drug design* ne s'est illustré que depuis une dizaine d'années.

Les halogènes lourds, chlore, brome et iode, sont des atomes comportant une répartition d'électrons très spécifiques. Ces atomes sont dits anisotropes ; en fonction de la région observée, les propriétés électroniques de ces atomes vont être différentes (voir Figure 16). Ces atomes sont systématiquement liés de manière covalente à un seul atome, et dans le prolongement de cette liaison sur l'halogène se trouve une zone électropositive nommée la région σ. L'existence de cette région est due à une distribution anisotrope des électrons, situés de manière orthogonale autour de l'halogène par rapport à la liaison covalente. Ce

phénomène n'ayant été décrit à l'heure actuelle que chez ces 3 halogènes, le fluor (F) étant considéré comme trop électronégatif pour présenter une région σ.

La région σ interagit avec des bases de Lewis qui peuvent être des accepteurs de liaisons hydrogènes ou autres atomes présentant un moment électronégatif. Ce type d'interaction est plus communément appelé liaisons halogènes. Là encore, les paramètres géométriques ont une importance considérable. Energétiquement, une liaison halogène est équivalente à une liaison hydrogène entre deux molécules d'eau dans sa géométrie optimale. Cette dernière correspond à une distance égale à la somme des rayons de van der Waals et l'ensemble des atomes sont dans une configuration linéaire avec la région σ (voir Figure 17). Une augmentation de distance de 1,0Å contribuerait à une diminution de moitié de l'énergie d'interaction. Une diminution semblable est observée dans le cas d'une variation de l'angle sigma de 25 à 30° par rapport à un angle linéaire. De même, la liaison halogène est plus forte en présence d'un iode que d'un chlore de par l'intensité du *σ-hole* (4 kcal/mol à 1,6 kcal/mol).

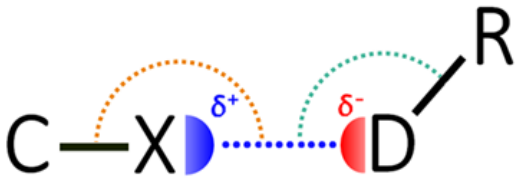


Figure 17 Schéma d'une liaison halogène entre un donneur de liaison halogène (D) et un halogène (X).

E. Contacts hydrophobes

L'effet hydrophobe joue un rôle prépondérant lors du mécanisme de repliement spécifique des protéines par l'intermédiaire des groupements hydrophobes présents à la surface de ces dernières [40]. Il joue aussi un rôle important dans le mécanisme d'interaction protéine - ligand, puisque la conception médicamenteuse requiert la présence de groupements hydrophobes afin que la molécule soit soluble dans le sang ou le cytoplasme [41]. La non-exposition au solvant de ces groupements contribue favorablement à l'énergie libre du système, à la fois du côté protéine mais aussi du ligand. En effet, l'exposition à la surface de ces groupements induit un coût entropique associé à leur organisation par rapport au solvant. L'enfouissement de ces groupements réduit alors considérablement ce coût entropique, favorisant l'énergie libre du système. Tout groupement est généralement considéré comme

hydrophobe si celui-ci est apolaire et/ou alkyle. Ainsi, le groupement méthyle CH_3 est le groupement hydrophobe le plus couramment rencontré dans les biomolécules.

Plusieurs études ont mis en avant la corrélation entre l'affinité du ligand et la surface hydrophobe du récepteur recouverte par le ligand. D'autres études ont empiriquement montré que le volume d'une poche hydrophobe doit être rempli à 55% par le ligand. Il est cependant ardu d'approximer le gain en énergie libre d'une désolvatation par la présence d'un groupement hydrophobe comparé à un atome polaire.

F. Rôle de l'eau

Les molécules d'eau, au-delà de leur contribution dans les phénomènes hydrophobes, ont un autre rôle dans le mécanisme d'interaction. Elle peut interagir en tant qu'accepteur ou donneur de liaisons hydrogènes en fonction de son orientation par rapport au ligand et à la protéine et servir ainsi d'intermédiaire au sein d'une interaction.

Ces « ponts aqueux » sont formés par la présence d'une molécule d'eau entre un atome d'une macromolécule et un atome provenant d'une macromolécule distincte, à distance d'interaction entre ces deux macromolécules. La Figure 18 montre un exemple précis d'une interaction médiée par une molécule d'eau l'interaction entre l'inhibiteur *Borolog2* et une α-thrombine. En l'absence de molécule d'eau intermédiaire, les deux oxygènes seraient en interaction défavorable. Ainsi, la dualité interactive, accepteur et donneur de liaisons hydrogènes, de la molécule d'eau va contribuer à la stabilisation de l'interaction protéine - ligand. Cependant, cette contribution est difficile à quantifier de par l'aspect interactif, mais aussi entropique lié à l'immobilisation d'une molécule d'eau entre deux solutés. L'optimisation des inhibiteurs de facteur anticoagulant Xa représente un cas concret où la substitution d'une molécule d'eau intermédiaire par un atome du ligand (chlore) a des effets bénéfiques, diminuant l'IC_{50} d'un facteur 200 [42]. Dans d'autres cas, concevoir un ligand qui déplacerait une molécule d'eau du site de liaison peut provoquer une perturbation entropique trop importante et déstabiliser l'interaction entre le ligand et son récepteur.

De nombreuses études tentent de déterminer si des molécules d'eau spécifiques font partie intégrante du récepteur, couplées à la protéine à travers plusieurs structures mais aussi plusieurs ligands [43]. Ces molécules d'eau peuvent avoir un rôle prépondérant dans la conservation structurale du site de liaison. Leur identification permet d'adapter et optimiser

la conception de molécules actives afin de ne pas déstabiliser l'intégrité structurale de la protéine.

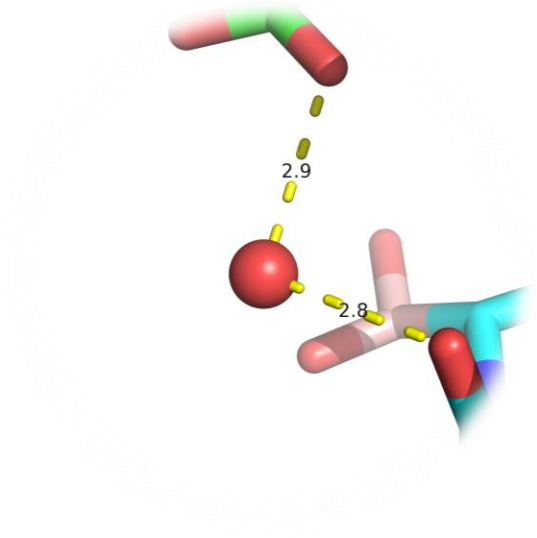


Figure 18 Illustration 3D d'une interaction impliquant une molécule d'eau intermédiaire entre l'inhibiteur Borolog2 et la Glutamine 192 d'une α-thrombine (code PDB 1a3e).

G. Doublet non liant - π

Les interactions décrites comme doublet non liant – π peuvent prêter à confusion. En effet, la dénomination π est communément employée pour décrire des cycles aromatiques ou système π, et leur moment électronégatifs correspondants. Or dans ce cas, π désigne en réalité des *π-holes*, déficient en électrons (voir Partie 1.2.C) [44]. Le partenaire d'interaction de ces *π-holes* déficient en électrons est généralement un doublet non liant, en anglais *lone pair* (lp), présent notamment sur l'oxygène et l'azote. Ces interactions, pouvant impliquées des anions, sont très fréquentes dans les interactions protéines - ADN ou impliquant des cofacteurs telle la flavine. De nombreuses études notamment statistiques ont été menées ces dernières années pour mettre en avant l'existence de ce type d'interaction [45]. Certaines suggèrent une interaction proche de la covalence entre un anion et un *π-hole* en termes d'énergie [46]. La géométrie requise pour cette interaction requiert un alignement linéaire entre la paire d'électrons libres et le système π électro-déficient. La Figure 19, extraite de [47], illustre différents cas d'interactions *lp* – π impliquant une molécule d'eau et des aromatiques de différentes compositions. L'importante différence énergétique entre les deux interactions

eau – imidazole, Figure 19A et C, est dû à la protonation de l'imidazole de la structure de l'amino-peptidase (code PDB 1c24) favorisant la déficience en électrons.

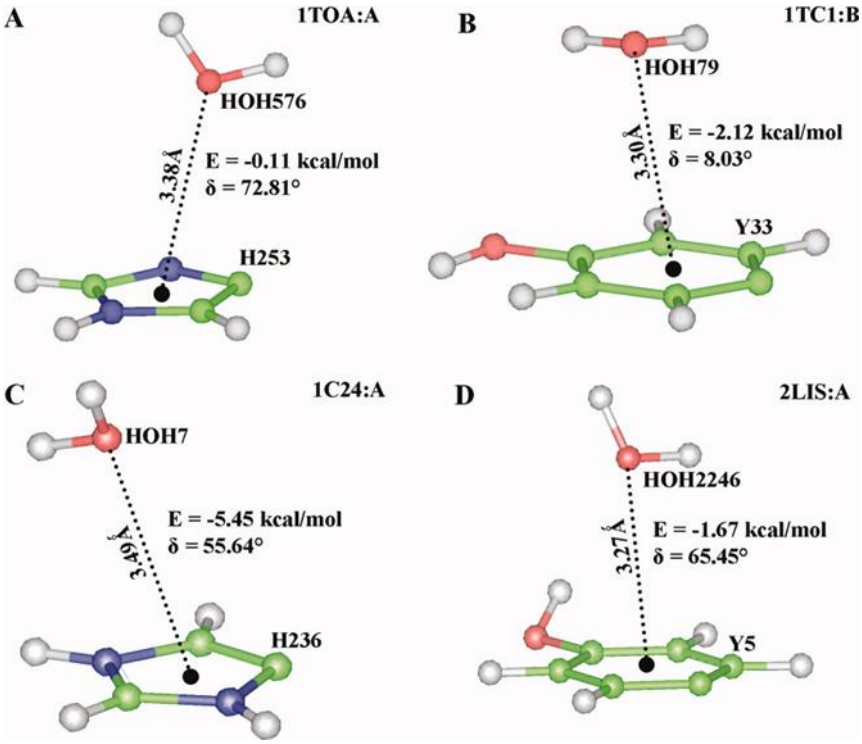


Figure 19 Exemples 3D de géométries favorisant les interactions lp - π impliquant le doublet non liant des molécules d'eau et des aromatiques contenant des groupements électroattracteurs [47].

H. Ponts salins et liaisons ioniques

Les ponts salins et liaisons ioniques correspondent à toute interaction impliquant une entité portant une charge positive ou négative. Les ponts salins peuvent être parfois assimilés à des liaisons hydrogènes dans le cas où un donneur de liaison hydrogène fait face à une entité portant une charge négative. Les liaisons ioniques sont généralement identifiées comme deux éléments portant des charges opposées en contact comme des groupements guanidine et carboxylate, comme illustré sur la Figure 20. La force de cette interaction est très variable et dépend à la fois des éléments mis en jeu mais aussi de la distance. Ainsi, à courte distance les valeurs énergétiques peuvent atteindre jusqu'à 100 kcal/mol et être assimilées à des liaisons covalentes [48]. Elles peuvent aussi être potentiellement observées à longue distance [49].

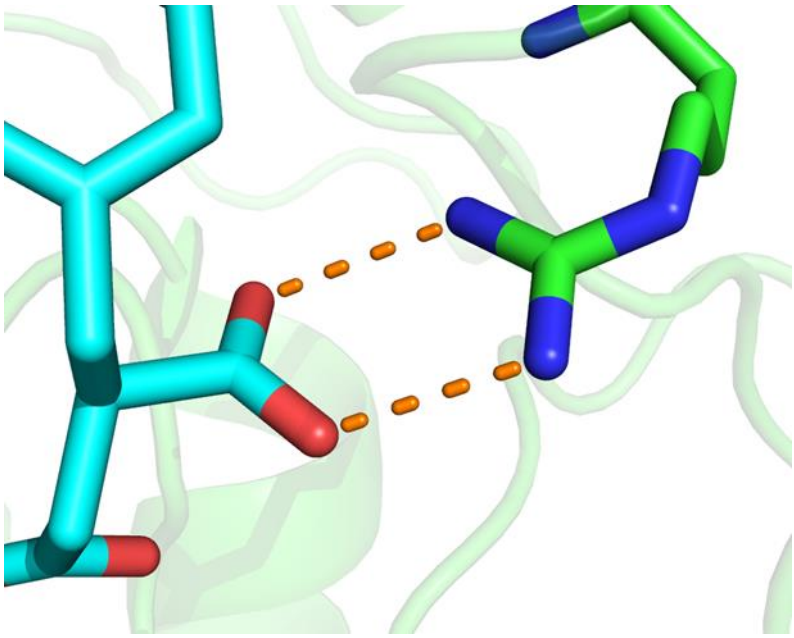


Figure 20 Interaction ionique entre un groupement guanidine et un groupement carboxylate dans un complexe carboxypeptidase et un acide benzylsuccinique (code PDB 1cbx, ligand BZS et l'arginine 145)

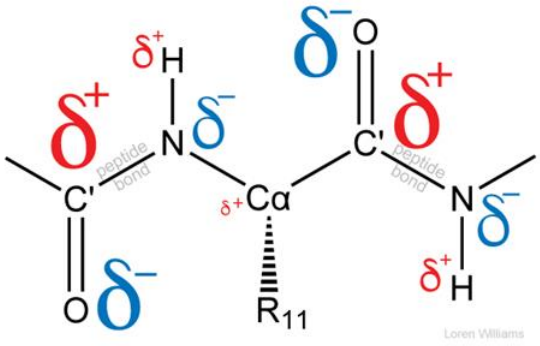
I. Interactions peu décrites

Les interactions décrites dans les parties précédentes correspondent toutes à des interactions acceptées et reconnues par la communauté scientifique. Elles ont régulièrement fait l'objet d'analyses statistiques et quantiques décrivant les différentes géométries et énergies correspondantes. Cependant, de nombreux éléments polaires ou potentiellement interactifs n'ont été que peu explorés jusqu'à présent et sont décrits dans ce chapitre.

Parmi les interactions attractives peu décrites, le rôle des halogènes en tant qu'accepteur de liaison hydrogène est souvent négligé. Une de leur spécificité réside dans leur répartition anisotrope des électrons. Le nuage d'électrons autour de la région σ peut aussi être considéré comme accepteur de liaison hydrogène. Cette hypothèse, explorée pour la première fois par Lin et collaborateurs en 2017, met en avant des énergies d'interaction entre 2 et 14 kcal/mol [50] et améliore l'affinité par un facteur 250 dans le cas de l'inhibiteur de polymérase ARN du virus de l'hépatite C [51]. Le fluor est aussi de manière inconsistante considéré comme un faible accepteur de liaison hydrogène.

D'autres éléments attractifs sont relativement peu explorés d'un point de vue interactif. Les interactions impliquant un groupement carbonyle sont généralement traitées comme des

interactions carbonyle – carbonyle [52]. Or, le carbone du squelette peptidique (annoté C' sur la Figure 21) présente un moment dipolaire important pouvant interagir avec n'importe quel moment positif. La délocalisation importante des électrons au niveau de l'oxygène de l'azote contribue à la présence d'une charge partielle électropositive pouvant interagir avec une paire d'électrons libres ou accepteurs de liaison hydrogène. Cette interaction, nommée interaction dipolaire orthogonale dans certaines études [53, 54], a été étudié notamment par Fischer et collaborateurs en 2008 au cours duquel une énergie d'interaction proche d'un π-*stacking* a été mesurée [55].



-- Partial Charges on the Atoms of a Peptide --	
Atom	partial charge (e^-)
N	-0.42
HN	+0.27
Cα	+0.12
C'	+0.60
O	-0.57

Figure 21 Illustration du moment dipolaire présent sur le squelette peptidique induisant une charge partielle positive du carbone du groupement carboxyle. Les valeurs de charge partielle correspondantes sont décrites dans le tableau associé.

Le sulfure est aussi un élément attracteur pouvant adopter plusieurs rôles en termes d'interactions. Tout d'abord, au sein même de la structure protéique, il va permettre la création de ponts disulfures covalents entre deux cystéines et sera alors, en théorie, incapable d'interagir avec d'autres atomes. Cependant, il est considéré aussi comme une base de Lewis et est donc capable de jouer le rôle d'accepteurs faibles de liaisons hydrogènes ou de liaisons halogènes. La plupart des études centrées sur cet atome se focalisent sur l'impact de l'ajout d'un sulfure sur la stabilité du ligand [56, 57]. En 2015, Zhang et collaborateurs analysent un autre type d'interaction potentielle impliquant ces atomes. Ils démontrent qu'une interaction sulfure-oxygène pouvait atteindre 6,2 kcal/mol (voir Figure 22). Cette étude bien entendu se limite à des cas très spécifique de sulfure : le thiophène, le thiazole et le thiadiazole [58]. Aucune étude exhaustive sur l'interactivité du sulfure dans le contexte protéine - ligand n'a été réalisée à ce jour.

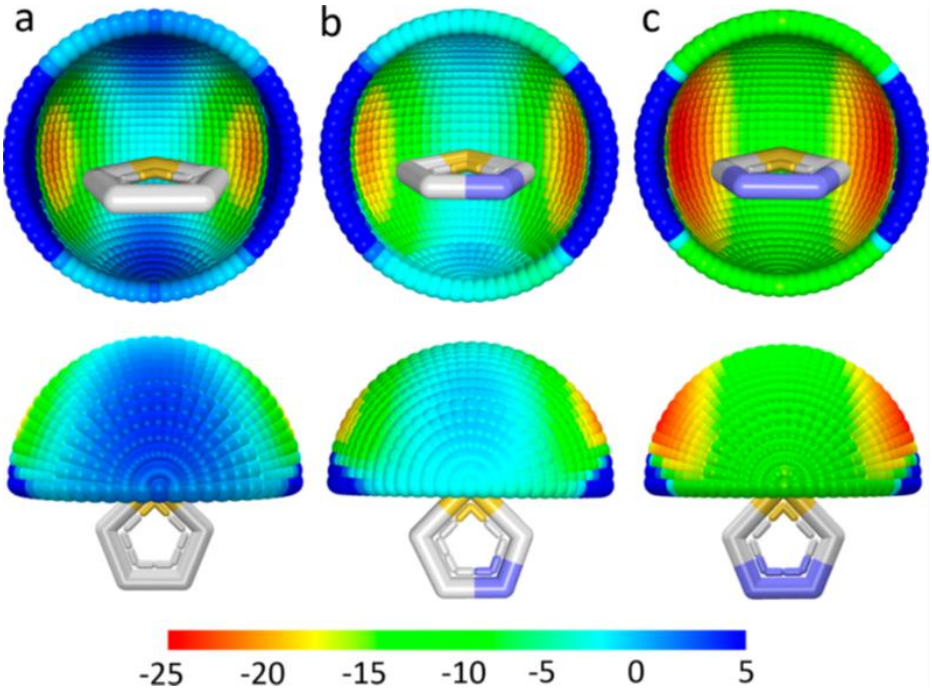


Figure 22 Représentation sphérique de l'énergie d'interaction en kJ/mol entre un doublet non liant d'un oxygène et le sulfure d'un A. thiophène B. thiazole C. thiadiazole [58].

Contrairement aux cycles aromatiques, les cycles aliphatiques ne sont pas regroupés en une entité interactive à part entière. Leurs interactions sont généralement considérées de la même manière qu'une chaine aliphatique, chaque élément du cycle interagira avec un élément correspondant se situant en face de lui. Il n'existe à l'heure actuelle aucune étude exhaustive sur les interactions impliquant des cycles aliphatiques.

Enfin, parmi ces interactions peu, voire pas répertoriées, les interactions dites défavorables ou répulsives sont rarement décrites dans la littérature. Deux moments électrostatiques de même nature, par exemple positifs ou deux doublets non liant se faisant face, vont généralement constituer ces répulsions électrostatiques. Il n'existe à l'heure actuelle pas d'études exhaustives dans ce domaine, ni même de quantifications énergétiques bien que leurs existences soient connues de tous, notamment dans le contexte protéine - ligand.

Chapitre 3 : Outils et applications pour l'étude des interactions

Ces interactions moléculaires, précédemment décrites, peuvent être identifiés sur des structures protéine - ligand 3D. Il existe à l'heure actuelle différents logiciels et outils permettant la détection et caractérisation de ces interactions. Sont décrits dans ce chapitre les différentes approches employées, ainsi que les différents outils recensés à travers la littérature.

A. Mécanisme général de détection

La détection d'interactions protéine - ligand est généralement effectuée en deux étapes successives. La première repose sur la détection de paires d'atomes dont la distance les séparant est en dessous d'un seuil prédéterminé, souvent définis comme contacts atomistiques. Ce seuil varie selon les logiciels et études d'interactions moléculaires considérées. La deuxième étape consiste à caractériser les différents contacts établis entre atomes selon les deux éléments impliqués et leurs configurations géométriques respectives.

La première étape est commune à la quasi intégralité des méthodologies, le traitement de la seconde est particulièrement spécifique en fonction des définitions et seuils de distance sélectionnés. Ainsi, entre deux différentes méthodes de détection, certains vont considérer un angle halogène pouvant aller jusqu'à 150° pour des liaisons halogènes, d'autres vont tolérer un angle de 120°. Il en est de même pour la distance.

Un autre point de divergence important est notamment retrouvé dans la classification des interactions caractérisées. Selon les études et autres logiciels employés, cette dernière peut significativement varier et ainsi impacter la manière de décrire la relation protéine - ligand. Selon les études, il est courant de voir les liaisons hydrogènes regroupées et ou dissociées en deux sous-catégories : les liaisons hydrogènes faibles et les liaisons hydrogènes fortes en fonction des donneurs, accepteurs présents.

Néanmoins, la grande majorité des outils de détection et caractérisation utilisent des définitions et seuils de distance semblables proches des descriptions observées dans la littérature. A l'heure actuelle, il existe un nombre limité de logiciels accessibles proposant la

détection et caractérisation d'interactions protéine - ligand au sein d'une structure tridimensionnelle.

B. Notion fondamentale : le *fingerprint*

A travers cette thèse, la notion de *fingerprint* (empreinte digitale) reviendra régulièrement. Ce procédé consiste à concaténer l'information dans un vecteur à une dimension. Les *fingerprints* peuvent être binaires ou quantitatifs. Un *fingerprint* dit binaire contient des valeurs (ou *bits*) binaires, 1 ou 0, indiquant la présence ou absence d'une propriété à une position définie. Un *fingerprint* quantitatif correspond à un vecteur dont les valeurs représentent une quantification d'une certaine propriété. L'élaboration de *fingerprints* repose sur la connaissance à priori des éléments présents dans chaque *bit*, induisant une taille fixe pour chaque *fingerprint*. Un *fingerprint* moléculaire par exemple, tel ECFP (*Extended Connectivity FingerPrint*) [59], va décrire une molécule en identifiant toutes les sous-structures moléculaires possibles à différentes distances (exprimé en nombre de liaisons), comme illustré sur la Figure 23.

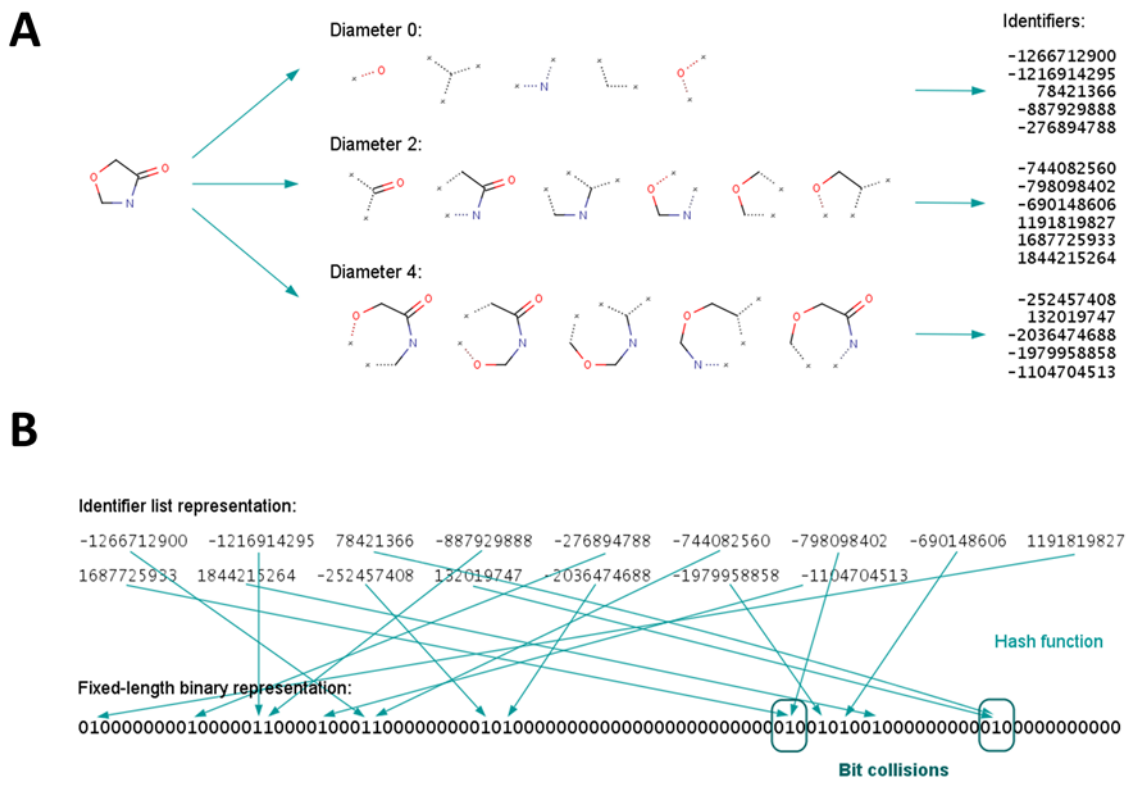


Figure 23 Génération d'un fingerprint moléculaire selon la méthode ECFP A. Identification des sous-structures à différentes distances de liaisons B. Création du fingerprint en assignant toutes les sous-structures recensées (images extraites de ChemAxon : https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP).

C. Base de données et visualisation

Plusieurs outils sont accessibles à la communauté scientifique permettant de (i) d'identifier les interactions moléculaires dans un complexe protéine - ligand, et de (ii) stocker dans une base de données

a. *LIGPLOT* et *LIGPLOT+*

LIGPLOT est un programme informatique publié en 1996 générant automatiquement des représentations sous forme de diagramme 2D des interactions protéine - ligand [60]. La détection des contacts à partir de fichiers PDB s'effectuait par le logiciel *HBPLUS* publié en 1994 [61] permettant la détection des liaisons hydrogènes et des contacts hydrophobes. L'algorithme implémenté dans *LIGPLOT* « écrasait » ensuite le ligand passant d'une représentation 3D à 2D et assignait ensuite les interactions. Sa version moderne, *LigPlot+* [62] publié en 2011, permet d'afficher et aligner sur un plan 2D plusieurs complexes protéine - ligand similaires. La Figure 24 illustre un exemple de contacts hydrophobes et liaisons hydrogènes identifiés pour un complexe impliquant la protéine HSP90α et le ligand 9-Butyl-8-(2-chloro-3,4,5-trimethoxy-benzyl)-9H-purin-6-ylamine.

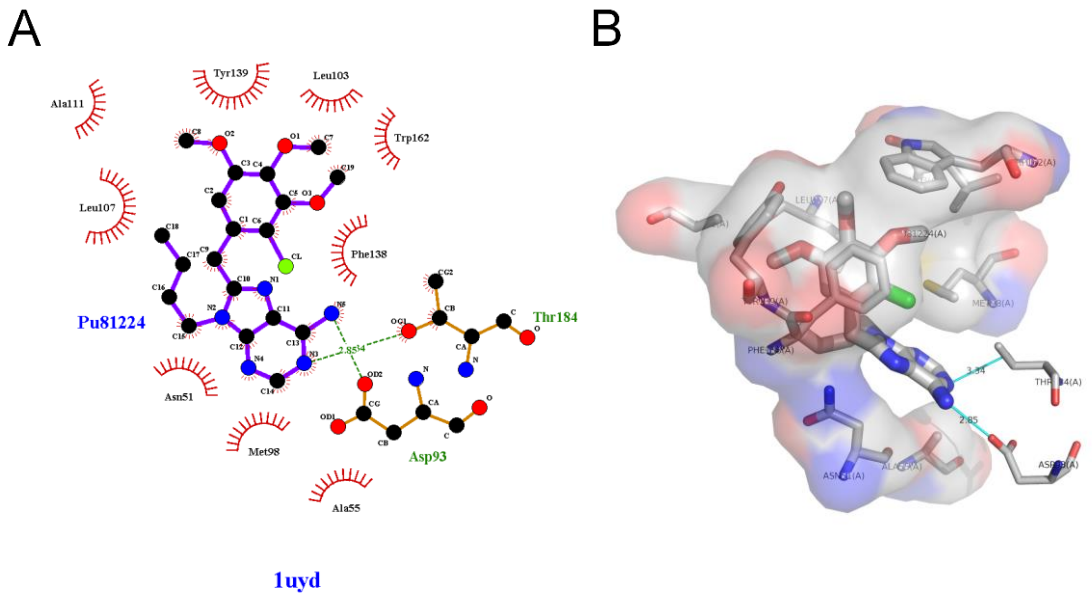


Figure 24 Résultats obtenus par LigPlot+ 1.4.5 pour la protéine HSP90α (code PDB 1uyd) A. Diagramme d'interaction hydrophobes (demi-cercle rouge) et liaisons hydrogènes (verts). Représentation 3D générée à partir de LigPlot+ et PyMOL.

b. *CREDO*

CREDO est une base de données relationnelle publiée en 2013 dans laquelle est stockée les structures PDB [63]. Les acides aminés de chaque structure sont annotés aux séquences UniProt auxquelles elles sont rattachées. De plus, une fragmentation moléculaire de chaque ligand a été effectuée. *CREDO* contient les interactions détectées différentes macromolécules (protéines, acides nucléiques, petites molécules, …) stockées et caractérisées sous format *SIFt* amélioré décrivant ainsi 13 types d'interactions distincts (description de *SIFt* dans la Partie 1.3.D.a). Cependant, *CREDO* ne semble plus être mis à jour depuis Juin 2014.

c. *Protein Ligand Interaction Profiler (PLIP)*

Dans une optique purement descriptive, en 2015 a été publié et mis à disposition gratuitement une interface web permettant la détection et caractérisation automatique au sein d'un complexe protéine - ligand de ces interactions. *PLIP*, pour *Protein-Ligand Interaction Profiler*, élaboré par Salentin et collaborateurs, détecte dans un premier temps les contacts par un critère de distance non précisée [64]. La description interactive des atomes, par exemple donneur de liaisons hydrogènes, est effectuée par la solution informatique *OpenBabel* (http://openbabel.org/).

Cependant, *OpenBabel* [65] ne considère pas un atome d'halogène comme étant un accepteur de liaisons hydrogènes. De même, le fluor est un donneur de liaison halogène contrairement aux présomptions émises dans la littérature et tous les atomes d'oxygène, azote et soufre sont considérés comme accepteur de liaison halogène. La caractérisation se divise en 7 interactions différentes au moment de sa publication : interactions hydrophobes, liaisons hydrogènes, le *stacking* aromatique, les interactions π-cations, les ponts salins, les interactions *water-mediated*, et enfin les liaisons halogènes. La visualisation des interactions peut se faire directement dans une visionneuse 3D sur le serveur web. Les résultats peuvent aussi être récupérés et téléchargés sous divers formats (fichiers .xml, image .png et session *PyMOL*). Cependant, les seuils de distance et d'angle utilisés lors de la détection et définition d'interactions donnent des résultats inconsistants pour certains complexes, notamment la tyrosine kinase 2 illustrée en Figure 25.

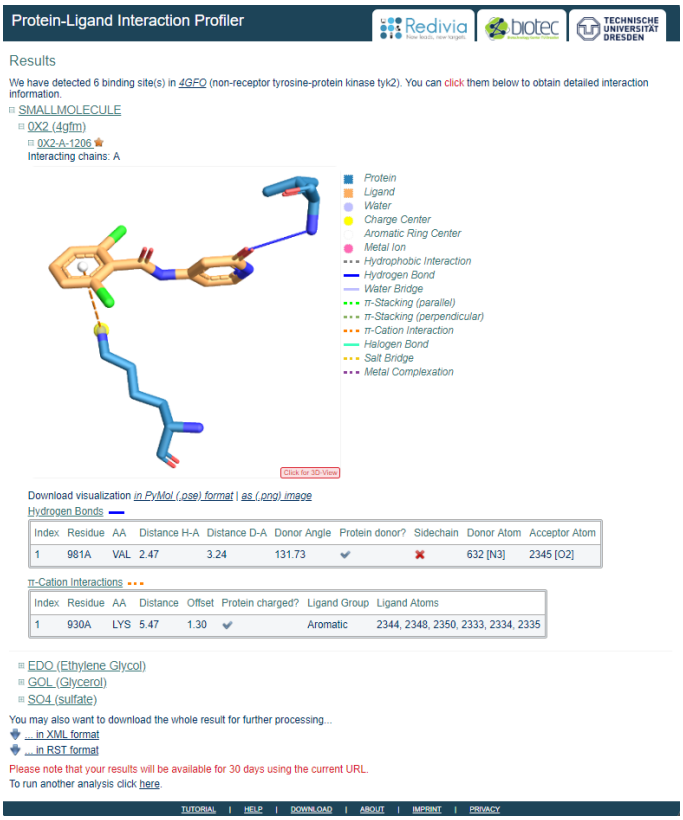


Figure 25 Résultats obtenus par le serveur web PLIP dans la détection d'interaction pour la protéine Tyrosine Kinase 2 (code PDB 4gfo).

d. Arpeggio

Arpeggio, publié en 2017 [66], est très semblable à *CREDO*, des mêmes auteurs, et *PLIP* mais a pour avantage de proposer un nombre plus important de types d'interaction notamment les coordinations métalliques et les interactions impliquant les aromatiques tel des interactions halogènes-π, ou encore carbone-π. Les résultats sont là uniquement visibles en 3D par l'intermédiaire d'une image ou téléchargeables sous forme de sessions *PyMOL*. Les paramètres géométriques pour déterminer les différentes interactions varient là encore, avec pour exemple une interaction halogène défini par un angle compris entre 70° et 170° degrés. Contrairement aux 2 interactions détectées par *PLIP*, *Arpeggio* détecte un nombre significativement plus important d'interactions pour la protéine tyrosine kinase 2 avec un total de 206 contacts dont 20 hydrophobes et 7 liaisons hydrogènes faibles, plus pertinents pour expliquer la liaison du ligand sur sa cible (voir Figure 26).

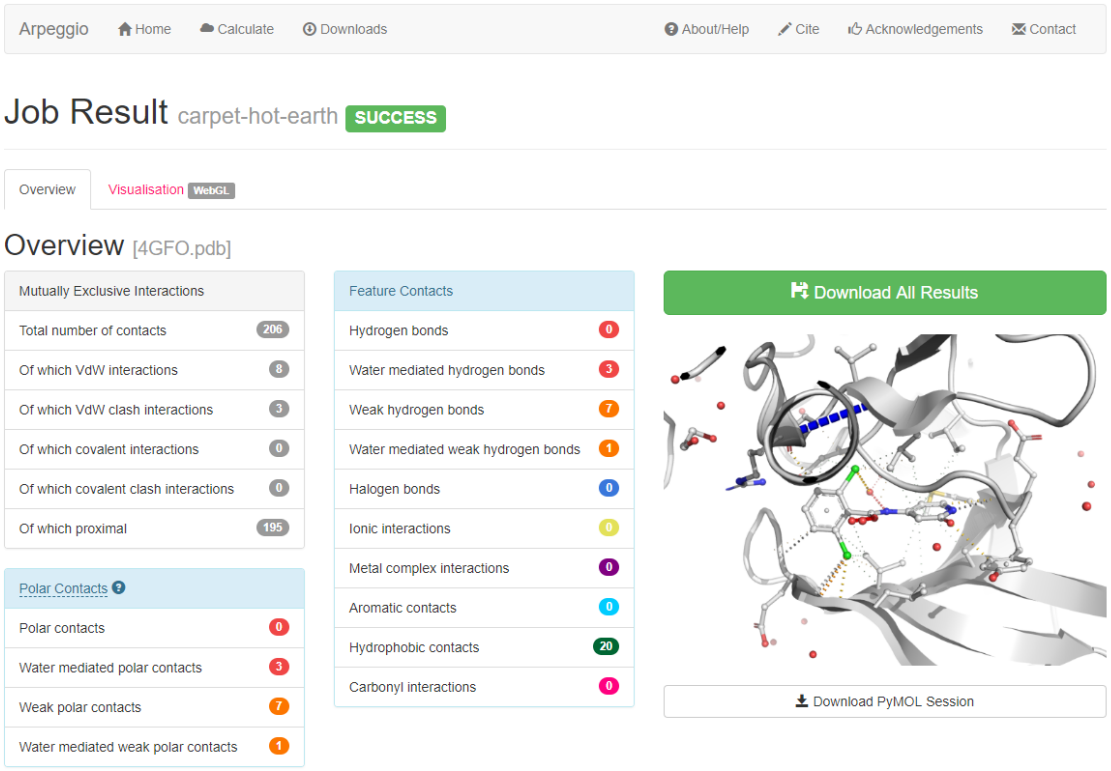


Figure 26 Descriptions quantitatives des interactions et contacts détectés par Arpeggio et représentation 3D des interactions dans le contexte protéine - ligand pour une Tyrosine Kinase 2 (code PDB id 4gfo).

D. Méthodes de descriptions de l'interaction

Les logiciels et plateformes décrits précédemment détectent et offre des outils de visualisation des interactions protéine - ligand. Cependant, aucun de ces outils ne permet la comparaison de *pattern* d'interaction. Différentes approches ont été mis en place ces dernières années et sont décrites à travers ce chapitre.

a. *Structural Interaction Fingerprint* (*SIFt*)

La méthode précurseur de la caractérisation des interactions est *SIFt* pour *Structural Interaction Fingerprint*, publié en 2004 par Deng et collaborateurs [67]. Pour caractériser un ensemble d'interactions entre un ligand et une protéine, cette approche repose sur deux étapes relativement simples. La première consiste à identifier les acides aminés voisins du ligand dans l'espace par détection des atomes protéiques en interaction avec une distance seuil de 4.5Å et mesure de l'accessibilité au solvant par le programme AREAIMOL [68] et HBPLUS [61].

Ensuite, chaque résidu est ensuite caractérisé selon (i) contact ou non avec le ligand, (ii) contact de la chaine principale, (iii) contact avec la chaine latérale, (iv) interaction polaire ou non, (v) interaction apolaire ou non, (vi) présence sur le résidu d'accepteur de liaison hydrogène, et (vii) présence de donneur de liaison hydrogène. Chaque résidu sera ainsi annoté sur un vecteur binaire de 7 bits. L'ensemble de ces vecteurs est concaténé selon l'ordre croissant de la numérotation acides aminés dans la séquence dans un *fingerprint*. La comparaison des *fingerprints* se fera par le rapport des éléments de chaque vecteur en commun sur le nombre total d'éléments de chaque vecteur, appelé coefficient de Tanimoto [69] :

$$Tanimoto\ (X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

avec X et Y les deux *fingerprints* issus de complexes distincts. *SIFt* permet notamment de comparer deux ligands différents complexés à des protéines très similaires, à condition que leurs *fingerprints* soient de taille identique (que le nombre de résidus impliqués soit le même).

De nombreux dérivés et évolutions ont ensuite été développés, dont les nouveautés consistent majoritairement à l'ajout de descripteurs dans la caractérisation du récepteur. *CREDO* [63], notamment, 8 descripteurs supplémentaires comme la distinction entre faibles et fortes liaisons hydrogènes, présence d'accepteur halogène ou encore d'aromatique. Ces approches permettent notamment l'analyse de résultats issus de *docking* ou la comparaison d'interactions dans des complexes impliquant des protéines très similaires. La Figure 27 illustre un cas concret de l'utilisation de ces *fingerprints* dans le regroupement hiérarchique de 89 complexes de protéines kinases.

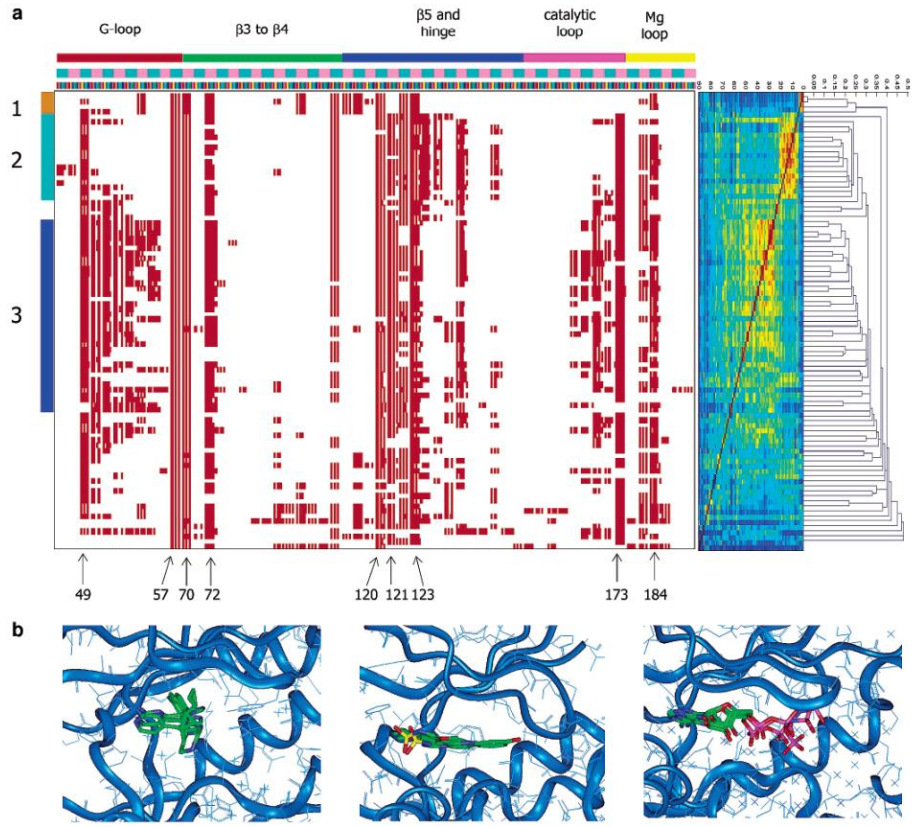


Figure 27 Exemple d'utilisation du fingerprint SIFt dans le cadre de protéines kinase A. Regroupement hiérarchique effectué sur les SIFt générés sur 89 protéines kinases, chaque ligne correspond à un fingerprint d'interaction. Les résidus fréquemment identifiés dans ces kinases sont représentés par des flèches. B. Représentations 3D des 3 groupes majeurs issus de ce regroupement hiérarchique, affiché à gauche du dendrogramme, soulignant des poses de ligands identiques entre différents complexes. Image extraite de l'article de Deng et collaborateurs [67].

b. Triplet de Fingerprint d'Interaction (TIFP) et *Graph Matching of IPA* (Grim)

Cette méthode, développée en 2013 par Desaphy et collaborateurs, repose sur des arrangements locaux d'interactions [70]. Les atomes sont dans un premier temps décrits sous 7 catégories : hydrophobe, aromatique, donneur ou accepteur de liaison hydrogène, métallique, composé ionisable positivement ou négativement. Puis 7 interactions sont caractérisées selon ces descriptions et des paramètres géométriques distincts (distance maximum pour une liaison hydrophobe de 4.5Å contre 3.5Å pour une liaison hydrogène, 1ère étape de la Figure 28). L'interaction est ensuite décrite par un pseudo-atome (IPA) définit soit au niveau du ligand, soit au niveau du récepteur ou soit au niveau du centre géométrique de l'interaction. Des triplets de pseudo-atomes auxquels sont assignés leurs distances respectives catégorisées (6 groupes) vont être comptabilisé (2ème étape de la Figure 28). 391 triplets fréquents (plus de 10 occurrences) sur 12 508 combinaisons possibles ont été conservés après

analyse sur 9 877 complexes protéines-ligands. Ces triplets sont regroupés dans un *fingerprint* final composé de 210 bins représentant des triplets récurrents (3ème étape de la Figure 28). La comparaison entre complexes se fait par similarité de *fingerprint*.

Un algorithme d'alignement d'interaction, Grim pour *Graph Matching of IPAs*, fut développé à partir des pseudo-atomes définis précédemment. Chaque complexe est défini par un graphe où chaque *IPA* est défini comme un nœud et tous les *IPA* de même nature, liaison hydrogène par exemple, sont liés entre eux par des arêtes. Le produit des deux graphes est par la suite calculé, donnant la liste de toutes les combinaisons possibles entre deux ensembles d'*IPA*s. L'algorithme de Bron−Kerbosch est ensuite utilisé pour détecter la plus grande clique, soit le sous-ensemble de sommets le plus grand où 2 nœuds sont toujours adjacents [71]. Chaque IPA d'un complexe référent est ensuite aligné à l'IPA correspondant d'un autre complexe par calcul d'une matrice de rotation et translation [72].

L'avantage de cette méthode repose sur la possibilité de comparer et aligner sur leur *pattern* d'interaction. Néanmoins, elle repose sur des interactions typées, telles les liaisons halogènes ou le *π-stacking* par exemple, et omet donc des interactions peu connues qui peuvent être importantes dans certains complexes [70].

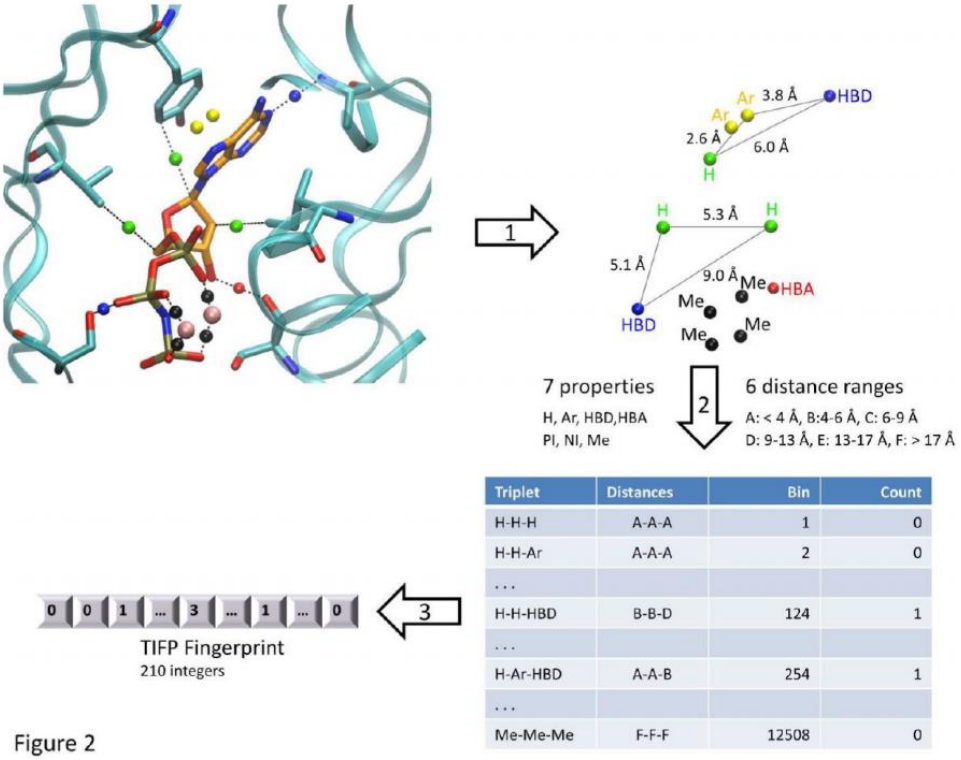


Figure 28 Illustration du processus de caractérisation de l'interaction [70]. 1. Création des triplets d'interactions 2. Recensement des triplets à travers 9 877 complexes. 3. Elaboration du fingerprint d'interaction final.

c. *Protein Ligand Interaction Explorer (PROLIX)*

PROLIX est une base de données développée en interne chez Hoffmann-La Roche recensant les interactions protéine - ligand [73]. L'outil, inaccessible au public, offre une interface graphique permettant à l'utilisateur de rechercher un complexe en sélectionnant un des résidus, des types d'interaction associés ainsi que des contraintes géométriques.

11 types d'interactions distinctes sont décrites ici, allant des plus récurrentes comme la liaison hydrogène ou le *π-stacking* mais aussi les interactions impliquant des cations de type cation-dipôle et cation-π. L'aspect innovant de cette base de données repose sur l'intégration des interactions dites non favorables, i.e. interactions électrostatiques répulsives, les clashs stériques ainsi que les interactions impliquant une forte pénalité de désolvatation. La caractérisation globale d'un complexe repose sur la somme de l'ensemble de ses interactions.

Les recherches dans la base de données se font en 3 étapes : (i) une recherche par *fingerprint* des résidus du site de liaison, (ii) une recherche par *fingerprint* des interactions moléculaires, et (iii) filtre jusqu'à un certain seuil de distance induit par l'utilisateur. Selon les options sélectionnées par l'utilisateur, des résultats partiels peuvent aussi être obtenus.

d. *Structural Protein Ligand Interaction Fingerprint (SPLIF)*

Da et Kireev ont mis au point en 2014 [74] une méthode décrivant de manière implicite les interactions moléculaires. La Figure 29 décrit les différentes étapes de cette approche. L'ensemble des contacts entre deux atomes avec un seuil de distance de 4,5Å est dans un premier temps détecté. Un *fingerprint* moléculaire (ECFP2) décrivant l'environnement local de chaque atome en contact est ensuite généré, décrivant un fragment local. Les coordonnées de chaque fragment sont conservées. Le *fingerprint* d'interaction final représente l'ensemble des couples de *fingerprints* moléculaires observé dans un complexe.

La comparaison de deux *fingerprints* d'interaction se fait ensuite en deux étapes. Dans un premier temps, l'identification de *fingerprints* moléculaires identiques entre les deux complexes est réalisée. Dans un second temps, l'alignement 3D des fragments locaux considérés comme identiques est effectué, et l'écart-type quadratique moyen (RMSD) permet d'évaluer la similarité des deux complexes. Une RMSD inférieur à 1,0Å catégorise l'interaction comme identique.

La description se fait ainsi de manière implicite, les interactions ne sont pas définies par des types prédéfinis tels que liaisons hydrogènes ou halogènes mais uniquement deux fragments moléculaires. Cependant, la comparaison entre fragments moléculaires nécessite une identité totale et non une certaine similarité, dissociant ainsi tous les accepteurs de liaisons hydrogènes entre eux par exemple.

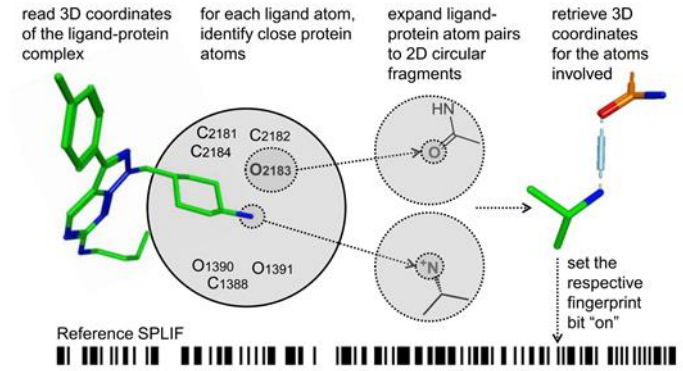


Figure 1. Essential steps of building a reference SPLIF.

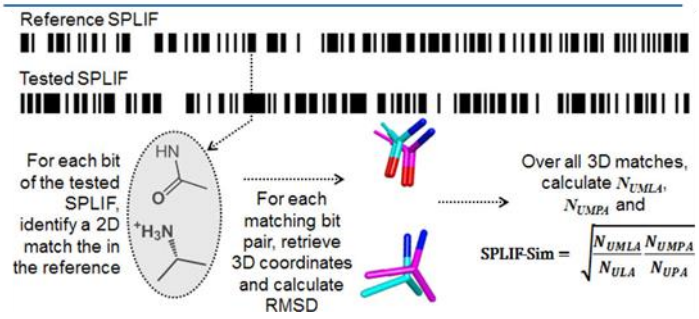
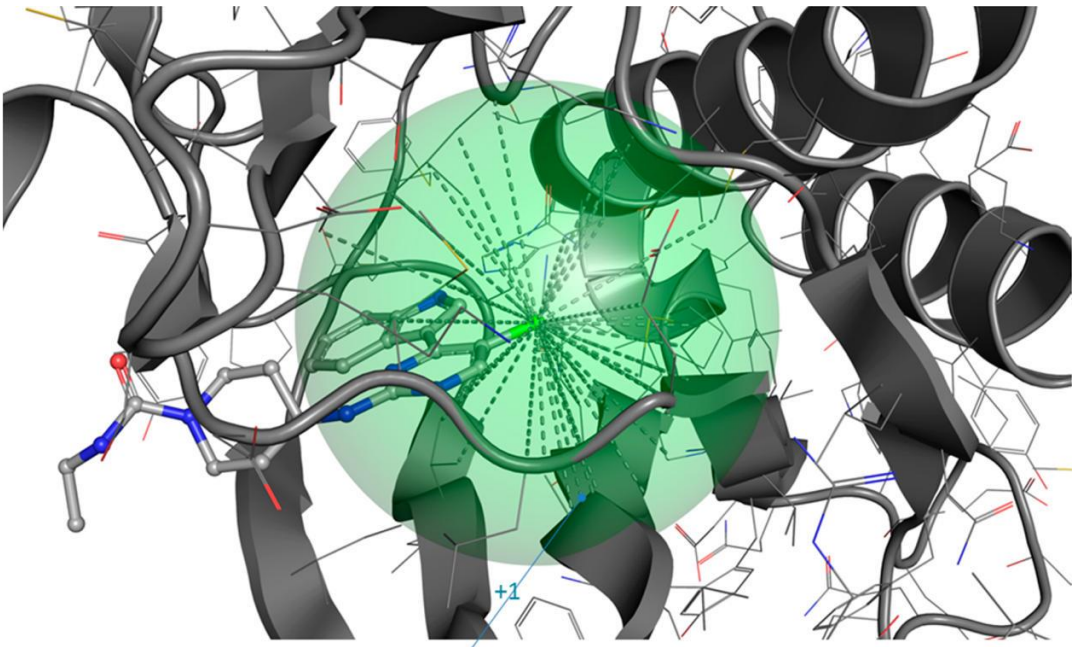


Figure 29 Description de la méthodologie SPLIF : détection des atomes de la protéine proche d'un atome du ligand, caractérisation de chaque atome par un fingerprint moléculaire (fragment local), stockage du couple de fingerprints moléculaires et des coordonnées 3D correspondantes. Enfin, comparaison entre deux complexes par l'identification de fragments moléculaires et de configuration géométrique identiques [74].

e. *Elements* et *Sybyl*

En 2014, Ballester et collaborateurs ont proposé plusieurs méthodes de description du complexe protéine - ligand en mettant en relation leur complexité respective et la prédiction de l'affinité [75]. La détection de l'ensemble des contacts se fait par l'intermédiaire d'un seuil de 6,0Å, sans considération de l'accessibilité ou d'angles comme illustré sur la Figure 30. Le *fingerprint* d'interaction dit *Elements* recense quantitativement l'ensemble des contacts observés en fonction des deux éléments mis en jeu. Ainsi, pour un même atome de chlore, un nombre important de contacts avec le carbone peut être détecté et insérer dans le *fingerprint*.

Le *fingerprint* d'interaction dit *Sybyl* est construit de manière identique à l'exception que chaque élément est décrit selon le schéma *Sybyl* [76]. Ce schéma distingue les différents types de carbones par exemple en fonction de leur hybridation et du nombre de liaisons covalentes (décrit sur http://www.tripos.com/mol2/atom_types.html). Le carbone est décrit sous 6 entités distinctes : C+, C1, C2, C3, Cac et Car. Le nombre de contacts décrit étant plus important, le *fingerprint* d'interactions est considérablement plus grand que le *fingerprint Elements*.



$pK_{d/i}$	C.C	...	C.Cl	...	C.I	N.C	...	I.I	PDB ID
5.70	95		30		0	73		0	2p33

Figure 30 Description de la création du fingerprint Elements [75]. L'ensemble des atomes de la protéine dans un rayon de 6,0Å autour de l'atome du chlore du ligand sont détectés et ajoutés quantitativement dans le vecteur recensant tous les types de contacts interatomiques possibles.

La création de ces deux *fingerprints* d'interaction a eu pour but d'évaluer la complexité d'un *fingerprint* dans la prédiction de l'affinité. Un modèle de régression *Random Forest* a été utilisé pour établir une fonction de score à but prédictif entrainé sur 1105 complexes. Son évaluation sur 195 complexes différents du set d'apprentissage a montré qu'une description plus simple comme *Elements* donnait de meilleurs résultats.

E. Comparaison des méthodes

A l'heure actuelle, peu d'études ont comparé les différents types de descripteurs d'interactions. En 2016, Lenselink et collaborateurs [77] ont effectué un travail de recherche centré sur 5 protéines transmembranaires couplés aux petites protéiques G (des *GPCRs*) différentes : ils ont ainsi testé 5 méthodes de caractérisation et comparaison d'interactions (*SIFt*, *CREDO*, *Elements*, *Sybyl* et *SPLIF*) et évaluer leur capacité à retrouver des molécules actives par *Virtual Screening*. Chaque protéine décrite par une structure protéine - ligand de référence se voyait *docker* 100 molécules agonistes pour 5000 leurres générés par *DUD-E* [78]. Ces leurres comportent les mêmes groupements fonctionnels que les agonistes mais dans une autre disposition.

La meilleure pose de *docking* selon la méthode de *scoring* de GLIDE a été conservé pour évaluer l'efficacité de chaque méthode. Les 50 meilleures poses ont été testé séparément afin de tester la diversité des poses générées. Une méthode est considérée comme efficace dans sa capacité à classer les agonistes parmi les complexes les plus similaires à la structure de référence. Un poids plus important a été accordé aux premières places du classement dans l'élaboration du score, 80% du score global sur le premier pourcent du classement selon la méthode de score BEDROC (*Boltzmann-enhanced discrimination of receiver operating characteristic*) [79].

Les résultats obtenus montraient une meilleure performance de la méthode SPLIF sur l'ensemble des cibles étudiées. Cependant, les résultats ne sont pas homogènes sur les 5 protéines testées. Il est cependant important de noter que la génération des leurres par DUD-E ne certifie en aucun cas que ces dits leurres n'interagiront pas *in vivo* avec la protéine, ce qui limite la fiabilité de cette évaluation.

F. Difficultés et obstacles

L'ensemble de ces descriptions présente à la fois des avantages et des complications propres à chaque méthode.

L'utilisation des interactions décrites de manière explicite telles qu'une liaison hydrogène ou un contact hydrophobe par exemple requiert avant tout une mise en évidence de leur existence. Elle repose sur une définition d'ordre physique, chimique mais aussi géométrique

qui impose des seuils et contraintes parfois différents selon la méthode utilisée. Ainsi, de nombreuses interactions peuvent être négligées de par l'absence de leur caractérisation dans la littérature ou pour des paramètres géométriques sélectionnés trop stricts.

D'autre part, l'utilisation d'un *fingerprint* implique la description de l'ensemble de ces interactions explicites de manière binaire, présence ou absence, ce qui ne reflète pas les différentes nuances énergétiques. En fonction des différents éléments mais aussi de la géométrie, une liaison hydrogène peut avoir une valeur énergétique oscillant entre 1 kcal/mol et 5 kcal/mol, une différence significative dans la contribution globale de l'affinité.

Les définitions et regroupements effectués pour catégoriser les interactions selon leur type introduise des biais dans leur détection et analyse. En effet, l'utilisation de ces règles et contraintes limitent la détection d'interactions plus rares et méconnues. Or certaines de ces interactions peuvent avoir un rôle important dans le processus de liaison. De plus, la catégorisation selon leur force théorique peut être préjudiciable dans certains cas d'amélioration de l'affinité. Par exemple, une interaction halogène est censée être énergétiquement moins forte qu'une liaison hydrogène. Toutefois, l'optimisation des inhibiteurs de facteurs anticoagulant a montré qu'un atome de chlore par sa liaison halogène améliorait significativement l'affinité par rapport à un donneur de liaison hydrogène. Ainsi, certains types d'interaction vont avoir plus d'importance sur certaines protéines en fonction de la composition et structure locale de sous-régions dans la poche. Il est ainsi difficile d'évaluer l'impact concret d'une interaction spécifique sur le complexe entier.

La description des interactions est complexe car les deux approches majeures ont toutes un biais spécifique. La difficulté principale repose à trouver un juste milieu entre simplification des interactions par regroupements et spécificités de l'ensemble des contacts existants. Une approche dite explicite des interactions, où les types d'interactions sont recensés, font abstraction des éléments impliqués et ne considèrent que la nature électrostatique de l'interaction. Par exemple, un donneur de liaison hydrogène peut être un oxygène ou un azote. L'interprétation des interactions observées dans un complexe protéine - ligand est ainsi facilité. A contrario, les descriptions dites implicites, où le type d'interaction n'est pas prédéfini tel SPLIF, est plus pertinente vis-à-vis de la complexité de la liaison du ligand sur sa protéine. Toutefois, leurs visualisations et interprétations s'en retrouvent affectés car bien plus ardue. De plus, la comparaison de ces interactions implicites nécessite généralement que

les deux éléments soient identiques, or un oxygène et un azote présente tous deux un doublet non liant par exemple.

Partie 2 : Détection des contacts et applications

La spécificité d'une thèse avec une convention CIFRE est de lier une recherche d'intérêt scientifique avec des intérêts industriels de qualité. Ma thèse s'est donc déroulée entre l'Institut National de la Transfusion Sanguine (INTS, unité INSERM UMR_S 1134) et la société Discngine. Mes contributions scientifiques ont été très liées dans cette entreprise au développement du projet 3decision®.

Chapitre 1 : Projet 3decision®

L'idée initiale à l'origine du développement de 3decision® provient d'une observation de l'exploitation insuffisante, ou au moins souvent inadaptée, des données structurales dans le milieu industriel. Cette difficulté provient de plusieurs facteurs notamment de la dispersion des données au sein d'une même entreprise, voire d'une même équipe. Le manque d'une plateforme collaborative similaire à ce que propose la PDB pour la recherche académique, résulte dans des analyses fastidieuses, le plus souvent incomplètes, et contraignantes en termes de temps, notamment au regard des structures disponibles au sein d'une même entreprise pharmaceutique. De plus, ces données sont fréquemment conservées uniquement au sein d'une même équipe, ainsi peu d'informations sont partagées entre les différents groupes de recherche.

Le développement du projet 3decision® a débuté en 2014 sous la direction de Peter Schmidtke et a pour objectif d'harmoniser le stockage, mais aussi optimiser l'utilisation des structures tridimensionnelles dans le cadre de projet de développement pharmaceutique. L'infrastructure de 3decision® repose sur 3 entités distinctes, toutes dépendantes les uns des autres (voir Figure 31) : (i) une base de données relationnelle Oracle, (ii) un *framework* de programmation visuelle, BIOVIA Pipeline Pilot [80] et (iii) un *backend* et *front-end* basé sur Node.js et Angular.js respectivement.

La base de données Oracle conserve les informations extraites d'une structure protéique (PDB), des informations calculées par 3decision® et aussi des informations extraites de bases de données telles que *UniProt* [16] ou *ChEMBL* [81] et les met en relation.

BIOVIA Pipeline Pilot sert actuellement de cadre de développement des flux de traitement de l'information scientifique. Il est utilisé dans un grand nombre d'entreprises pharmaceutiques de premier plan et peut-être considéré comme leader de son marché.

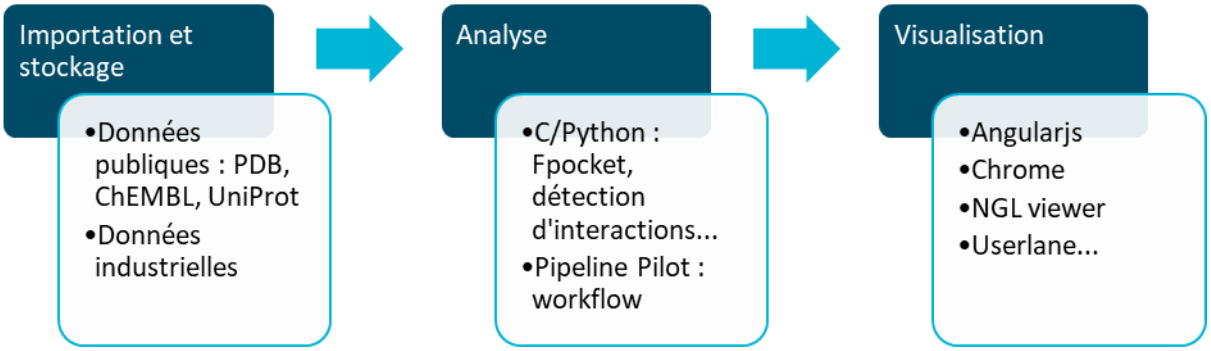


Figure 31 Diagramme récapitulatif de l'importation d'une structure cristallographique dans 3decision® ainsi que les outils utilisés.

A. Contexte industriel

Le développement du projet 3decision® s'inscrit dans une collaboration étroite entre les entreprises Discngine et Abbvie©, sous la forme d'un co-financement du projet. Sa réalisation s'effectue avec l'approche Agile, méthode de travail reposant sur des cycles de développement courts, itératifs et adaptatifs. Cette approche permet l'implémentation de nouvelles fonctionnalités à court/moyen terme en réponse continue aux besoins exprimés par les utilisateurs. Ainsi, le projet est en constante évolution, mais reste sensible en cas d'implémentation majeure. L'application est actuellement développée sur 3 serveurs distincts, le premier servant d'une part au développement ainsi qu'à la phase de beta test et donc conceptuellement plus instable. Le second est dédié au test fonctionnel et déploiement. Le troisième est l'environnement de production.

Il n'existe pas de logiciel distribué gratuitement et libre de droit capable de proposer une approche similaire dans l'étude des structures protéiques. Sur le marché industriel, deux acteurs peuvent être considérés comme concurrents directs au projet 3decision®. *Proasis*,

développé par la société australienne *DesertScientific* (http://www.desertsci.com/products/proasis4/), a permis notamment le développement de *PROLIX* [73]. Cependant, peu d'informations publiques sont disponibles sur le contenu et les fonctionnalités du logiciel. Il semblerait que la structure du logiciel ait des similitudes fortes à 3decision®, la finalité étant centrée sur le *drug design* comme illustré sur la Figure 32. *Proasis* contient une base de données structurale, une partie analyse et une partie visualisation, le tout traité dans un logiciel de *workflow* et accessible sur un serveur *cloud*. Le développement semble être réalisé de manière classique par l'intermédiaire de sorties (*releases*) annuelles.

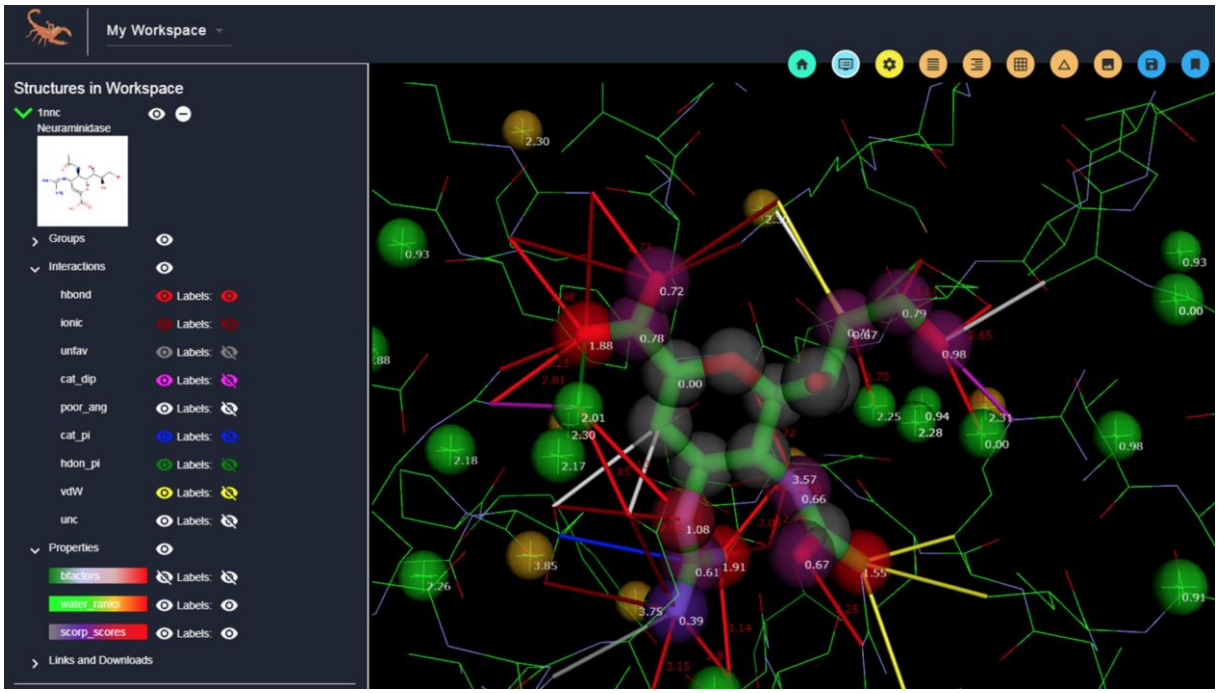


Figure 32 *Capture d'écran de l'interface web de Proasis4 décrivant pour la neuraminidase (code PDN 1nnc) les propriétés interactives à la fois du ligand mais aussi du récepteur.*

Schrödinger propose un ensemble de solutions pouvant permettre de reproduire certaines fonctionnalités de 3decision® (https://www.schrodinger.com/pldb et https://www.schrodinger.com/livedesign). Ces solutions sont capables d'interagir ensemble par le biais d'éléments externes à Schrödinger tels que des extensions KNIME (logiciel de *workflow* concurrent Pipeline Pilot) et d'API Python. En 2018, *LiveDesign* a été lancé sur le marché comme « plateforme collaborative nouvelle génération en matière de *drug design*. L'outil possède quelques similarités avec 3decision® avec une orientation plus chimie, malheureusement peu d'informations sont disponibles. PLDB est une base de données très lié

à Schrödinger Maestro, permettant l'étude des interactions protéine - ligand des structures de la PDB et des structures privées.

Relibase [82] et sa suite *Relibase+* [83], publié en 1998 et développés par le CDCC (*Cambridge Data Crystallographic Centre*) est une base de données de structures cristallographiques issues de la PDB affichant les interactions protéine - ligand (voir Figure 33). Les interactions détectées par le logiciel sont : les interactions ioniques, les liaisons hydrogènes, les métaux de coordination, l'interaction cation - π, les interactions aromatiques, les interactions hydrophobes ainsi que les collisions stériques. Cependant, certaines publications font aussi mention d'interactions moins communes pouvant être recherchées dans la base telle que les interactions entre groupements carbonyle [83]. Une version académique fut disponible gratuitement. Toutefois, *Relibase* sera mis hors service à la date du 31 Décembre 2018, la CDCC estimant que des outils accessibles au public comme la PDB ou PLIP permettent d'obtenir des informations identiques.

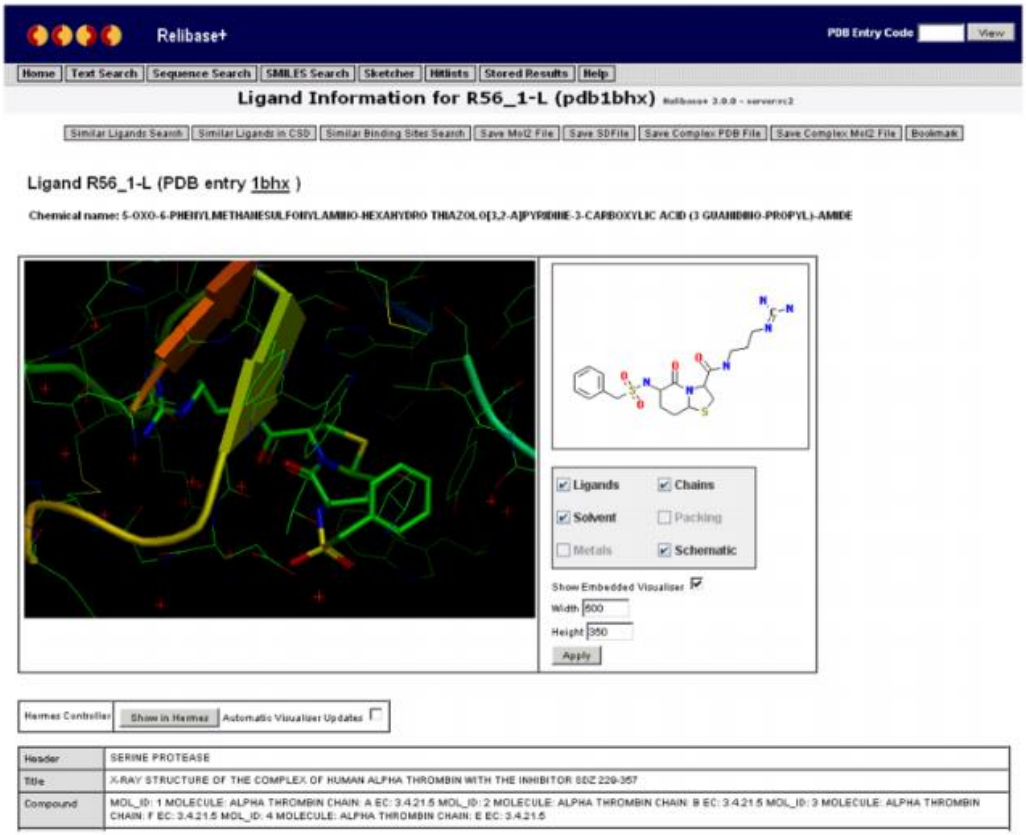


Figure 33 Illustration de l'interface de Relibase (source : http://relibase.ccdc.cam.ac.uk/documentation/relibase/relibase.1.41.html).

Psilo, solution développée par le *Chemical Computing Group* (CCG), est considéré comme un concurrent direct de 3decision®. Extension du logiciel MOE (*Molecular Operating Environment*), il contient une base de données de structures cristallographiques provenant à la fois du domaine publique (PDB) mais aussi du secteur privé. Les annotations provenant de plusieurs librairies d'annotations telles que les domaines SCOP [17] ou les annotations GO (*Gene Ontology*) [84] sont assignées aux structures. Des interactions entre protéine et ligand sont également calculées, toutefois aucune information sur la nature ou les critères géométriques utilisée ne semble être disponible sur leur site ou dans la littérature (voir Figure 34).

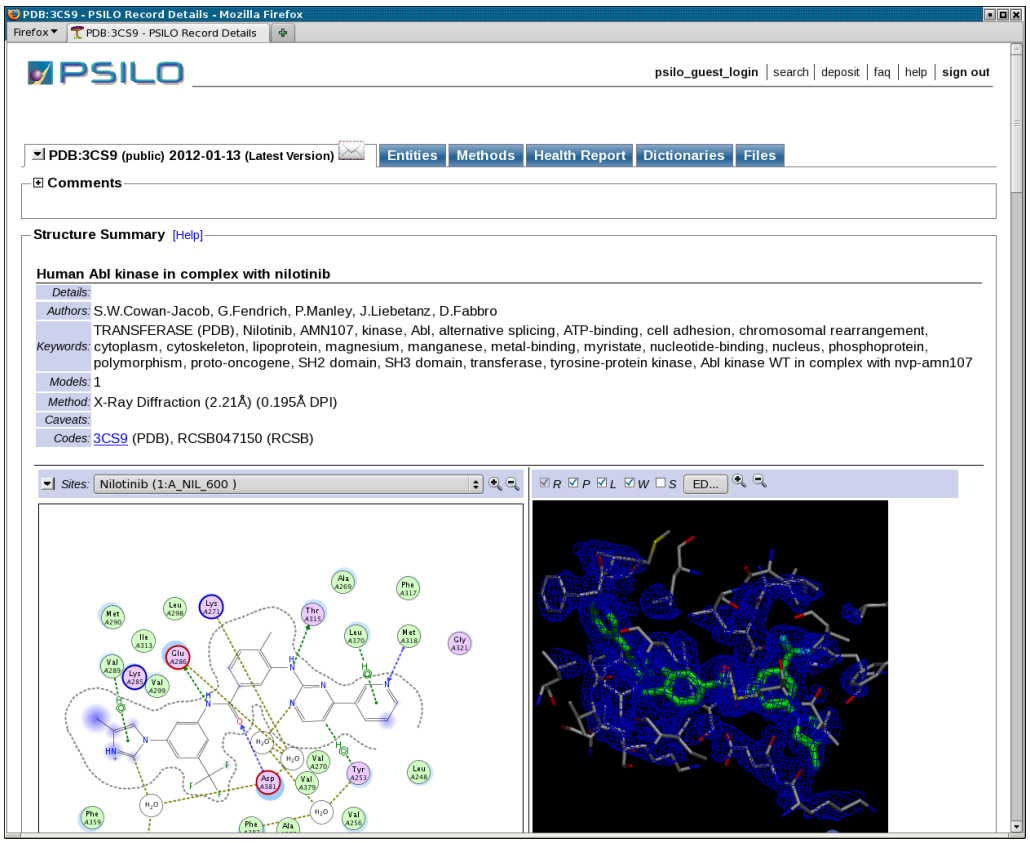


Figure 34 Interface web de la solution Psilo sur la structure 3cs9 (image provenant de https://www.chemcomp.com/img/psilo/).

B. Enregistrement des structures et stockage dans 3decision®

L'ensemble des structures issues de la PDB et de la base de données *SwissProt* [85] sont sauvegardées dans une base de données sécurisée maintenue par le système de gestion de données de base relationnelle *Oracle Database* (Oracle DB). Lors du processus d'importation

de la structure tridimensionnelle, de nombreuses vérifications sont effectuées. Tout d'abord, une identification et une annotation des éléments présents dans la structure enregistrée sont accomplies.

Chaque chaine protéique importée va être identifiée par l'intermédiaire d'une recherche de séquence *BLAST* [86] dans la base de séquences *UniProt* intégrée à 3decision® et sera annotée avec son code *UniProt* correspondant. Les résidus manquants ne sont pas complétés.

Un processus similaire sera effectué pour les petites molécules, annotées hétéroatomes dans un fichier pdb (*HETATM*), qui seront identifiées et regroupées par *fingerprint* moléculaire SMILES [87]. Toutefois, l'intégration des ligands provenant de la PDB dans Pipeline Pilot induit dans certains cas une mauvaise assignation des liaisons covalentes ou de leur ordre correspondant. Les ligands sont alors retraités afin que la stéréochimie soit correctement respectée. En cas d'atome manquant, le retraitement de ces molécules est alors limité et les atomes en question ne sont donc pas reconstruits.

Les informations extraites des macromolécules et petites molécules sont ensuite enregistrées dans la base de données comme par exemple l'identifiant résidu du ligand (généralement un code à 3 lettres), sa position dans le fichier pdb déterminée par le numéro de résidu ou la conformation enregistrée en cas de conformères alternatifs.

Lors de l'enregistrement de la structure, une première série d'analyse est effectuée. Dans un premier temps, la détection automatique de poches protéiques par le programme informatique *fpocket* [88] est effectuée. Ce procédé repose sur une détection géométrique du site de liaison protéique par la méthode des sphères α. Chaque sphère, vide, doit être en contact avec 4 atomes sur sa périphérie, ses 4 atomes reflétant la courbure locale du site de liaison. Les sous-poches sont ensuite regroupées par classification hiérarchique.

La détection d'interaction, sujet de ce travail de thèse, est aussi effectuée durant ce processus, la méthode étant décrite dans une partie ultérieure.

C. Visualisation

3decision® est une solution dite *cloud*, les données et autres calculs informatiques sont stockés et traités sur des serveurs informatiques externes. L'affichage est réalisé par l'intermédiaire d'un navigateur web connecté au serveur dédié.

La recherche de structure peut se faire par l'intermédiaire de codes PDB, nom de protéines ou sous-structures moléculaires par exemple. Une fois la recherche effectuée, l'affichage des différentes informations issues d'une structure est réparti sous forme de panneaux, illustré sur la Figure 35. Chaque panneau constitue soit un outil tel qu'une recherche par similarité de séquence ou un module de visualisation d'une propriété physico-chimique par exemple. De nombreux modules sont disponibles : un module résumant les informations sur la structure (résolution, nom de la protéine), un panneau affichant sous forme de graphe l'ensemble des poches présentes, un module de création de ligand, etc. Un exemple illustré sur la Figure 35 montre la visionneuse 3D (label 4) ainsi qu'un panneau d'annotations de séquences (label 5). Ces modules sont interactifs : la sélection d'un résidu sur le module d'annotation de séquence est mise en avant sur la visionneuse (en anglais *viewer*) 3D.

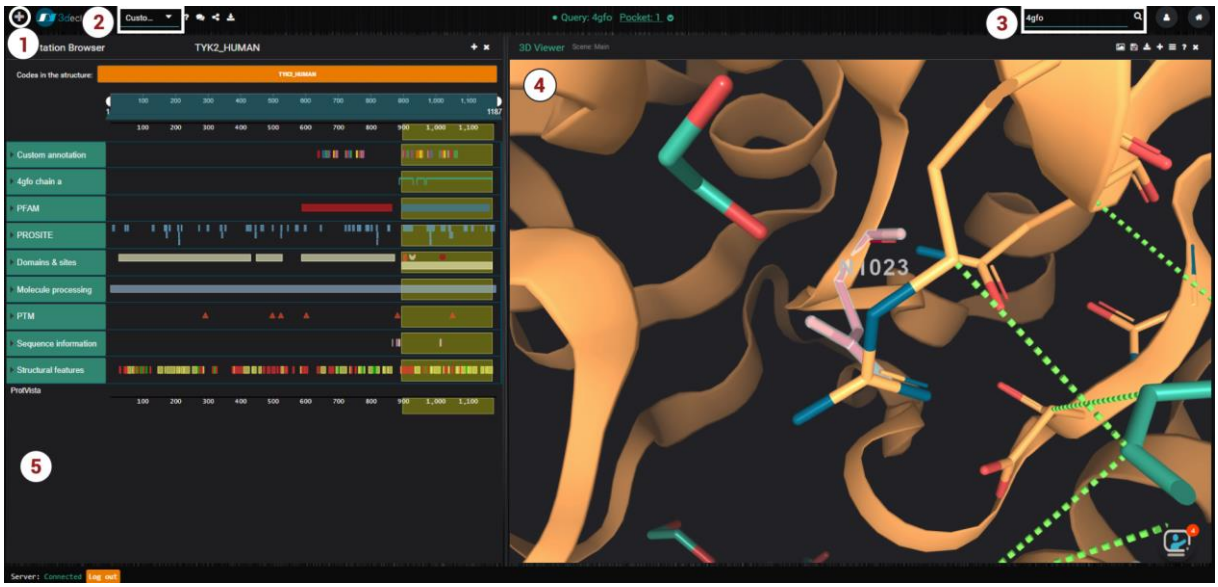


Figure 35 Exemple de sélection d'un résidu sur le panneau d'annotations et affichage sur la structure 3D par 3decision®. 1. Ajout d'un nouveau panneau 2. Arrangements de panneaux prédéfinis 3. Barre de recherche 4. Viewer 3D 5. Panneau d'annotation de séquence.

D. Analyse et outils de *drug design*

De nombreuses analyses sont disponibles pour l'utilisateur afin d'affiner les recherches pour une cible thérapeutique spécifique. Le programme de *docking rDock* [89] est ainsi directement implémenté dans 3decision®. L'ensemble des ligands présents dans une poche identique à travers la base de données peut être affiché. Leurs propriétés physico-chimiques telles que le poids moléculaire, le nombre de donneurs de liaisons hydrogènes etc. sont notamment calculables et peuvent être visualisés à l'aide d'une représentation statistique tel un histogramme. Une recherche par similarité de sous-poches a été intégrée selon un algorithme identifiant dans un premier temps les pharmacophores de poches en commun, puis dans un second temps va aligner et évaluer la qualité de la superposition des pharmacophores. Les travaux effectués ici s'inscrivent dans la détection automatique des interactions moléculaires lors de l'enregistrement d'une structure protéique. Leurs caractérisations vont notamment à faciliter la visualisation et l'interprétation du complexe protéine - ligand.

Chapitre 2 : Implémentation de la détection de contacts dans 3decision®

Jusqu'alors, la détection des interactions protéine - ligand dans 3decision® était réalisé par l'intermédiaire de *PLIP*. Cependant, l'utilisation d'un programme tierce (i) n'était pas satisfaisante dans certains cas (voir Partie 2.3.D), (ii) et amenait de trop fortes contraintes sur le processus de détection et définition d'interaction. Dès lors, l'implémentation d'un programme informatique propre à 3decision® s'est imposée comme une nécessité.

Les interactions moléculaires sont avant tout des contacts proches entre atomes provenant de deux éléments distincts. La détection de contacts est un processus se déroulant en plusieurs étapes. Tout d'abord, il est nécessaire de différencier dans notre cas la molécule qui sera considérée comme ligand du reste de la structure, défini comme récepteur. Une fois cette étape réalisée, la définition et détection des contacts se fera par l'identification des atomes du récepteur se situant dans un rayon prédéfini autour d'un atome du ligand.

La définition et classification des interactions moléculaires dépend des spécificités chimiques et de l'arrangement spatial entre deux atomes impliqués dans un contact. Cette étape de caractérisation des deux atomes en question est alors réalisée à posteriori, notamment pour des raisons computationnelles et de performances. Plus la description de chaque atome est détaillée, plus pertinente devrait être la description de la nature de l'interaction.

Ces différentes étapes sont décrites et détaillées dans les parties suivantes.

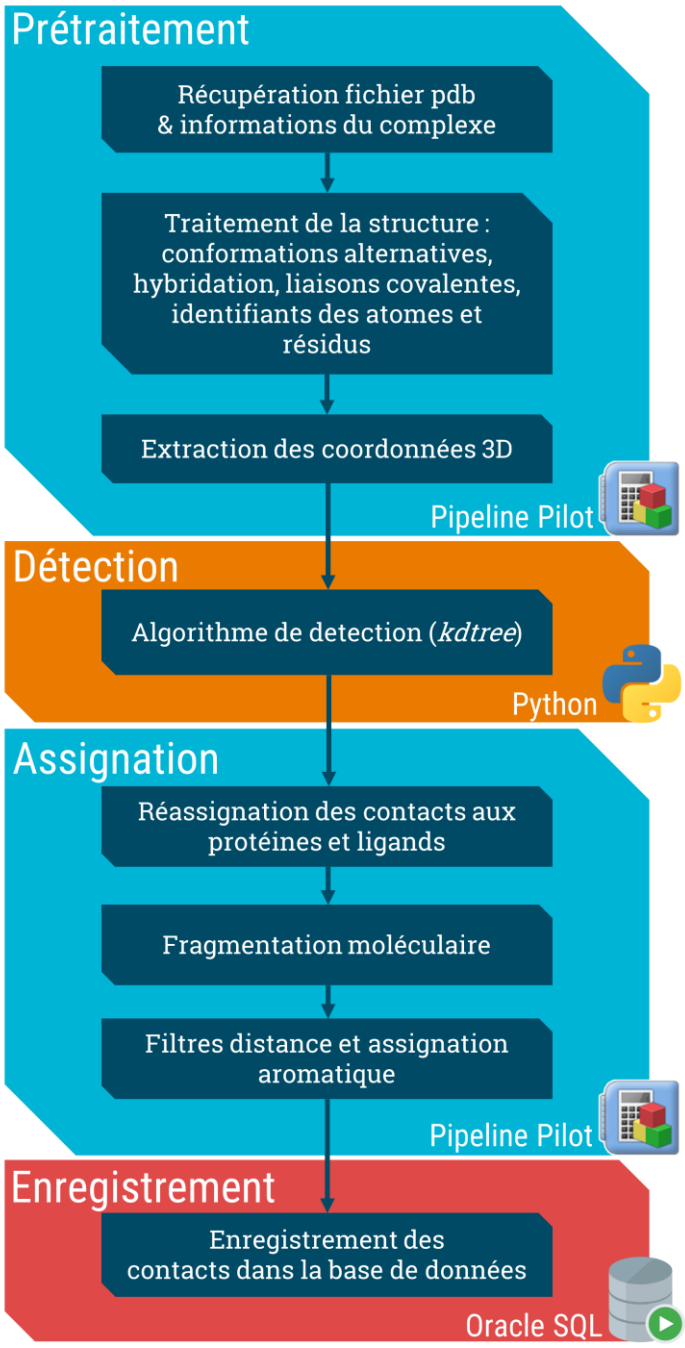


Figure 36 Schéma récapitulatif du processus de détection de contacts et des outils informatiques correspondants.

A. Prétraitement et récupération des coordonnées

La détection des contacts dans un complexe protéine - ligand nécessite l'ensemble des coordonnées atomiques 3D à la fois du ligand mais du reste de la structure. Les coordonnées du ligand sont récupérées à partir de la table de données correspondantes. Cependant, les coordonnées restantes telles que le récepteur protéique, les co-facteurs et molécules d'eau ne sont pas présentes dans la base de données pour des raisons de performances nécessitant le retraitement du fichier PDB par Pipeline Pilot (voir Figure 36).

Chaque complexe est considéré comme étant une conformation spécifique d'un ligand en contact du reste de la structure. Un ligand est annoté comme hétéroatome, *HETATM* dans un fichier pdb (1 sur la Figure 37), alors qu'un récepteur peut être une protéine, annoté *ATOM* dans ce même pdb (2 sur la Figure 37), ou n'importe quel autre hétéroatome. Ainsi, les éléments pouvant être considérés comme appartenant au récepteur sont généralement des protéines, molécules d'eau, ions, cofacteurs, acide nucléique ou un autre ligand.

Une gestion des conformations alternatives est aussi prise en compte. Les structures RMNs par exemple comporte plusieurs modèles distincts d'une même structure dans un même fichier (label 3 sur Figure 37). Chaque modèle va être traité isolément. De même, en cas de résolution approximative de certains atomes, plusieurs conformations sont présentes, différenciées par des identifiants spécifiques ainsi que des valeurs d'occupation (label 4 sur la Figure 37). Cette valeur d'occupation, *occupancy* en anglais, désigne le pourcentage de conformations identiques observées par le cristallographe dans un cristal (label 5 sur la Figure 37). Dans notre cas, seule la conformation la plus fréquente sera considérée et pour des valeurs égales (0,5), la première conformation par ordre d'apparition sera sélectionnée.

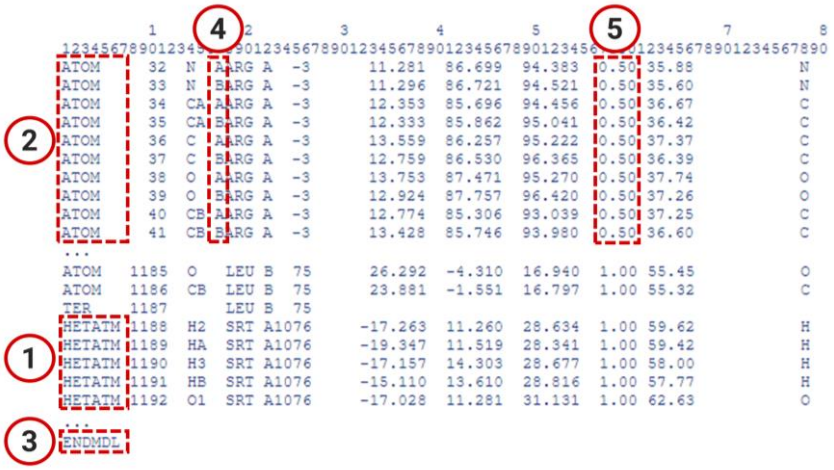


Figure 37 Exemple des données présentes dans un fichier pdb. 1. Champ hétéroatome 2. Champ atome protéique 3. Fin du modèle, généralement présent dans les structures RMN 4. Identifiant de la conformation alternative 5. Occupation.

B. Algorithme de détection

Pipeline Pilot est un logiciel dont l'utilisation première reste le traitement à haut débit de données chémoinformatiques. De ce fait, la gestion des éléments 3D a quelques limitations du fait (i) du langage informatique employé dans Pipeline, et (ii) des outils chémoinformatiques, limités dans le traitement de l'information 3D. Le langage interne, dérivé du PLSQL, n'inclut pas des fonctions mathématiques avancées notamment des calculs sur vecteurs multidimensionnels. Les contraintes de performances imposées par les standards industriels, telles que le traitement de poses de *docking*, nous ont obligé à externaliser partiellement l'algorithme de détection en dehors du pipeline, illustré par l'étape de détection sur la Figure 36.

Les coordonnées 3D du ligand et du récepteur sont alors extraites du fichier pdb dans *Pipeline Pilot* puis enregistré dans un fichier texte temporaire sur le serveur Linux de Pipeline. Un script Python, à l'aide du *package numpy*, permet la réorganisation des coordonnées sous forme d'arbre à k-dimensions (k = 3 dans notre cas), autrement appelé arbre k-d ou *kdtree* en anglais.

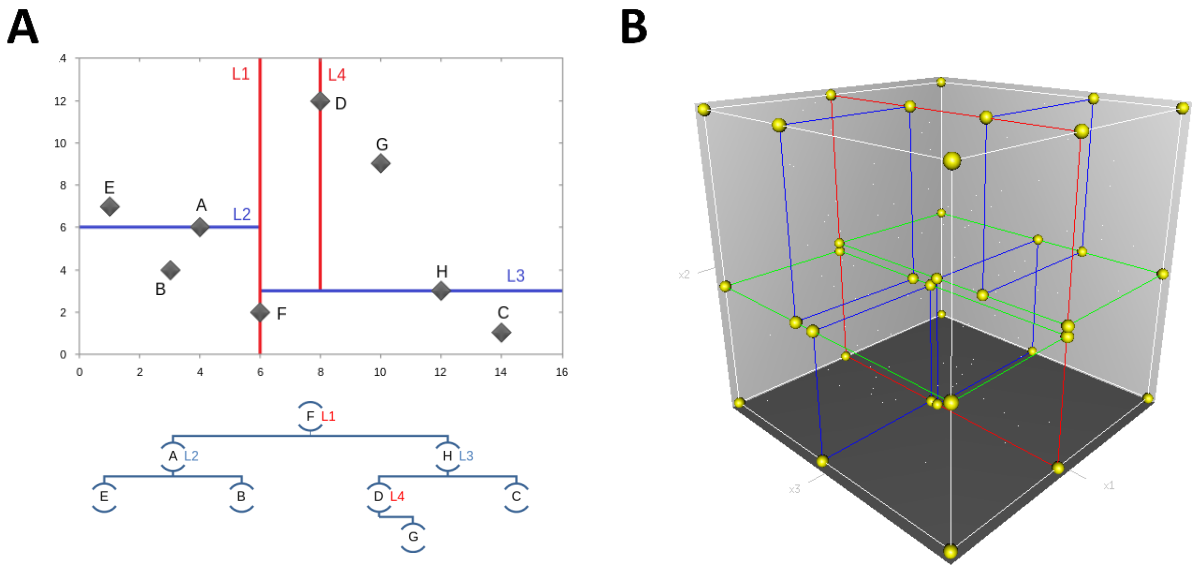


Figure 38 Illustration des arbres k-d A. Schéma de l'élaboration d'un arbre k-d à 2 dimensions B. Arbre k-d à 3 dimensions (provenant de Wikimedia Commons).

Ces arbres k-d sont des structures de données binaires à k-dimensions permettant d'effectuer des recherches rapides par plage. Dans le cas de coordonnées tridimensionnelles, des nœuds sont générés séparant l'espace dimensionnel en plusieurs sous-espaces correspondant à un ensemble d'atomes proches, schématisé sur la Figure 38A. Ces sous-espaces forment une branche de l'arbre regroupant dans des « feuilles » des éléments spatialement proches (dendrogramme de la Figure 38A). Le nombre de recherche à effectuer est alors limité par un nombre restreint de branches à comparer. Chaque atome du ligand va ainsi être soumis à une recherche dans l'arbre k-d définit pour le récepteur selon une distance de 6,0Å initialement ici. Ce processus est effectué dans un script *python* opérant sur le serveur *Linux* de Pipeline Pilot prenant en entrée un fichier texte contenant les coordonnées 3D du ligand et du récepteur. Le fichier de sortie issu du programme python contient uniquement les paires d'atomes du ligand et du récepteur en contact sans valeur de distance.

Les résultats sont ensuite importés dans le protocole Pipeline Pilot pour être réassignés à leurs données respectives, à savoir le numéro du résidu correspondant, le nom de l'atome du ligand, sa chaine associée. Un second seuil de distance est appliqué par la suite éliminant les contacts dont la distance est supérieure à la somme des rayons de Van der Waals des deux

éléments impliqués plus un delta défini comme 1,0Å ici. Les contacts à la fois protéine - ligand mais aussi ligand-eau, ligand-ADN et ligand-ligand sont calculés.

C. Retraitement et caractérisation des contacts

La détection des contacts étant effectuée, il est nécessaire de caractériser les deux atomes mis en jeu dans chaque contact afin de déterminer la nature de l'interaction moléculaire. Cette caractérisation est multiple, elle prend en compte à la fois l'agencement spatial mais aussi les propriétés chimiques et électrostatiques de chaque atome. L'étape de caractérisation est effectuée dans un protocole Pipeline Pilot (Assignation dans Figure 36), à posteriori de la détection de contacts afin de limiter le nombre de caractérisation à effectuer (seulement sur les atomes en contacts).

a. Fragmentation

La fragmentation moléculaire va permettre de (i) définir l'environnement local ou groupement fonctionnel, d'un atome en contact (ii) à plus long terme considérer les interactions moléculaires par l'intermédiaire de fragments et non plus d'atomes à des fins de *Fragment-Based Drug Design*.

La fragmentation moléculaire consiste à regrouper des atomes d'une molécule en un ensemble de fragments de taille réduite déterminé en fonction de certains paramètres [90]. De manière arbitraire, les paramètres sélectionnés lors de ce procédé garantissent la conservation de l'intégrité des groupements fonctionnels et des groupements terminaux. Néanmoins, les longues chaines aliphatiques ainsi que les groupement multi-cycliques ont été séparés. A titre d'exemple, le groupement adénine du ligand *PU8* illustré sur la Figure 39 (*9-Butyl-8-(2-chloro-3,4,5-trimethoxy-benzyl)-9H-purin-6-ylamine*) voit ses deux aromatiques séparés de manière distinctes.

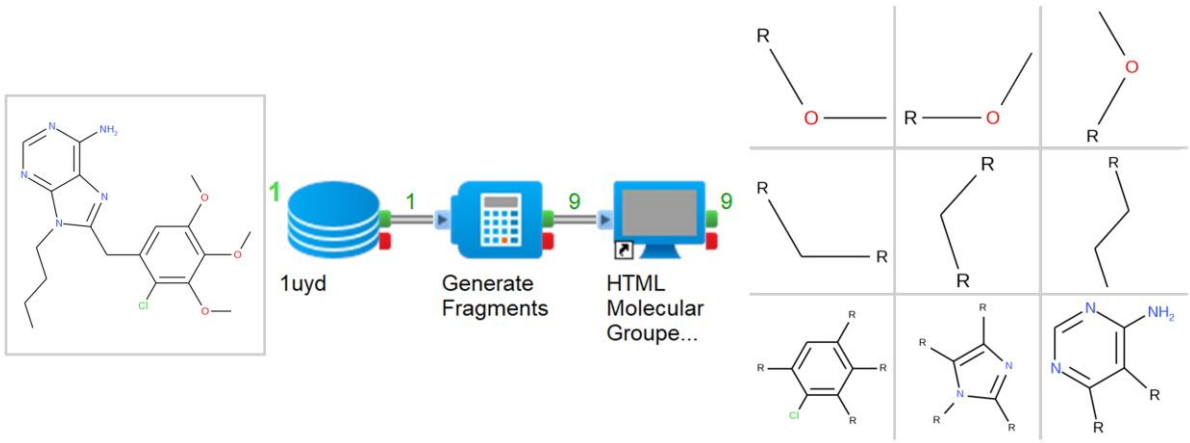


Figure 39 Exemple de fragmentation d'un inhibiteur de HSP90 (code PDB HETATM PU8) en 9 fragments distincts dans Pipeline Pilot.

b. Descripteurs géométriques

La définition des descripteurs géométriques est réalisée à 2 fins spécifiques : (i) affiner la définition de la nature de l'interaction, et, (ii) permettre l'alignement et la superposition des atomes interagissant avec un atome en commun à travers plusieurs structures. Différents descripteurs sont calculés pour définir l'orientation géométrique relative de chaque atome, à la fois du côté du récepteur mais aussi du ligand. Chaque atome va être caractérisé par un vecteur dépendant du nombre d'atomes covalents rattachés à cet atome.

Tout d'abord des angles nommés angles d'élévation vont être déterminés à l'aide de vecteurs présents du côté du ligand et du récepteur. Ces angles sont dits d'élévation car ils définissent la position d'un atome par rapport au plan horizontal de l'atome référent soit sa hauteur. Cependant, il ne caractérise pas un point précis autour de l'atome mais une région circulaire, illustré par le cercle vert sur la Figure 40.

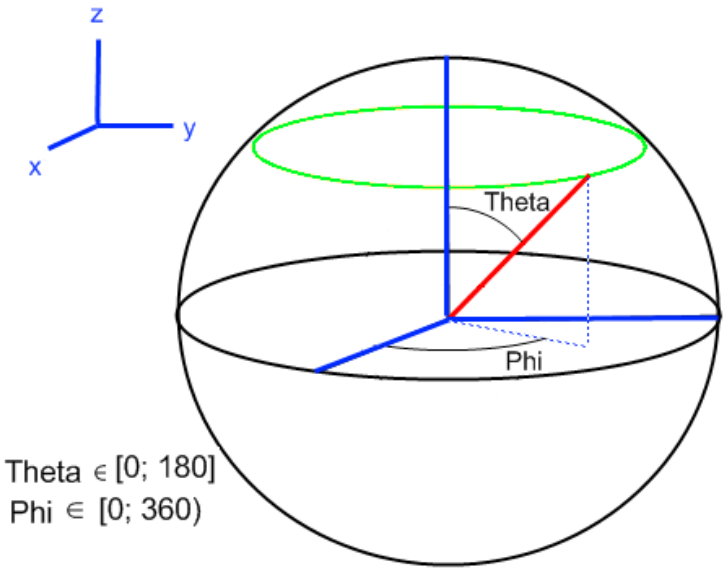


Figure 40 Description d'un angle d'élévation (theta) pouvant designer tous points présents sur le cercle en vert. Sans repère local directeur x et y, il est impossible de localiser sur ce cercle la position précise du point.

Dans notre cas, cet angle n'impliquera pas un plan mais un vecteur directeur défini en fonction du nombre de liaisons covalentes auquel l'atome de référence est rattaché. Ainsi, pour une seule liaison covalente, le vecteur est déterminé le long de la liaison covalente (voir Figure 41A). En présence de deux liaisons covalentes, le centre géométrique des deux atomes covalents est déterminé et servira d'origine au vecteur, schématisé sur la Figure 41B. En présence de 3 liaisons covalentes si le fragment n'est pas de type plan, alors le vecteur est déterminé de la même manière qu'avec 2 liaisons covalentes (Figure 41C). Si les quatre atomes sont dans un même plan, alors deux vecteurs normaux sont générés par rapport à l'atome en interaction, le vecteur final sera dirigé vers le partenaire de l'interaction (Figure 41D).

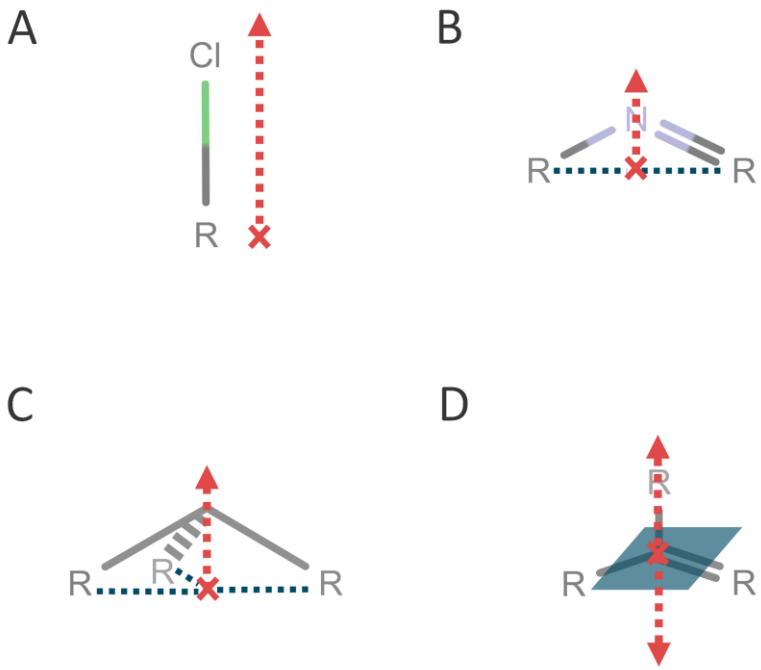


Figure 41 Schéma de la création des vecteurs nécessaires au calcul d'angles d'élévation.

Ces vecteurs directeurs couplés au vecteur entre les deux atomes permettent de déterminer 3 angles. L'angle entre le vecteur directeur défini sur le ligand et le vecteur $\overrightarrow{Atome_{lig}Atome_{rec}}$ va déterminer la position circulaire (cf. cercle vert de la Figure 40) de l'atome du récepteur par rapport à l'atome du ligand, dit angle *ligand*. La même opération est répétée du côté du récepteur définissant l'angle *récepteur*. Enfin, l'angle défini entre les vecteurs directeurs du ligand du récepteur va quantifier la direction relative des deux atomes, l'angle d'*orientation*.

Les angles ligands et récepteurs varient entre 0 et 180°, 90° correspondant à un atome positionné de manière orthogonale par rapport au vecteur directeur, 180° étant un atome en face du vecteur directeur. Les angles d'orientation oscillent entre 90° et 180°, la première valeur indiquant un atome dont la direction est orthogonale par rapport à la direction définie par l'autre atome. Les deux atomes sont inversement colinéaires pour une valeur de 180°. La Figure 42 illustre quelques exemples d'angles.

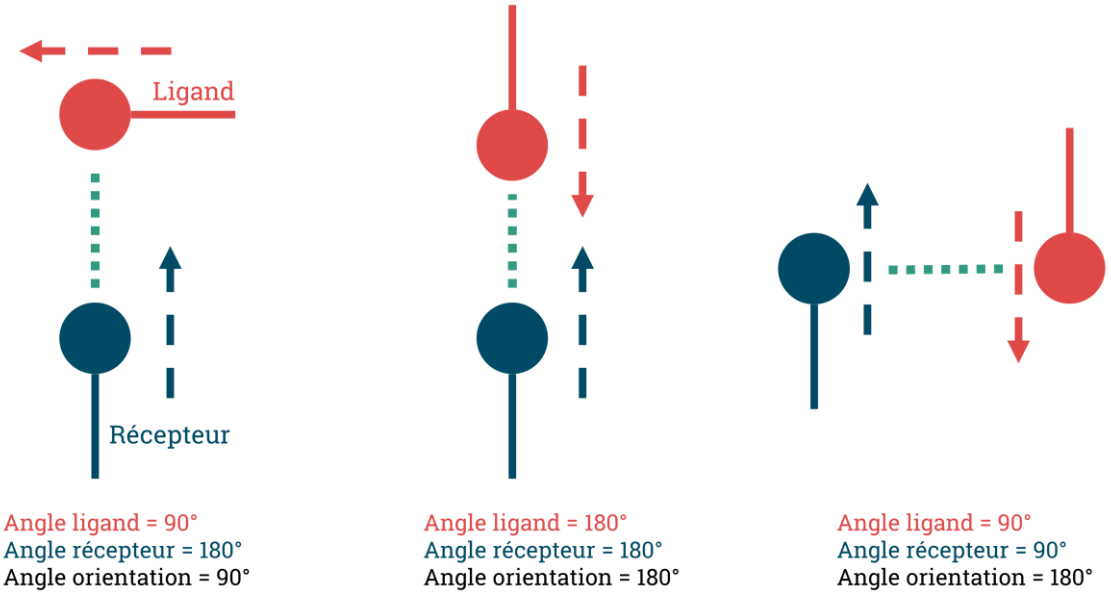


Figure 42 Exemples d'arrangement géométrique 2D et leurs angles correspondants.

c. Caractérisation chimique

Jusqu'à maintenant tout traitement sur les contacts était fait sans typage. Néanmoins, notamment en *drug design* et en analyse d'interactions moléculaires, des types d'interactions sont utilisés afin de caractériser un contact. La classification des contacts en interactions moléculaires prédéfinis requiert la description des propriétés électrostatiques de chaque atome impliqué dans le contact. Les composants présents dans Pipeline Pilot permettent de définir un certain nombre de propriétés physico-chimiques. L'hybridation de chaque atome est ainsi relevée et assignée aux atomes correspondants. L'hybridation décrit la configuration électronique d'un atome ou l'état d'occupation des orbitales. Dans certains cas, elle indique le nombre et la nature des liaisons covalentes effectuées par l'atome en question. La Figure 43 illustre les diverses hybridations pour un azote : sp^2 correspond à une liaison double et une liaison simple, sp^3 correspond à trois liaisons simples.

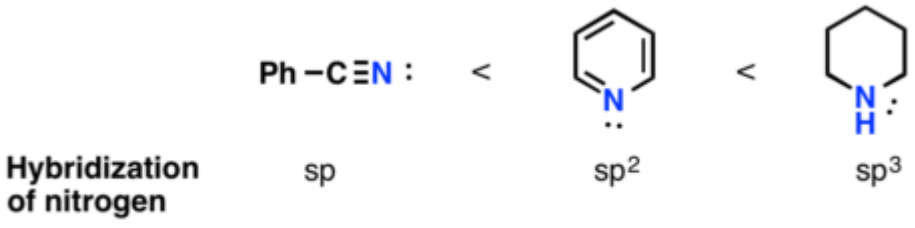


Figure 43 Description de l'hybridation de l'atome d'azote en fonction des différents types de liaisons covalentes.

D'autres descripteurs sont assignés à chaque atome. Des fonctions dédiées dans Pipeline Pilot permettent de définir le caractère donneur ou accepteur de liaison hydrogène d'un atome par exemple. Le nombre et les éléments liés de manière covalentes à l'atome d'interaction sont aussi considérées. L'appartenance à un groupement aromatique ainsi que le nombre d'hydrogènes implicites est aussi pris en compte dans les descripteurs chimiques. Enfin, la charge partielle de l'atome est calculé selon l'algorithme de Marsili-Gasteiger [91].

Cet algorithme définit les charges partielles en trois étapes. Tout d'abord, dans le cas d'un ligand dont les hydrogènes n'ont pas été résolu, ces derniers seront ajoutés temporairement à la molécule. Par la suite des charges partielles initiales sont attribuées à chaque atome en fonction de l'élément présent, par exemple -0.5 pour un oxygène carboxylé. Enfin, une étape itérative d'optimisation est réalisée durant laquelle les charges partielles vont être distribué le long des liaisons covalentes en fonction de l'électronégativité des éléments liés. Bien que cette méthode ait été développé pour comparer la réactivité théorique des groupements fonctionnels entre différentes molécules, les résultats obtenus dans la définition des accepteurs et donneurs de liaisons hydrogènes a été jugé comme satisfaisante.

D. Contacts aromatiques

A posteriori de l'étape de caractérisation de contacts, une distinction va avoir lieu entre les contacts impliquant deux entités aliphatiques et ceux impliquant au moins un élément aromatique. Les groupements aromatiques étant considérés dans la littérature comme une entité d'interaction à part entière, tout contact impliquant au moins un atome caractérisé comme aromatique sera dupliqué et traité dans un protocole distinct. Afin de limiter la redondance de ces contacts, tous les contacts appartenant à un même cycle aromatique sont regroupés ensemble. Ainsi, pour chaque contact atome – aromatique ou aromatique – aromatique, une seule entrée sera sauvegardée dans la base de données, réduisant dans certains cas l'information par un facteur 15, illustré sur la Figure 44.

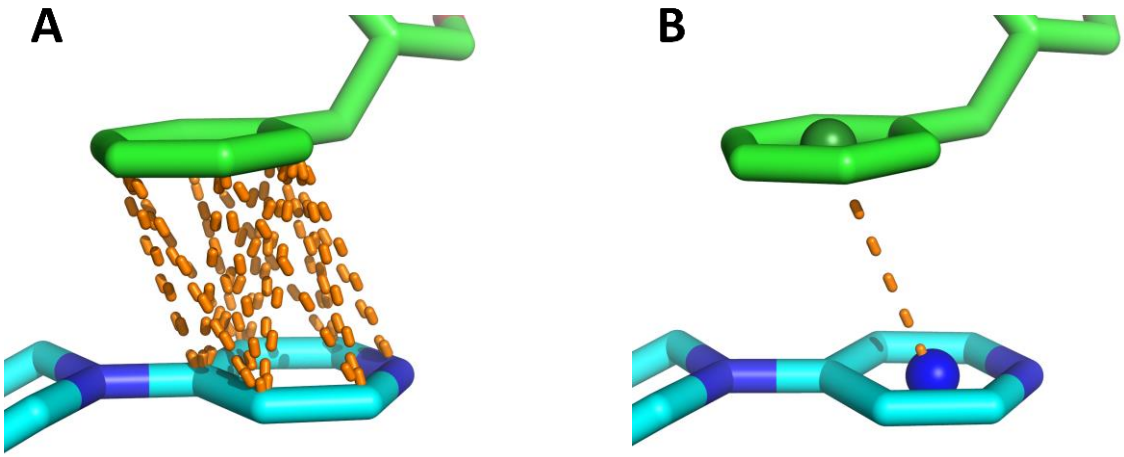


Figure 44 A. Ensemble des contacts aromatiques enregistrés individuellement dans la table de contacts. B. Description de l'interaction aromatique enregistrée dans la table de contacts aromatiques.

La description de l'aromatique va être désigné par son centre de masse qui servira comme point de référence pour l'ensemble des descripteurs géométriques. Par conséquent, un nouveau vecteur normal est généré afin d'obtenir les angles sphériques et coordonnées locales correspondantes. Ce vecteur normal sera toujours dirigé dans la direction de l'atome ou du cycle d'interaction. La taille et composition des aromatiques sont aussi conservées comme caractéristiques.

E. Enregistrement dans la base de données

Chaque contact est individuellement sauvegardé dans une table de données de la base SQL (voir Annexe 1). La représentation textuelle de chaque fragment sous format *Chemical Table* (ou *ctab*) [92] est aussi enregistrée, permettant la préservation des coordonnées 3D ainsi que des groupements radicaux. Les contacts impliquant les cycles aromatiques sont enregistrés dans une table de données séparée. Au mois de Novembre 2018, plus de 40 millions de contacts sont enregistrés dans 3decision®.

F. Accessibilité des données

Actuellement, l'intégralité du protocole de détection de contacts, à l'exception de l'algorithme d'arbre k-d, se fait dans le *framework* de développement Pipeline Pilot, et est par conséquent inaccessible à tout personne externe au développement de 3decision®. A la publication de ces travaux, le script python permettant d'effectuer une recherche par arbre k-d à partir d'un fichier de coordonnées sera disponible à l'adresse suivante : https://github.com/Discngine/. Une version python de l'intégralité du protocole, à savoir la caractérisation géométrique 2D et 3D mais aussi un *parser* de fichier pdb, est en cours de développement et sera prochainement accessible à l'ensemble de la communauté scientifique à la même adresse.

Chapitre 3 : Visualisation et description des contacts

La visualisation des interactions moléculaires pour un ligand se fait par l'intermédiaire de deux requêtes SQL sur la table de contacts et de contacts aromatiques. Un ensemble de règles décrites ci-dessous va permettre de filtrer les contacts les plus pertinents et être affichés comme interactions communément définies.

	Accepteur	Donneur	Distance
Liaisons hydrogènes	• Doublet non-liant • Charge partielle négative • Système π d'un aromatique et halogènes lourds (angle sphérique < 110°)	• Nombre d'hydrogènes implicites > 0 • Atome polaire ou charge partielle négative	3,5Å
Liaisons halogènes	• Doublet non-liant • Charge partielle négative • Système π d'un aromatique	• Halogène lourd (Cl, Br, I) • Angle sphérique > 140°	Σ(rayons de VdW) + 1,0Å
Métaux de coordination	• Doublet non-liant • Charge partielle négative • Système π d'un aromatique et halogènes lourds (angle sphérique < 110°)	• Cation métallique	2,5Å
Ponts salins (interactions ioniques)	Atome portant une charge positive	Atome portant une charge négative	Σ(rayons de VdW) + 1,0Å
Contacts polaires	• Doublet non-liant • Charge partielle négative • Système π d'un aromatique et halogènes lourds (angle sphérique < 110°)	• Nombre d'hydrogènes implicites > 0 • Atome polaire ou charge partielle négative	> 3,5Å
Contacts hydrophobes	Tout contacts impliquant un carbone porteur dont l'un des hydrogènes (position estimée par géométrie VSEPR) est orienté vers un autre atome.		Σ(rayons de VdW) + 1,0Å

Tableau 1 Définitions des interactions moléculaires décrites dans l'outil de visualisation 3D de 3decision®.

A. Interactions polaires

a. Liaisons hydrogènes

Les donneurs et accepteurs de liaisons hydrogènes sont identifiés dans un premier temps. Les atomes caractérisés comme potentiels accepteurs sont discernés selon plusieurs critères : (i) soit par la fonction Pipeline Pilot, (ii) soit par une caractérisation comme polaire ou d'une charge partielle de Gasteiger inférieure à -0.20 ou, (iii) soit par la présence d'un doublet non liant. De même, un atome sera déterminé comme donneur de liaison halogène si (i) le composant Pipeline Pilot le définit comme tel, ou, (ii) l'atome est polaire ou dispose d'une charge partielle inférieure à -0.20 et est rattaché à des hydrogènes implicites.

Tout accepteur et donneur de liaison hydrogène présent à une distance inférieure à 3,5Å fut considéré comme une potentielle interaction hydrogène. Ce critère de distance correspond à une énergie d'interaction modérée et est fréquemment utilisée dans des analyses de structures secondaires notamment [93, 94]. Le Tableau 1 récapitule les angles considérés pour les donneurs et accepteurs de liaisons hydrogènes. Ainsi, l'oxygène d'un groupement carbonylé est caractérisé comme accepteur de liaison hydrogène par une surface d'interaction dans un intervalle compris entre 110° et 180°. De même, un atome d'azote hybridé sp^2, illustré sur la Figure 43, aura une surface d'interaction dont l'angle sphérique sera compris entre 140° et 180°. Une exception a été mise en place pour les halogènes lourds, à savoir le chlore, le brome et l'iode dont la distribution anisotrope des électrons induit une surface acceptrice de liaisons hydrogènes à un angle sphérique compris entre 70° et 110°. L'ensemble de la surface du fluor a été considéré comme potentiel accepteur de liaison hydrogène. Un exemple de liaison hydrogène détecté dans 3decision® est illustré sur la Figure 45.

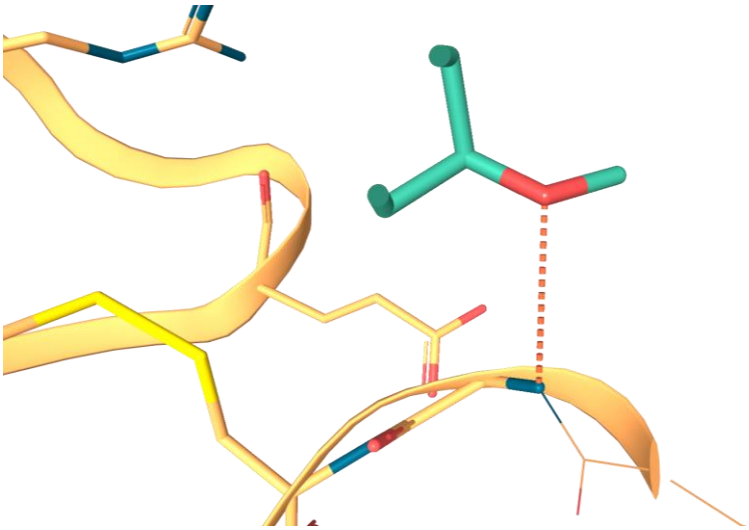


Figure 45 Détection et affichage d'une liaison hydrogène faible entre la glycine 215 et le groupement méthoxyle dans un inhibiteur de facteur anticoagulant Xa par 3decision® (code PDB 2pr3).

La surface d'interaction des donneurs de liaisons hydrogènes respecte la géométrie VSEPR (*Valence Shell Electron Pair Repulsion* en anglais, ou Répulsion des Paires d'Électrons des Couches de Valence), la position des atomes d'hydrogènes implicites dépend donc du nombre de liaison covalentes engagé par l'atome porteur de l'hydrogène. Ainsi, l'angle sphérique permet de couvrir l'ensemble des positions adoptés par un hydrogène sur un groupement hydroxyle, soit un angle de 109,25° en théorie. Aucune restriction d'angle n'est appliquée sur les molécules d'eau, l'absence d'hydrogène dans une grande partie des structures rend l'approximation de leur position ardue.

b. Liaisons halogènes

Seuls les halogènes lourds, soit le chlore, le brome et l'iode, ont été considérés pour ce type d'interaction, le fluor étant assimilé à un accepteur de liaison hydrogène. Une distance inférieure à la somme des rayons de Van der Waals plus 1.0Å est requis pour la considération d'une liaison halogène (soit 4,5Å pour une interaction iode – oxygène par exemple). L'accepteur de liaison halogène, plus communément appelé *base de Lewis*, doit présenter un moment polaire négatif, par exemple un doublet non liant.

Ces bases de Lewis sont en réalité identiques aux accepteurs de liaisons hydrogène. Dès lors, l'oxygène, le sulfure, l'azote ainsi que la surface des aromatiques pauvres en groupes électroattracteurs ont été définis comme potentiels accepteurs de liaisons halogènes. Leurs géométries d'interactions est orienté de telle sorte à ce que le doublet non liant soit orienté

vers l'halogène, soit un angle sphérique compris entre 150° et 180° pour un cycle aromatique par exemple. Une partie de la surface d'interaction de l'halogène, le *σ-hole* en anglais, dépourvue d'électrons nécessite un angle sphérique linéaire par rapport à sa liaison covalente pour interagir. Un intervalle autour de 160° ± 20° a été considéré pour ce groupe d'atomes lourds, sachant qu'une déviation de 30° par rapport à un arrangement linéaire réduit théoriquement de moitié la force de l'interaction [95].

c. Autres types d'interactions

Le processus de détection d'interactions permet également de mettre en évidence des interactions impliquant des métaux. Ces métaux 'interactifs', appelés métaux de coordination, n'ont pas de valeur d'angle prédéfinie car Pipeline Pilot ne permet que partiellement la détermination correcte de ces liaisons ioniques métalliques. La caractérisation de ce type d'interaction nécessite toutefois une distance inférieure à 2.5Å ainsi qu'un accepteur de liaison hydrogène dirigé vers cet atome [96, 97].

Les interactions impliquant des groupements chargés, positivement et négativement, sont aussi calculées et affichés lors de la visualisation des interactions. Ces interactions, notamment les ponts salins (liaison hydrogène chargée et interaction ionique), impliquent un élément portant une charge distante d'un autre élément portant une charge partielle de nature opposée. L'approximation de la position de la charge est dépendante du nombre de liaisons covalentes présentes sur chaque atome et est déterminé de manière identique aux liaisons hydrogènes. Les interactions cation - π font aussi parti de cette catégorie et sont explicités dans une partie ultérieure.

Des contacts dits polaires sont aussi pris en compte dans la visualisation. Les carbones présentant une charge de Gasteiger dont la valeur absolue est supérieure à 0.20 est définit comme accepteur ou donneur de contacts polaires. Le carbone du groupement amide du squelette peptidique est par exemple identifié comme accepteur d'interactions polaires par son moment dipolaire présent entre l'oxygène, le carbone et l'azote. Ces interactions, généralement étudiés à travers les interactions carbonyle – carbonyle, se font dans ce cas avec toute atome présentant un doublet non liant ou une charge partielle négative. La surface d'interaction définie pour ce type d'interactions est dépendante du moment dipolaire, pour le carbone de l'amide, la charge partielle est approximée comme étant au-dessus du plan de

l'amide, vers le carbone (voir Figure 46). Toutes liaisons hydrogènes dont la distance est comprise entre 3,5Å et le seuil de distance (somme des rayons de Van der Waals + 1,0Å) furent aussi traités comme contacts polaires.

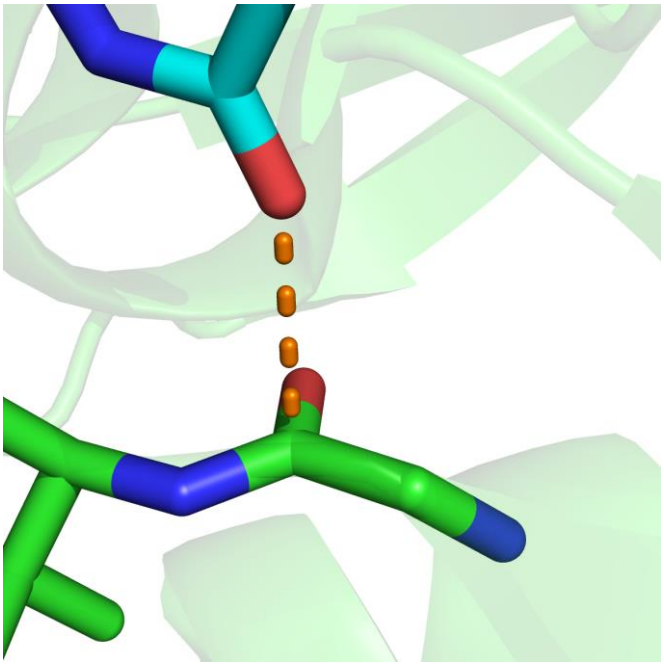


Figure 46 Contact polaire identifié entre le moment dipolaire d'un carbone du squelette peptidique (glycine 455) et le groupement carbonyle du ligand de l'adenylosuccinate synthetase (code PDB 1lny).

d. Contacts polaires non considérés

Les ponts aqueux ne sont à l'heure pas pris en compte dans la visualisation des interactions. Ces derniers nécessitent le recalcul des contacts effectués par chaque molécule d'eau en interaction avec le ligand. Pour des contraintes computationnelles et de temps, ce type d'interaction sera implémenté dans un futur proche. Les interactions longues distances ne sont pas non plus pris en compte de par le seuil de distance imposé initialement (somme des rayons de Van der Waals + 1,0Å). Agrandir le seuil de distance pour ce type d'interaction nécessiterait d'identifier l'absence d'atome dans la trajectoire séparant les deux atomes se faisant face. Or cette étape est très coûteuse en termes de performances informatiques et ne permettrait le recensement que d'un faible nombre d'interaction de ce type. De plus, les cycles aliphatiques sont considérés de la même manière que les groupements aliphatiques. Or, il serait intéressant de les considérer distinctement [98].

B. Contacts hydrophobes

Les contacts hydrophobes sont aussi calculés et affichés. Tout contact est considéré comme hydrophobe dès lors qu'un atome de carbone non ou faiblement polarisé est impliqué dans le contact. Des groupements méthyles -RCH_3 ou méthylènes -R_2CH_2 par exemple font partis des éléments hydrophobes fréquemment rencontrés dans les contacts hydrophobes. L'orientation théorique des hydrogènes implicites détermine la géométrie de ces contacts. Par conséquence, un groupement riche en hydrogènes comme le méthyle présentera un angle dont la valeur seuil minimale sera proche de 90°. A contrario, un groupement méthylène présent sur une longue chaine aliphatique sera limité à un angle de 120°. Les cycles aromatiques peuvent aussi être impliqués dans des contacts hydrophobes si le contact est réalisé avec la périphérie du cycle.

C. Interactions aromatiques

a. *π-stacking*

Les interactions impliquant deux aromatiques sont distinguées en fonction de l'orientation et de la position de chaque aromatique l'un par rapport à l'autre. Les angles dit ligand et récepteur décrivent ainsi la position relative d'un cycle aromatique par rapport à l'autre, une valeur de 180° correspondant à un alignement du centre de masse d'un cycle aromatique sur le vecteur normal de l'autre. Cependant, si l'un des angles aromatiques est linéaire, l'orientation l'autre cycle peut être mené à des interactions de nature distinctes. Pour rappel, un cycle aromatique tel le benzène présente une densité électronique importante de part et d'autre de son plan mais une périphérie dépeuplée d'électrons.

Ainsi, l'angle d'orientation va décrire l'arrangement entre deux aromatiques. Un angle de 180° indique deux plans se faisant mutuellement face tandis qu'une valeur de 90° correspond à deux aromatiques dans un arrangement dit en forme de 'T', en anglais *T-shape*. Les valeurs entre ces deux extrema montrent différentes inclinaisons de l'un des aromatiques. A partir de ces différentes valeurs d'angle, des configurations caractéristiques ont été mises en place.

Deux aromatiques en conformation dites *sandwich* sont associés généralement à deux plans se faisant parfaitement faces (voir Figure 47A), soit un angle sphérique supérieur à 165° ainsi qu'un angle d'orientation supérieur à 165° (Tableau 2). A l'opposé, les conformations dites *T*-

shape sont caractéristiques d'une forme de 'T' où le bord externe d'un des aromatiques interagit avec la surface de l'autre (Figure 47F). Cet arrangement se traduit par l'un des angles sphériques supérieur à 160° et un angle d'orientation proche de 90° (± 15°) (Tableau 2). Un arrangement appelé parallèle décalé, en anglais *parallel-displaced*, est caractérisé par un angle d'orientation supérieur à 160° et un angle sphérique entre 140° et 160°. Une configuration proche de la configuration dite *sandwich* mais dont l'un des plans est incliné sera nommée *face-to-face* (110° < angle orientation < 150° et angle sphérique > 165°), illustré sur la Figure 47D. Une valeur d'angle sphérique inférieure à 140° sera décrite par une configuration *edge-to-face*. Enfin, cas rare, si les plans de chaque aromatique sont alignés l'un à côté de l'autre, soit un angle d'orientation proche de 180° et un des angles sphériques proche de 90° sera déterminé comme bord à bord, en anglais *edge-to-edge.*

	Configuration	Angle orientation	Conditions
Parallèle	***Sandwich***	≥ 165°	Angle ligand + récepteur ≥ 165°
	Parallel-displaced	≥ 165°	135° < Angle ligand/récepteur < 165°
	Edge-to-edge	≥ 165°	Angle ligand/récepteur ≤ 100°
Incliné	***Face-to-face***	105° < x < 165°	Angle ligand/récepteur ≥ 165°
	Edge-to-face	105° < x < 165°	140° < Angle ligand/récepteur < 165°
Forme T	***T-shape***	75° < x ≤ 105°	Angle ligand/récepteur ≥ 160°

Tableau 2 Tableau récapitulatif des différents arrangements aromatiques considérés ainsi que leur paramètre géométrique.

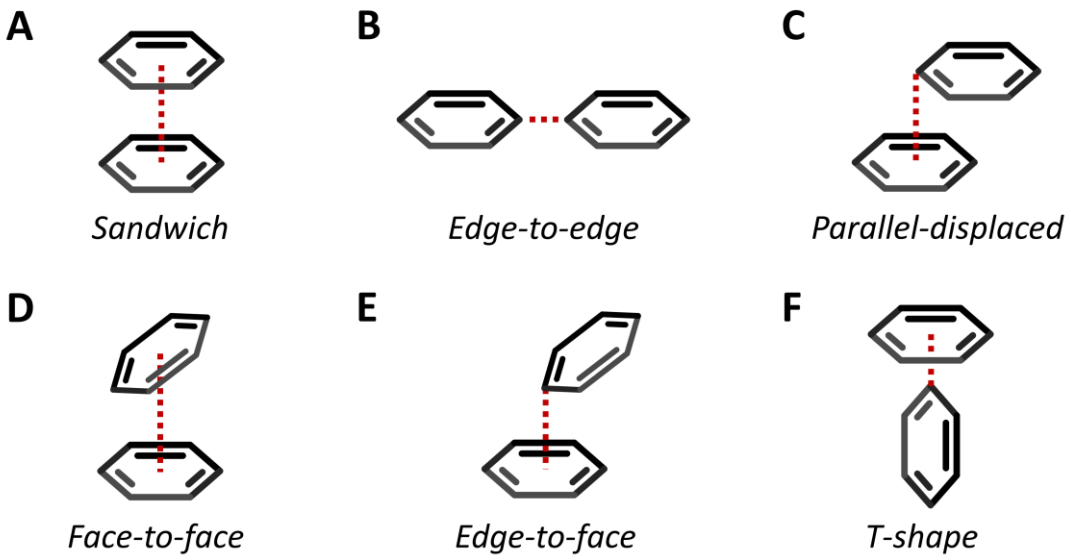


Figure 47 Schéma récapitulatif des interactions aromatiques considérées dans la base de données 3decision®.

Ces configurations sont à la fois validées par visualisation 3D des différents arrangements distingués mais aussi par le graphique des distances entre centres de masses (*offset*) quantifiés entre les deux aromatiques, représentées sur la Figure 48. Les arrangements *sandwich* et *face-to-face* requièrent deux aromatiques dont la distance (*offset*) entre leurs centres de masses soit proche de 0Å comme présent sur le graphique. De même, deux aromatiques l'un à côté de l'autre, *edge-to-edge*, doivent avoir des distances de décalage égales et supérieures à 4Å. Sur certaines zones de la Figure 48, différents types d'arrangements aromatiques sont distinguées sur des valeurs d'offset identiques, notamment pour des valeurs de distance de 2Å et 6Å. L'angle d'orientation entre aromatiques permet de distinguer les différentes conformations entre *T-shape* et *edge-to-face* dans cet exemple.

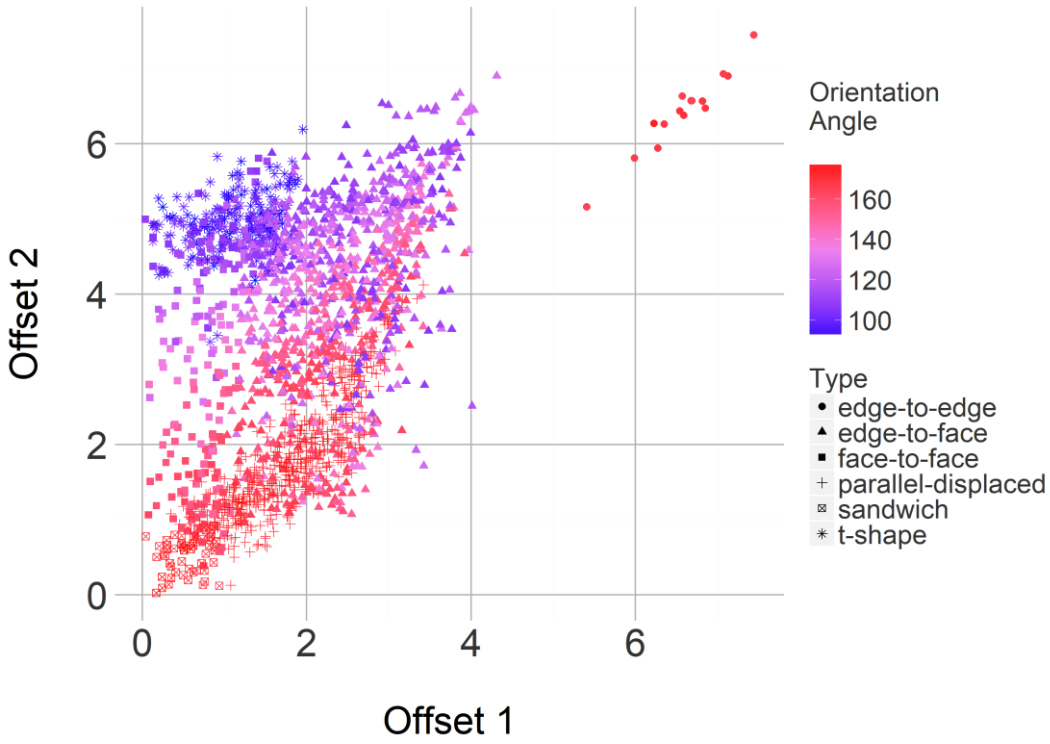


Figure 48 Distribution des distances séparant les centres de masses d'aromatiques, 2000 distances (offset en Å) ont été mesurées. Leur classification est représentée distinctement par leur forme alors que l'angle d'orientation est défini par un gradient de couleur (bleu : orientation orthogonale, rouge : aromatiques parallèles).

Les autres contacts aromatiques restent présents dans la base, pouvant être considérés comme des contacts hydrophobes ou de potentiels accepteurs de liaisons hydrogènes

b. Cation - aromatique

Les interactions cation – aromatique, ou cation – π, sont aussi considérées lors du processus de visualisation. Toute interaction impliquant un cation dirigé vers une surface aromatique apparaitra lors du processus de visualisation. Ces interactions se traduisent par un angle sphérique caractérisé par l'angle aromatique proche de 180° (± 30°), soit un positionnement linéaire par rapport au vecteur normal de l'aromatique. La définition géométrique des cations dépend de leur appartenance à une molécule ou non. Des ions dits libres n'auront ainsi pas de seuil géométrique appliqué, à l'inverse de cation présent dans une molécule où des configurations linéaires sont favorisées (180° ± 30°).

D. Comparaison avec *Protein Ligand Interaction Profiler*

Malgré des définitions d'interaction relativement proches de celles utilisés par PLIP [64], des différences notables sont à signaler. L'exemple illustré sur la Figure 49A, un complexe de protéine tyrosine kinase 2 (domaine JH1), n'affichait que 2 interactions protéine - ligand (une liaison hydrogène et une interaction π-cation) après extraction des résultats de PLIP. L'affichage des interactions moléculaires selon nos critères résulte dans une visualisation plus représentative du mécanisme d'interaction protéine - ligand, illustrée sur la Figure 49B. L'absence de l'interaction π-cation de notre affichage est due à nos critères de détection séparant les deux éléments. Cette interaction semble toutefois contestable de par la distance importante entre les atomes, 5,5Å dans le cas présent. Il est important de noter que cette structure dispose d'une résolution modérée (2,3Å) dont la carte de densité suggère une position approximative du ligand dans son site de liaison. Néanmoins, cette observation ne devrait pas influencer le nombre ou la nature des interactions détectées.

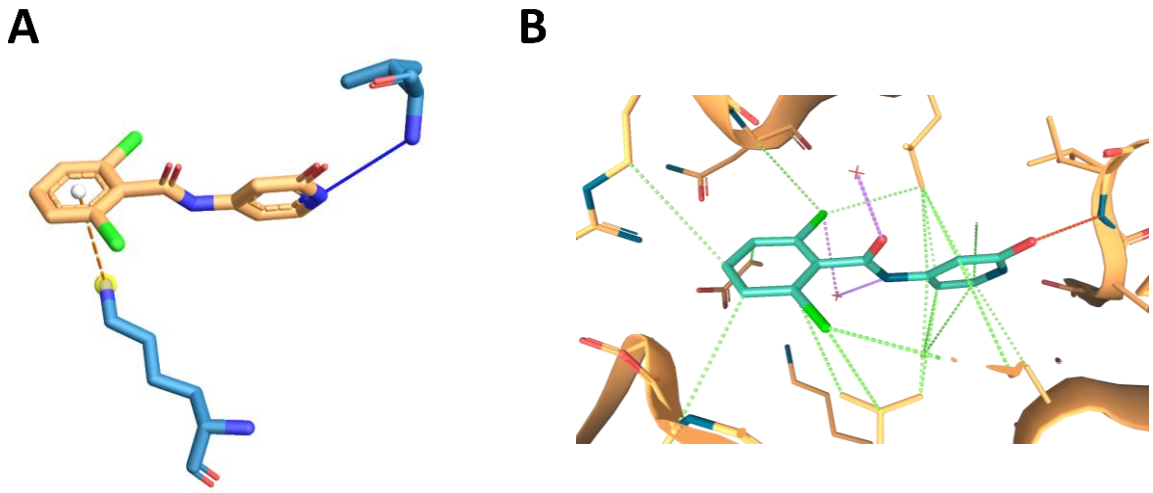


Figure 49 Détection et visualisation des interactions moléculaires d'une tyrosine kinase 2 (code PDB 4gfo) issus de A. PLIP [64] (π-cation en orange, liaison hydrogène en bleue) B. 3decision® (contacts hydrophobes en vert, contact polaire en orange, liaisons hydrogènes avec une molécule d'eau en violet).

Chapitre 4 : Superposition de contacts

A l'aide des contacts calculés sur l'ensemble de la base de données, l'idée de proposer des solutions d'optimisation médicamenteuse aux chercheurs s'est traduit par le développement d'une méthode d'alignement de ces contacts. Ce projet repose sur la capacité à pouvoir guider l'utilisateur dans le développement d'un médicament en indiquant les régions de contacts les plus denses autour d'un atome dans 3decision®. Pour cela, il devait être possible de savoir pour un acide aminé ou une partie d'un acide aminé quels sont les types d'atomes sur un potentiel ligand les plus fréquents dans l'environnement proche de ces atomes et où ils se trouvent. Une manière d'amener ce type d'information est de projeter une grille de densité sur l'atome d'intérêt de la protéine ou du ligand (en cas d'analyse basée sur le ligand) par l'intermédiaire de superposition d'atomes d'interaction par rapport à un atome de référence. Les approches développées pour parvenir à réaliser ces types d'analyses sont détaillées dans ce chapitre. Les différents calculs de repère orthonormés sont réalisés lors de l'enregistrement d'une structure.

A. Développement d'une méthode d'alignement de contacts sur un atome

L'alignement et superposition de deux éléments provenant de coordonnées cartésiennes 3D distincts nécessitent l'identification de repères communs afin de –généralement- calculer une matrice de rotation. Le calcul de cette matrice est un processus nécessitant de passer par un processus distinct dont l'exécution nécessite un certain temps. Ici, une méthode alternative a été envisagée. Cette méthode rapide repose sur l'identification de repères normées permettant la création d'un repère tridimensionnel local, une retransformation du repère local en coordonnées cartésiennes permettrait alors la superposition.

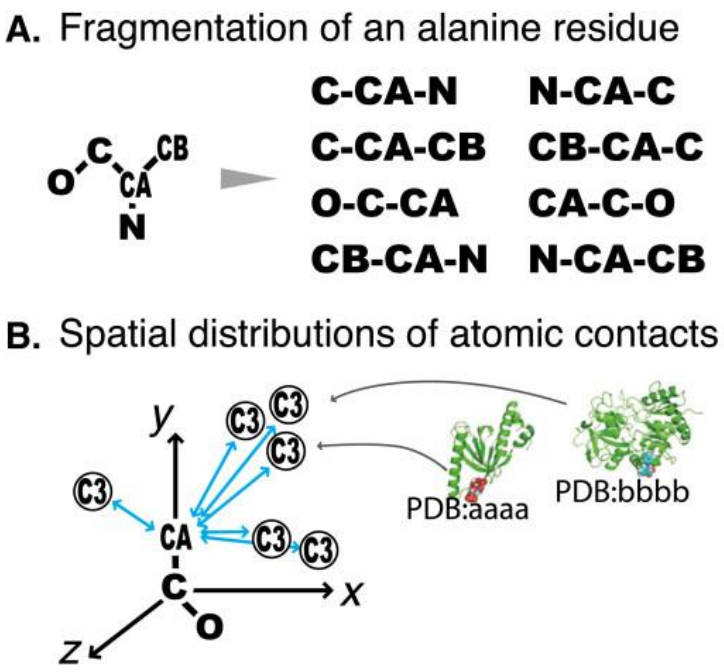


Figure 50 Approche de repère local utilisée par Kasahara et collaborateurs [99] pour définir un repère orthonormé normalisé pour chaque atome du récepteur. A. Dictionnaire prédéfini de fragments issu du récepteur B. Création du repère spatial et analyse de la distribution spatiale des atomes du ligands.

La méthode a été élaborée en deux temps. La première approche s'est faite en s'inspirant de la méthodologie proposée par Kasahara et collaborateurs [99, 100]. Cette approche consiste à élaborer des repères tridimensionnels locaux par l'intermédiaire de fragments de 3 atomes, tous liés de manière covalente, dont l'atome central représentera l'origine du système local (voir Figure 50).

Le développement de notre première approche permet la création de repères normalisés autour d'acides aminés. En effet, la structure répétée et connue des 20 acides aminés les plus courants résulte dans la prédéfinition de repères communs permettant d'établir ce repère normé sur chaque atome des acides aminés. Ce principe ne peut pas directement être appliqué sur un ligand car l'espace chimique représenté est beaucoup trop importante pour permettre la prédéfinition de ces repères.

Dans notre approche, l'atome du récepteur est défini comme l'origine du système local, correspondant aux coordonnées (0, 0, 0), l'atome d'azote sur la Figure 51A. Un dictionnaire comportant deux atomes du récepteur sélectionnés arbitrairement (deux atomes liés de manières covalentes à l'atome du récepteur par exemple) est utilisé dans le but d'assister la définition des vecteurs directeurs. Le vecteur directeur *x* (1, 0, 0) fut définit dans un premier

temps comme le vecteur entre l'origine et un des atomes du récepteur issu du dictionnaire, schématisé le Cα sur la Figure 51A. Le vecteur directionnel *y* (0, 1, 0), dont la sphère d'arrivée est affichée en rouge sur la Figure 51A, correspond au produit scalaire entre le vecteur directeur x et le vecteur défini entre l'origine (0, 0, 0) et le second atome du dictionnaire. Enfin, le produit scalaire des deux vecteurs directionnels *x* et *y* résultent dans le vecteur directeur *z* (0, 0, 1), schématisée par la sphère d'arrivée violette sur la Figure 51A.

A partir de ces vecteurs directionnels, trois angles peuvent être calculés entre le vecteur récepteur-ligand et chaque vecteur directionnel. Ces angles, nommé θ (*thêta*) 1 à 3, permettent ensuite de déterminer la position locale de l'atome d'interaction par rapport à ce vecteur (voir Figure 51B). Cette méthode a été utilisée pendant l'intégralité des travaux de thèse.

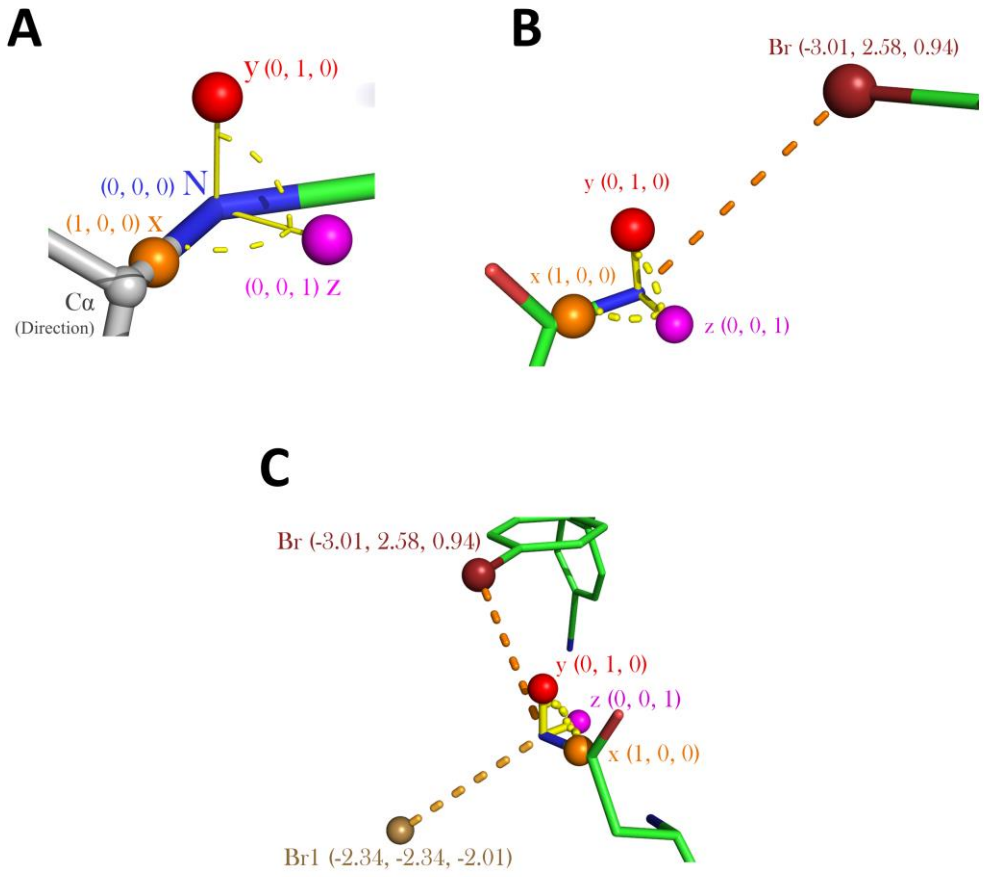


Figure 51 Illustration 3D de l'élaboration d'un repère local autour d'un acide aminé A. Création d'un repère spatial 3D généré selon la méthode de Kasahara utilisé dans 3decision® (coordonnées locales entre parenthèses) B. Exemple de coordonnées locales obtenus sur un atome de Brome (colorié en bordeaux, code PDB 5i74, ligand 69D, résidu ASN 177 de la chaine A). C. Transposition de coordonnées locales provenant d'une autre structure (code PDB 3bm9, ligand BXZ, initialement en contact avec le résidu ASN 51 de la chaine A) sur le système décrit sur l'image B.

Les coordonnées locales peuvent ensuite être retranscrites en coordonnées globales sans passer par une étape de calcul de rotation de matrice. Plusieurs méthodes peuvent être

utilisées pour effectuer cette transposition, compte tenu des limitations de Pipeline Pilot, la résolution d'une équation à 3 inconnues a été utilisé :

$$\begin{cases} a_{vx}x + b_{vx}y + c_{vx}z = x_{local} \\ a_{vy}x + b_{vy}y + c_{vy}z = y_{local} \\ a_{vz}x + b_{vz}y + c_{vz}z = z_{local} \end{cases}$$

avec (x, y, z) les coordonnées 3D transformées sur le nouveau repère, $(x_{local}, y_{local}, z_{local})$ les coordonnées sur le repère 3D local et (a_{vx}, b_{vx}, c_{vx}) le vecteur directeur x.

Une transposition par rapport à l'atome du récepteur permet par la suite de retrouver nos coordonnées retranscrites dans le nouveau repère cartésien comme illustré sur la X. Ainsi, la superposition de deux interactions nécessite qu'une requête dans la base de données et la résolution d'une équation à 3 inconnues selon la méthode du pivot de Gauss [101].

B. Amélioration de la création de repères locaux

Cette première approche, bien qu'efficace dans la translation de coordonnées provenant de repères distincts est (i) applicable uniquement aux acides aminés et (ii) ne permet pas une interprétation intuitive des coordonnées locales. En effet, le deuxième point fut rapidement corrigé lors de l'étape d'identification des repères.

Ainsi, une nouvelle approche a été développé dans 3decision® et fait l'objet actuellement d'une intégration dans l'interface. Le vecteur directeur utilisé pour le calcul des angles d'élévation fut tout d'abord définit comme le vecteur directeur z du repère local (0, 0, 1) (voir Partie 2.2.C.b). Ce vecteur peut être interprété comme la *profondeur* du système et fait face à la surface accessible de l'atome. L'identification des repères permettant de définir les vecteurs directeurs *x* et *y*, la longitude et latitude respectivement s'est fait différemment selon un acide aminé ou un ligand. Ce problème peut être assimiler à une assignation des points cardinaux sur une sphère cartographique : une fois le nord déterminé sur une sphère, par rapport à quel point orienté l'est et l'ouest ? Dans le cas des acides aminés, plus simple, le vecteur directeur x est toujours orienté vers le carbone sur une représentation 2D (voir Figure 52). Le vecteur directeur *y* (0, 1, 0) n'est que le produit scalaire des vecteurs directeurs *x* et *z* établis précédemment.

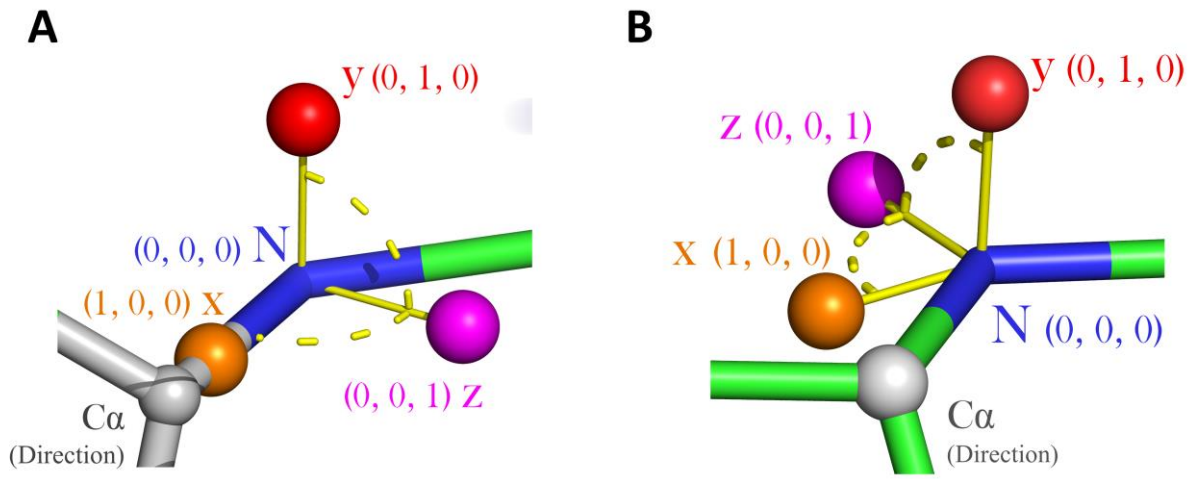
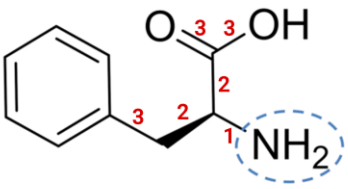


Figure 52 Exemple de repères locaux créés autour de l'azote du squelette peptidique selon A. la première approche inspirée de Kasahara B. la méthode améliorée.

En ce qui concerne le ligand, la détermination de l'orientation du vecteur *x* se fait par une autre approche consistant à identifier des asymétries de sous-structures. Dans cette méthode, l'atome d'interaction constitue toujours l'origine du système local (0, 0, 0). Le vecteur directionnel *z* est donc colinéaire au vecteur définit pour le calcul des angles d'élévation et débute au niveau de l'origine (0, 0, 0) (voir Figure 52B).

Dès lors, une recherche d'asymétrie est effectuée dans le ligand à partir de l'atome d'interaction. Cette recherche est effectuée uniquement pour les atomes à une ou deux liaisons covalentes. Elle repose sur le principe qu'une molécule peut être assimilée à un graphe et est donc déterminé par un nombre défini de liaisons covalentes. Ainsi, par itération le long des liaisons covalentes, la recherche d'asymétrie est déterminée par (i) un nombre différent de liaisons covalentes (ii) des éléments de nature différente. L'exemple présent sur la Figure 53 illustre un cas concret où l'asymétrie est déterminée au niveau d'une liaison d'ordre 3, un des carbones n'étant lié qu'à un seul atome, faisant apparaitre une asymétrie.



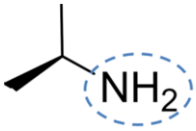
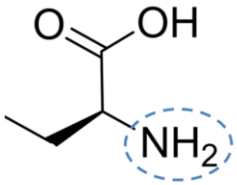
Bond Path		
1	2	3
		

Figure 53 Détermination de l'asymétrie moléculaire. La recherche d'asymétrie se fait itérativement le long des liaisons covalentes selon des règles spécifiques.

Le point référent de l'asymétrie permet alors de déterminer la direction du vecteur directionnel *x* (1, 0, 0) dont les coordonnées locales détermineront la latitude de l'atome en question (gauche et droite, étape 2 de la Figure 54). Enfin, le vecteur directionnel *y* (0, 1, 0) issu du produit vectoriel des vecteurs *x* et *z* représentera la position longitudinale, correspondant au repère 'haut-bas' de notre système local (étape 3 de la Figure 54).

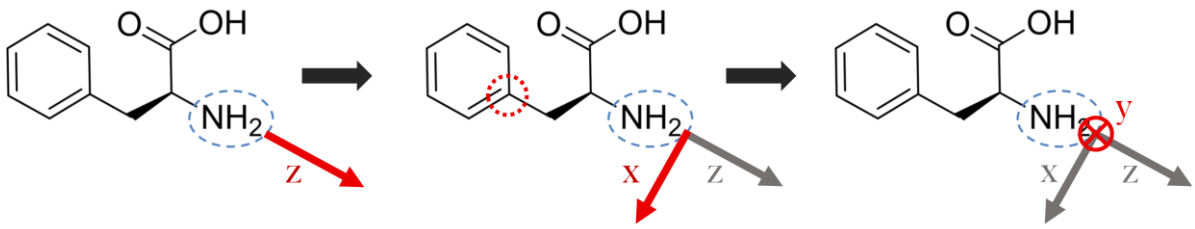


Figure 54 Schéma de la définition du nouveau repère local tridimensionnel selon la nouvelle approche.

Ainsi, les angles définis par le vecteur récepteur-ligand et les vecteurs *x* et *y* permettent de définir ainsi l'élévation et l'azimut, principes populaires notamment dans le domaine de l'astronomie.

C. Création de grilles de densité

La transposition de n'importe quel élément provenant d'un repère tridimensionnel distinct autour d'un atome permet l'affichage des grilles de densité dont les éléments les plus présents se traduiront par une densité plus importante. L'utilisation de ces grilles a pour objectif final de suggérer aux chercheurs des pistes exploratoires d'optimisation du ligand par l'ajout de nouveaux atomes à partir des contacts recensés dans la base. L'affichage de ces grilles peut être fait dans une cavité comportant un ligand, mais aussi en l'absence de ligand étant donné que les repères peuvent être normalisés sur chaque résidu.

Le ligand présent dans la structure de la protéine HSP90α (code PDB 1uyd) ne présente pas une affinité satisfaisante, IC_{50} supérieur à 200µM, et nécessite donc une phase d'optimisation [102]. Sur l'exemple présenté en Figure 55, les contacts impliquant un carbone du ligand hybridé sp^2 ont été transposé sous forme de grille de densité sur chacun des résidus du site de liaison de la protéine HSP90α (en bleu-gris). La grille de densité, dans ses valeurs les plus élevées, désigne une région spatiale où un nombre important de contacts impliquant un carbone sp2 ont été identifiés. Dans l'exemple, cette fréquence importante se traduit au niveau des groupements méthoxyles du ligand PU8. Cette observation peut donc laisser imaginer qu'un cycle aromatique fusionné puisse potentiellement se situer dans cette région. Il se trouve que cette suggestion est vérifiée de manière expérimentale : une autre structure de la protéine HSP90α (code PDB 4cwt, ligand IK9) contient un ligand comportant un cycle aromatique fusionné à cette position, représenté en orange sur la Figure 51. Casale et collaborateurs indiquent une constante d'inhibition (K_i) de 0,17µM [103].

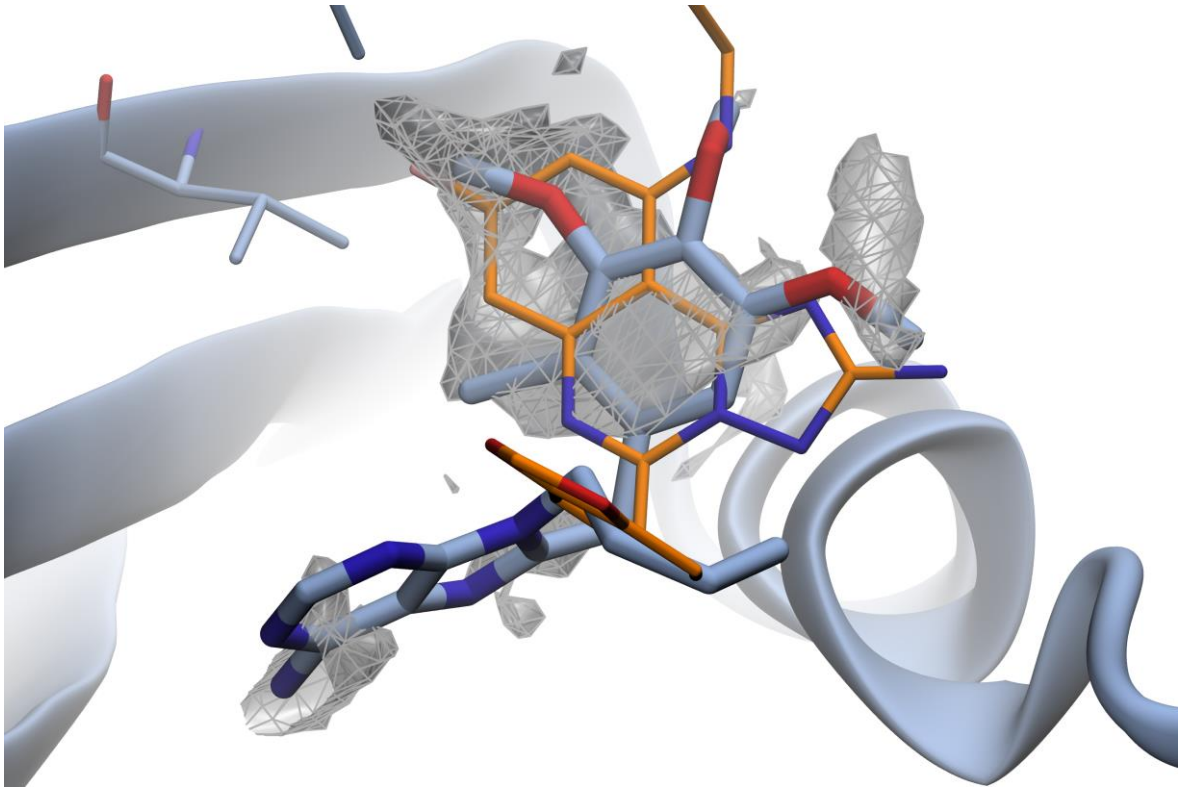


Figure 55 Illustration d'une grille de densité sur le site de liaison à l'ATP de la protéine HSP90α comportant le ligand PU8 (code PDB 1uyd, en bleu gris). La grille de densité représente les contacts observés dans la base de données impliquant un carbone sp2 du ligand transposé sur chaque résidu du site de liaison. La superposition du site de liaison de la structure 4cwt indique la présence du ligand aminotriazoloquinazoline (code résidu IK9) sur la région dont la grille suggérait une présence d'atome sp2 (image réalisée sur VMD).

Ces grilles, réalisées dans un premier temps par Pipeline Pilot puis en Python ensuite, peuvent aussi être affichés dans d'autres outils de visualisation 3D comme VMD [104], Chimera [105] ou PyMOL [106]. Les fonctions permettant le calcul de repère orthonormé, de transposition de coordonnées locales et de création de grilles tridimensionnels seront disponibles prochainement par l'intermédiaire de script python à l'adresse suivante : https://github.com/Discngine.

Chapitre 5 : Comparaison de contacts

La comparaison de complexes se fait généralement soit par la comparaison de ligand (*ligand-based*) soit par des approches centrées sur le récepteur (*receptor-based*). Ces deux approches sont très utiles pour comparer deux molécules sur le même site ou une même molécule sur deux protéines différentes. Toutefois, la comparaison d'interactions entre deux complexes est une étape importante permettant d'identifier de potentiels liaisons sur d'autres sites dits *off-*

target en anglais. 3decision® disposant déjà d'algorithmes distincts de comparaison de ligand et de poches, des approches ont été explorés se focalisant sur les contacts. La majorité des méthodes de comparaisons décrites dans la Partie 1.3, à l'exception de *Grim*, permet une comparaison quantitative des interactions protéine - ligand. La grande partie d'entre eux dépendent notamment d'un élément en commun, à savoir le récepteur, comme les méthodes *SIFt* et *CREDO* dont la construction du *fingerprint* dépend de la séquence en acide aminé. Or les récentes études de Ballester et collaborateurs [75] et Lenselink et collaborateurs [77] ont montré les limites de ces méthodes et du typage des interactions moléculaires telles que les liaisons hydrogènes. Les deux méthodes exposées ici ont été élaboré de manière exploratoire.

A. Approche fragmentaire

Une approche fragmentaire a été exploré dans un premier temps avant toute implémentation de contacts dans la base de données. Cette approche fut initialement testée sur les données fournit par Lenselink et collaborateurs [77]. Ce jeu de données est constitué de poses de docking générés sur 5 cibles de la famille protéine GPCR. 100 ligands distincts (selon le *fingerprint* FCFP4) récupérés sur ChEMBL et 5000 leurres générés par *DUD-E* [107]furent *dockés* via GLIDE [108] sur chacune des cibles (codes PDB 4eiy, 3rze, 3uon, 2rh1, et 4mbs). Les leurres produits par DUD-E consistent à réarranger la topologie des propriétés physico-chimiques de molécules actives. L'évaluation de la méthode se fait par le calcul du BEDROC, qui consiste à attribuer des pondérations plus fortes pour les complexes ayant le *pattern* d'interaction le plus proche de la structure initiale dans le calcul de la courbe ROC. Plus grande sera la valeur du BEDROC, meilleure sera la méthode à classer les molécules comme actives parmi les plus similaires en théorie.

Il est important de noter que cette méthode d'évaluation dispose de certains biais. Rien ne garantit que les leurres générés artificiellement par redistribution topologiques des groupements fonctionnels soient des molécules inactives. De plus, il est tout à fait possible qu'une molécule comportant un *pattern* d'interaction distinct puisse se lier sur le même site de liaison qu'une molécule active. Toutefois, il s'agissait au moment de l'expérimentation de la seule étude avancée de comparaison d'interaction.

Lors de l'élaboration de cette méthode, une définition différente des contacts fut utilisée (étape 1 de la Figure 56) comparé à la définition décrite dans les chapitres précédents. Ainsi,

un contact est défini comme deux fragments dont au moins deux atomes sont distants d'une valeur inférieure à la somme de Van der Waals + 1,5Å. Contrairement aux contacts présents dans la table d'interactions, à la fois les éléments pris en compte dans la définition du contact sont les fragments du récepteur et du ligand issus de la fragmentation moléculaire (voir Partie 2.2.C.a). D'autres caractéristiques ont été répertoriés pour dans la définition de chaque contact fragmentaire (étape 2 de la Figure 56) : (i) les deux éléments impliqués dans le contact avec la plus courte distance, (ii) la distance de ce contact court, (iii) la distance entre les centres de masses de chaque fragment, (iv) l'orientation angulaire entre les hyperplans de chaque fragments (entre les deux vecteurs normaux), (v) des *fingerprints* moléculaires descriptifs de chaque fragment, et (vi) le nombre total de contacts détectés.

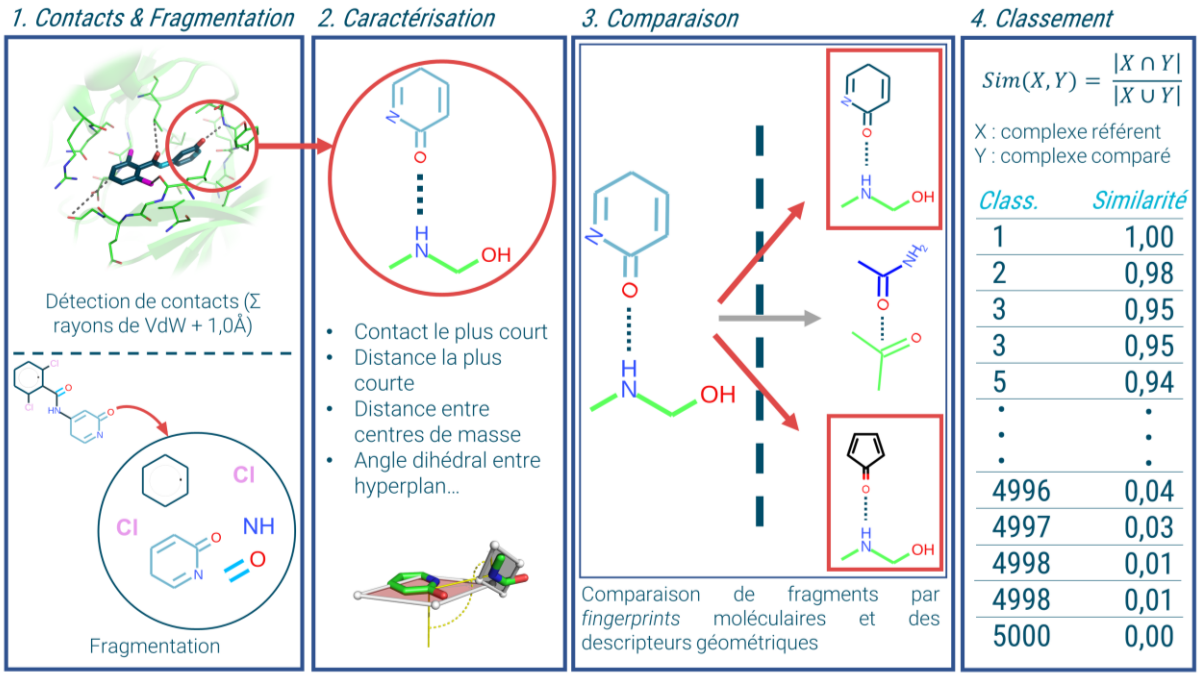


Figure 56 Schéma récapitulatif des étapes du protocole de comparaison fragmentaire.

La comparaison entre deux complexes fut réalisée par l'identification de contacts fragmentaires considérés comme identiques selon les différentes caractéristiques définies (étape 3 de la Figure 56). Le score final de similitude entre deux complexes fut calculé sur le rapport entre le nombre de contacts fragmentaires identiques et le nombre de contacts dans le complexe de référence (étape 4 de la Figure 56). Cette méthode, bien que proche de la méthode SPLIF, se différencie notamment par la notion de similitude et non l'identité exacte dans la comparaison de fragment. De plus, l'utilisation de critères géométriques dans notre

méthode en lieu d'un alignement structural et comparaison par RMSD permet d'obtenir des résultats plus nuancés.

Les résultats obtenus, notamment présentés lors de la *German Chemoinformatics Conference* 2016, furent peu concluants (voir Figure 57). En effet, l'un des principaux obstacles rencontrés reposait sur la corrélation entre méthode de fragmentation, caractérisation des fragments en *fingerprints* moléculaires et métriques de comparaison. Différents paramètres de fragmentation (conservation ou non des cycles fusionnés par exemple) ainsi que différents *fingerprints* moléculaires (ECFP2, FCFP2, FCFP4…) furent testés de manière extensive mais ne permettaient pas d'obtenir des résultats consistants à travers les 5 structures. La structure de l'histamine H1 (code PDB 3rze) comporte notamment un large cycle fusionné composé de 2 aromatiques et d'un cycle aliphatique peu présent sur les molécules actives fournies dans le jeu de données. Les résultats les plus consistants, annotés 'Fragments' sur la Figure 57, sont issus des paramètres de comparaison suivants : description du fragment par FCFP4 et deux contacts considérés comme similaires lorsque coefficient de Tanimoto > 0.5 entre les deux fragments, une différence d'angle planaire < 15.0° et une différence de distance minimale < 0.5Å. Ces problématiques sont aussi retrouvés dans SPLIF qui considère une interaction comme identique que lorsque deux couples de fragments sont identiques. Un cycle aromatique à 5 atomes est-il identique voire similaire à un cycle aromatique à 6 atomes ? A quel point un azote accepteur de liaison hydrogène contribue-t-il au mécanisme de liaison par rapport à un carbonyle ? Des approches prédéfinissant les interactions moléculaires répondent partiellement à cette question.

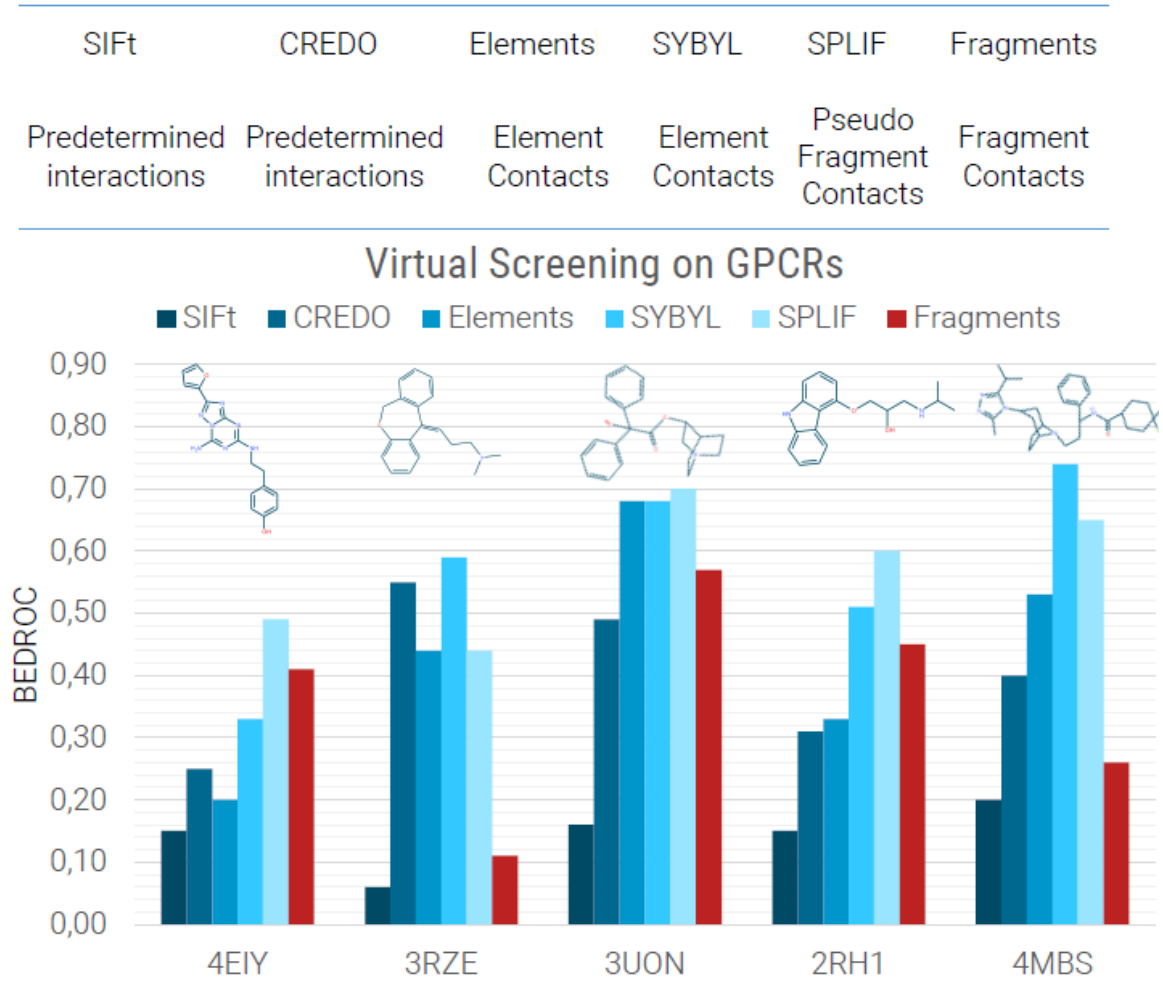


Figure 57 Résultats obtenus lors la comparaison de complexes issus de Virtual Screening sur des cibles GPCR [Lenselink2016]. Les méthodes SIFt [67] et CREDO [63] repose sur des définitions prédéfinies d'interactions (tels que liaison hydrogène, π-stacking). Elements et SYBYL [75] considèrent des contacts selon un critère de distance indépendamment de la nature des éléments mis en jeu. SPLIF [74] et notre méthode, Fragments, décrivent les interactions par deux fragments proches.

B. Approche par les contacts

Une méthode centrée sur la comparaison de contacts interatomiques fut ensuite explorée sur le même jeu de données. Cette approche, similaire aux approches *Elements* et *Sybyl*, repose sur un critère de distance, somme des rayons de Van der Waals + 1.5Å, séparant deux éléments, indépendamment de leur structure. La description de chaque contact interatomique fut caractérisée par (i) la nature des éléments impliquées, (ii) leur hybridation, (iii) les éléments liés de manière covalente à chaque atome, et (iv) les paramètres géométriques tels que la distance et l'angle sphérique. Tous les contacts d'un complexe référent furent ainsi comparés aux contacts d'un autre complexe, des valeurs seuils appliqués

sur les descriptifs précédemment cités permettent d'identifier les contacts comme similaires. Le score final entre deux complexes correspond au nombre de contacts similaires sur le nombre de contacts référents.

L'application de cette méthode sur les données fournies par Lenselink et collaborateurs a donné des résultats encourageants (catégorie 'Contacts' sur la Figure 58). En effet, les résultats indiquent des comparaisons consistantes dans la capacité à retrouver des complexes similaires sur les 5 protéines étudiées, les résultats les moins bons sur le récepteur de l'adénosine (code PDB 4eiy, BEDROC = 0,37) sont observées aussi sur les autres méthodes. La valeur moyenne des BEDROCs, 0,56 ici, classe cette approche au niveau des meilleurs méthodes, *SPLIF* (0,58) et *SYBYL* (0,57).

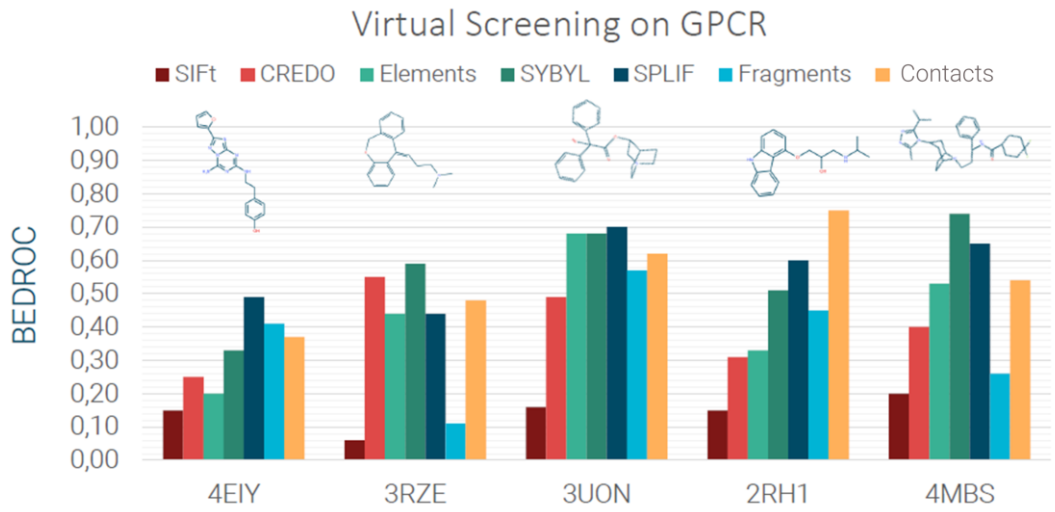


Figure 58 Nouveaux résultats obtenus par lors la comparaison de complexes issus de Virtual Screening sur des cibles GPCR [Lenselink2016]. La méthode 'Contacts' prend en compte avant tout un critère de distance séparant les deux éléments dans la définition des contacts.

C. Perspective 3decision®

Ces deux approches n'ont à l'heure actuelle pas été implémenté dans 3decision® pour des raisons notamment de performances sur la base de données de contacts. En effet, une recherche et comparaison rapide nécessite (i) un formatage des données de la même manière que des *fingerprints* d'interaction type SIFt [67] (ii) ou un archivage des données dans la mémoire RAM de la base (procédé 'In-memory' chez Oracle), très performants mais dont le coût est proportionnel à la quantité de données. La première solution est en opposition avec une des finalités désirées, à savoir la superposition des complexes sur des *patterns* de liaisons

identiques. Ce procédé nécessite l'identification des éléments en commun, ici des interactions, afin d'établir la matrice de rotation. Or, la concaténation des données dans un *fingerprint* ne permet généralement que des comparaisons d'ordre quantitative.

Une autre problématique se pose dans un deuxième cas de figure. Dans le cas où un nombre *i* d'interactions d'un complexe 1 sont jugés comme identiques aux *j* interactions d'un complexe 2, il est nécessaire d'établir une correspondance entre couples d'interactions identiques.

La superposition de deux complexes sur des *patterns* d'interaction communs sans utilisation du ligand et du récepteur comme repère est une problématique comparable à une comparaison de formes 3D. La comparaison et alignement de deux formes 3D nécessite d'obtenir la meilleure correspondance entre des points spécifiques permettant ensuite l'obtention d'une matrice de rotation optimale. Des méthodes de réduction de données décrivant des volumes 3D comme *Ultra fast Shape Recognition* ont été développés [109]. L'utilisation de ces descripteurs volumétriques comme repères permettant un alignement structural s'est révélée non concluante sur cette approche de par le nombre important de contacts à prendre en compte ainsi que la sensibilité des distances considérées dans la méthode. Des approches utilisant notamment le type *Deep Learning* ont été exploré dans la reconnaissance de forme 3D ces dernières années et pourront être étudiées à l'avenir [110].

Partie 3 : Analyse des interactions autour des halogènes

Le recensement des contacts dans la base de données 3decision® a mené à un travail d'analyse sur les halogènes. En effet, leur introduction dans la conception médicamenteuse croit de manière constante ces dernières années. Bien qu'il soit fréquemment utilisé dans le but d'améliorer des propriétés ADME (Absorption, Distribution, Métabolisme et Excrétion) d'une molécule, leur rôle est trop souvent limité aux liaisons halogènes classiques et plus récemment en au rôle d'accepteur de liaisons hydrogènes. La description des interactions se fait très fréquemment en ne considérant que deux atomes. Ce travail d'analyse explore une description prenant en compte l'environnement atomique et électrostatique autour des halogènes dans le contexte protéine - ligand. Cet article fut soumis dans le *Journal of Medicinal Chemistry* (du groupe ACS) et est actuellement en cours de révision. Les résultats obtenus et décrits à travers l'article sont proposés dans les chapitres suivants.

Chapitre 1 : Contexte

A. Propriétés électrostatiques et chimiques

Les atomes dits halogènes sont les éléments appartenant au groupe 17 de la table périodique. Dans le contexte du *drug design*, seuls les éléments fluor (F), chlore (Cl), brome (Br) et iode (I) sont utilisés, l'astate étant trop lourd et radioactif. Ces éléments sont tous électronégatifs, et attirent donc les électrons vers eux sur des liaisons covalentes. Leur électronégativité est inversement proportionnelle à leur taille : le fluor qui a la plus petite taille est l'atome le plus électronégatif du groupe.

La particularité des halogènes réside dans la distribution anisotrope des électrons autour du noyau (voir Figure 16). Un halogène, toujours lié de manière covalente à un carbone, va voir ses électrons être délocalisés de manière orthogonale autour du noyau. Cette délocalisation laisse place à une région déficiente en électron dans le prolongement de la liaison covalente. Cette région, chargée positivement, est plus communément appelée région σ (en anglais *σ-hole*). L'intensité positive de cette charge est dépendante de la taille de l'halogène. Le fluor,

trop électronégatif, ne semble pas présenté de région σ au contraire de l'iode qui dispose du σ-*hole* le plus positif et le plus large. L'intensité de cette région dépend aussi du fragment auquel l'halogène est rattaché [111]. Plus ce fragment est électroattracteur, plus la délocalisation des électrons sera importante et le σ-*hole* sera large et intense.

B. Capacité interactive des halogènes

La présence de cette région électropositive attribue des propriétés d'interactions non covalentes pour les halogènes dit lourds que sont le chlore, le brome et l'iode. Ces propriétés attractives des halogènes ont été mis en évidence pour la première fois dans les années 1950 [112, 39]. Le terme « liaison halogène » est employé dans la littérature lorsqu'une entité électronégative, par exemple le doublet non liant d'un oxygène, et la région σ électropositive de l'halogène se font face. Par définition donc, le fluor n'est pas capable de produire une liaison halogène. L'estimation énergétique d'une liaison halogène est de l'ordre de 6 kcal/mol pour un iode, proche de celle d'une liaison hydrogène considérée comme forte [38].

En plus de l'influence énergétique induite selon l'halogène impliqué, des paramètres géométriques sont aussi à prendre en compte dans la considération de la force d'une liaison halogène (voir Figure 59). La liaison halogène est la plus forte dans une configuration linéaire entre l'halogène, son carbone covalent et son donneur de liaison halogène, soit un angle halogène de 180°. De même, la distance optimale pour une liaison halogène est estimée comme la somme des rayons de Van der Waals des deux éléments interactifs. Wilcken et collaborateurs ont estimé qu'une variation d'angle de 25° à 30° ou une variation de distance de 1,0Å contribue à une diminution de 50% de l'énergie d'interaction [95]. Aucune liaison halogène ne serait observée à un angle halogène inférieur à 140° selon cette même étude.

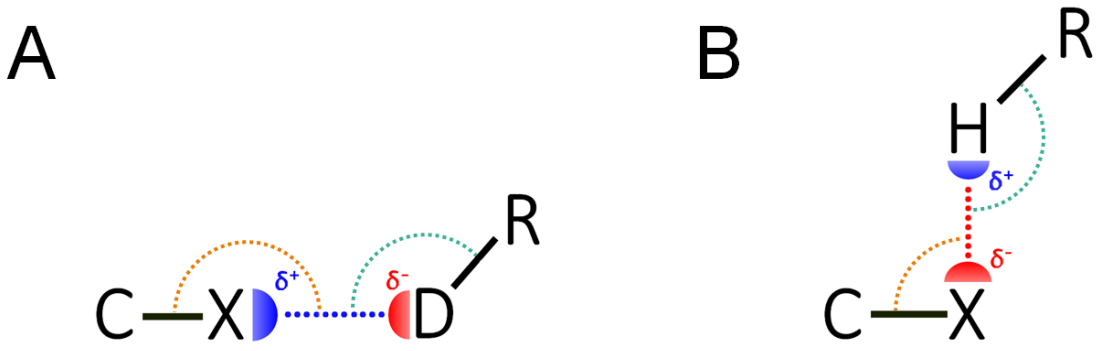


Figure 59 Représentation schématique d'interactions impliquant les halogènes A. Liaison halogène impliquant un halogène (X) et un accepteur de liaison halogène (D) et B. d'une liaison hydrogène entre un halogène et un donneur de liaison hydrogène (H). L'angle halogène est représenté en orange tandis qu'en vert est défini l'angle du résidu.

Mentionné pour la première fois par Murray-Rust en 1979 [113], la ceinture riche en électrons des halogènes lourds est aussi impliquée dans des interactions [114, 115, 116]. Ces interactions où l'halogène joue le rôle d'accepteurs d'électrons ont souvent été négligées, comme dans le travail d'Auffinger et collaborateurs en 2004 qui les caractérisant « d'interactions inhabituelles » [117]. En 2017, une étude extensive de Lin et collaborateurs évalue l'énergie d'interaction de ces liaisons hydrogènes impliquant un halogène lourd pouvant atteindre 14 kcal/mol dans des cas spécifiques. Ces valeurs sont supérieures aux valeurs décrites pour des liaisons halogènes.

Jusque-là majoritairement ignoré dans la considération des interactions moléculaires, un intérêt croissant commence à apparaitre pour le fluor. Depuis quelques années, le fluor est identifié dans certaines études comme accepteur faible de liaisons hydrogènes par sa forte électronégativité [118, 119]. L'énergie d'interaction estimée reste toutefois relativement faible, estimée à 1,7 kcal/mol [120]. Jusque-là majoritairement greffé aux molécules pour améliorer leurs propriétés ADME, le fluor a contribué à l'optimisation de l'affinité de certains composés. Rowlinson et collaborateurs [121] et Koch et collaborateurs [122] ont souligné l'impact de l'ajout d'atomes d'halogènes à des molécules ciblant la cyclo-oxygénase-2 et l'aldose réductase afin de modifier leur efficacité.

Une autre propriété du fluor liés aux interactions moléculaires consiste à modifier la distribution des électrons d'un cycle aromatique. Les groupements fluors, électronégatifs, vont attirer les électrons créant une délocalisation importante de ces derniers sur la périphérie du cycle aromatique, comme illustré sur la Figure 60 [123]. Les régions riches en électrons de part et d'autre du plan de l'aromatique, communément appelées régions π ou systèmes π, vont alors être électron-déficiente. Ces régions électropositives de part et d'autre du plan sont appelées *π-hole*. Elles sont notamment responsables des interactions *lp* – π.

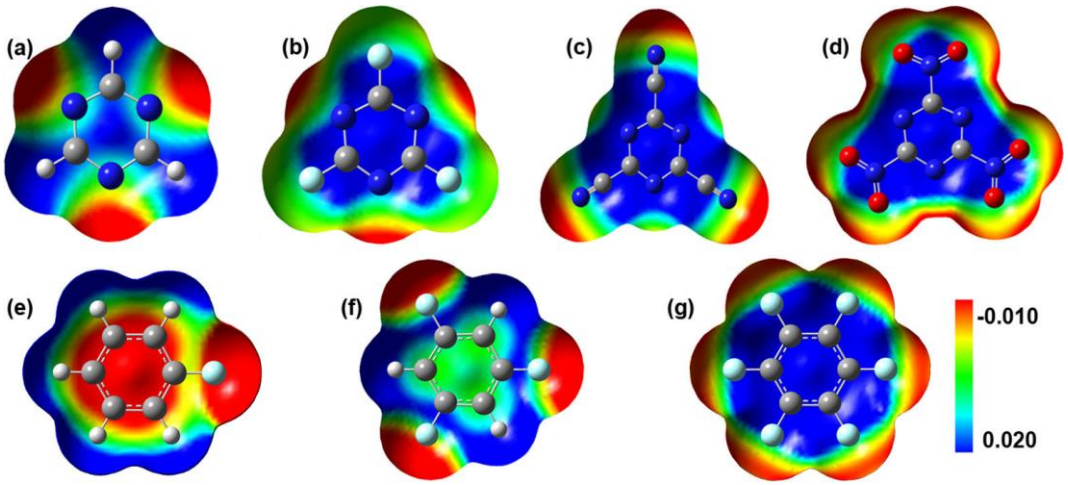


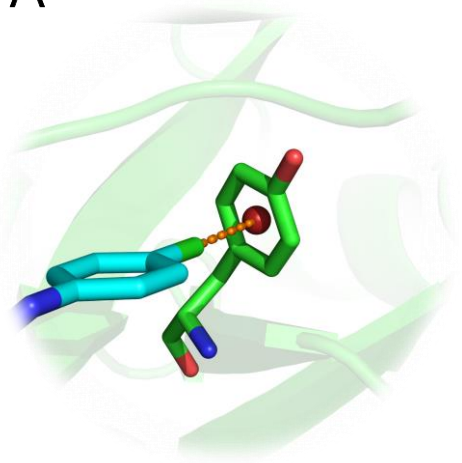
Figure 60 Surface de potentiel électrostatique de différents aromatiques A. 1,3,5-triazine B. trifluoro-1,3,5-triazine C. cyan-1,3,5-triazine D. nitro-1,3,5-triazine E. fluorobenzene F. 1,3,5-trifluorobenzene G. hexafluorobenzene. Image extraite provenant des travaux de Wang et collaborateurs [123].

C. Etudes récentes

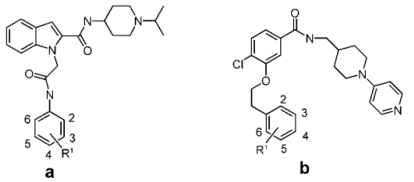
Les atomes d'halogènes sont couramment ajoutés aux ligands pour deux finalités distinctes. Le fluor et le chlore sont généralement greffés pour améliorer les propriétés d'absorption d'une molécule [124], tandis que le brome et l'iode sont ajoutés afin d'améliorer sa sélectivité [125, 126]. Néanmoins, de nombreuses études d'optimisation médicamenteuse font état d'une amélioration de l'affinité par introduction de chlore. Ainsi, les inhibiteurs de facteurs anticoagulants tel le facteur Xa ont vu leurs affinités grandement améliorées par l'ajout d'un chlore sur le cycle aromatique, comme illustré sur la Figure 61A et B [127, 128]. Expérimentalement, la liaison hydrogène entre un groupement bromophényle et le groupement donneur de l'Arginine 422 dans l'inhibiteur d'ARN polymérase chez le virus de l'hépatite C s'est traduite par une l'affinité accrue d'un facteur 250 [51].

Cette amélioration de l'affinité est en partie dû aux interactions moléculaires induit par la présence de ces halogènes, observable sur les structures cristallographiques. L'augmentation constante du nombre de ces structures disponibles dans la PDB, 300% en 10 ans, favorise les analyses d'interactions à grande échelle. Néanmoins, le nombre d'analyses se focalisant sur les interactions impliquant les halogènes est assez limité et sont généralement focalisées sur les liaisons halogènes [95, 117, 50, 129, 130].

A



B



a b

Entry	R^1	K_i [nM]	$\Delta\Delta G_{1a\rightarrow Xa}$ [kJ mol^{-1}]	K_i [nM]	$\Delta\Delta G_{1b\rightarrow Xb}$ [kJ mol^{-1}]
1	4-H	204	0	478	0
2	3-Cl	287	+0.84	1931	+3.89
3	4-Cl	3	−10.46	13	−8.51
4	4-F	63	−3.46	102	−3.40
5	4-Br	3	−10.46	36	−5.98
6	4-Me	56	−3.21	124	−2.92
7	4-CN	2855	+6.53	1552	+3.34

Figure 61 Illustration de l'effet induit par l'ajout d'un atome de chlore sur un inhibiteur de facteur Xa. A. Représentation 3D d'un inhibiteur de facteur Xa présentant une liaison halogène (code PDB 2pr3). B. Constantes d'inhibition observées (K_i) et estimations d'énergies libres ($\Delta\Delta G$) pour des bioisostères inhibiteurs de facteur Xa. En rouge, l'ajout d'un chlore en position 4 résulte dans la constante d'inhibition la plus faible, synonyme de plus grande efficacité [127].

Wilcken et collaborateurs ont mis en avant la rareté des liaisons halogènes dans la PDB ainsi que la relation entre données cristallographiques et calculs énergétiques [95]. Sirimulla a identifié les tendances préférentielles entre halogènes et acides aminés [129]. La dernière, plus prospective, s'intéresse aux liaisons hydrogènes sur les halogènes lourds. Ces interactions étaient alors soupçonnées mais n'ont été analysés statistiquement pour la première fois qu'en 2017 [50]. Des travaux de revues de Cavallo et Kolar explorent essentiellement la région du *σ-hole* [131, 38] .

D. Vers une analyse plus complète

Toutefois ces analyses, bien que particulièrement intéressantes, montrent après une lecture approfondie quelques biais qui peuvent-être problématiques, par exemple des acides aminés de type leucine considérés comme base de Lewis [129].

Les propriétés d'interaction des halogènes ne sont pas encore bien assimilées par la communauté scientifique. Le logiciel de détection d'interaction *PLIP* [64] ne considèrent les halogènes que par les liaisons halogènes, allant même à impliquer à tort le fluor. Or, les études récentes ont mis en avant de nouvelles propriétés interactives pour les halogènes telles que leur rôle comme accepteur de liaison hydrogène pour les halogènes lourds. Il est donc

nécessaire de décrire l'intégralité environnementale des halogènes afin d'avoir une meilleure description du contexte protéine - ligand. De cette description plus complète peut résulter de potentiels nouveaux motifs d'interaction. Ces analyses sont déterminantes. Il a été noté par Ballester et collaborateurs et Lenselink et collaborateurs qu'une description moins catégorique d'une interaction moléculaire contribuait à de meilleurs résultats dans la prédiction de l'affinité et la comparaison de complexes [75, 77]. Ces résultats suggèrent que des contacts voire interactions, peu décrites voire peu fréquentes, peuvent avoir un impact majeur dans l'interaction protéine - ligand et donc doivent être considérées.

Aussi, à travers l'étude exhaustive que j'ai pu effectuer, les diverses capacités d'interaction impliquant les atomes d'halogènes ont été étudiés. Ces résultats ont été confrontés aux travaux déjà publiés dans la littérature, mais aussi mis en relation avec des interactions peu décrites jusque-là. Pour cela, un jeu de données issue de la PDB a été extrait et filtré afin de limiter la redondance des observations. Un recensement des interactions connues et inexplorées a donc été réalisé afin d'examiner de nouvelles perspectives en matière de *drug design*.

Chapitre 2 : Données et méthodes

Lors de cette étude, une distinction est effectuée entre les termes *interactions* et *contacts*. Une interaction implique une composante électrostatique de la part des deux éléments impliqués dans l'interaction tandis que lors d'un contact, au moins un atome ne présente pas de polarité électrostatique, par exemple un groupement méthyle.

A. Jeu de données

Les structures extraites de la PDB présentes en mars 2018 ont été utilisées dans la construction du jeu de données. Seules les structures ayant une résolution inférieure à 2,5Å et comportant plus de 30 acides aminés sont considérés ici. Chaque structure étant traitée et enregistrée dans 3decision® (voir Partie 2.1). Un filtre au niveau des ligands a été appliqué dans un premier temps. Les molécules disposant d'un atome d'halogène, d'un poids moléculaire compris entre 250Da et 800Da, comportant au moins un cycle et ayant un pourcentage de liaisons covalentes simples inférieure à 90% ont été conservés. Ce procédé

permet de ne pas analyser des molécules peu considérées dans le domaine du *drug design* telles que les porphyrines, les staurosporines ou des lipides.

Afin de limiter la redondance structurale dans la PDB présente par exemple au sein des structures RMN mais aussi à travers plusieurs entrées PDB, une seule représentation par complexe protéine - ligand a été conservée. Les complexes identiques ont été réunis en fonction de leur identifiant UniProt ainsi que de leur représentation moléculaire *SMILES*.

L'ensemble des contacts impliquant au moins un atome d'halogène est aussi considéré. Le seuil maximal de distance toléré pour les contacts correspond à la somme des rayons de Van der Waals plus un delta de 1,0Å. Tous contacts dont l'angle halogène ou l'angle résidu est inférieur à 70° ont été ignorés. Ces valeurs correspondent à des configurations géométriques où, par exemple, l'atome du résidu est majoritairement en contact avec le carbone covalent de l'halogène dans le cas d'un angle halogène inférieur à 70°.

Les fragments générés lors du protocole de détection de contacts ont été conservés tels quels (voir Partie 2.2.C.a).

B. Description des interactions

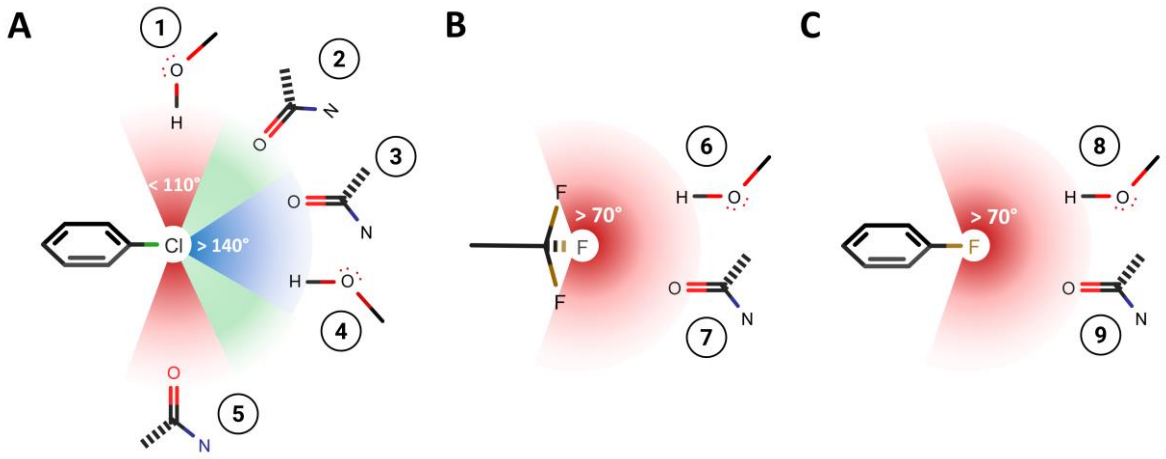


Figure 62 Résumé des interactions décrites autour des halogènes. Halogènes lourds (Cl, Br, I) : 1. Liaison hydrogène 2. Contacts polaires dans la région neutre 3. Liaison halogène 4. Interaction défavorable par répulsion de charge positive 5. Interaction défavorable par répulsion de charges négatives. Fluor aliphatique : 6. Liaison hydrogène 7. Interaction défavorable par répulsion de charge négative. Fluors aromatiques : 8. Liaison hydrogène 9. Interaction défavorable.

a. Liaisons halogènes

La liaison halogène s'effectue entre un moment électropositif de la région σ et une charge partielle électronégative comme un doublet non liant. La taille du σ-*hole* est dépendant de l'halogène impliqué. Cependant un intervalle d'angle identique pour les trois halogènes lourds, Cl, Br et I, a été considéré ici. Un contact est défini comme une liaison halogène si un angle halogène est compris entre 140° et 180° et est opposé à une surface électronégative (cas 3 de la Figure 62). La géométrie définit par l'angle résidu est dépendant de l'atome impliqué et son état d'hybridation, par exemple 120° ± 20° pour un oxygène et soufre et entre 140° et 180° pour l'azote de l'histidine et le vecteur normal d'un cycle aromatique.

b. Liaisons hydrogènes

Les liaisons hydrogènes ont été définies spécifiquement en fonction de l'halogène impliqué. Pour le fluor, l'angle halogène doit être supérieur à 70°, définissant ainsi une large région électronégative autour du noyau du fluor (cas 6 et 8 de la Figure 62). Pour ce qui est des halogènes lourds, l'angle halogène défini reflète la position orthogonale des électrons par rapport à la liaison covalente, soit une valeur oscillant autour de 90° (± 20°) (cas 1 de la Figure 62). L'angle avec le résidu, nommé *angle résidu*, est le même pour ce qui est du donneur de liaison hydrogène, l'hydrogène doit potentiellement faire face à l'halogène. Ne considérant pas les hydrogènes pour l'ensemble des structures, un groupement hydroxyle doit avoir un angle résidu compris entre 100 et 120° pour effectuer une liaison hydrogène.

Le carbone du groupement amide des acides aminés dispose d'une charge partielle positive induite par la délocalisation des électrons au niveau de l'oxygène et de l'amine. Dès lors, la surface d'un fluor ou la ceinture électronégative d'un halogène lourd peut interagir avec cette surface. Bien qu'aucun hydrogène ne soit impliqué, l'interaction est intégrée à ce type spécifique par la similarité électrostatique des acteurs impliqués. Ce carbone étant trivalent et plan, un vecteur normal au plan a été tracé afin de déterminer la position de l'halogène. L'angle résidu au niveau du carbonyle doit être compris entre 155° et 180°. Plusieurs configurations ont été décrites en fonction de l'orientation du fluor par rapport au plan de l'amine, de manière similaire à ce qui a été fait pour les aromatiques. Une conformation dite en T est décrite si l'angle halogène est supérieur à 165°, faisant face au plan de l'amide. Une conformation inclinée correspond à un angle halogène compris entre 120° et 165° et enfin, un

angle halogène en dessous de 120° correspond à une configuration dite parallèle, la liaison covalente de l'halogène et le plan de l'amide étant parallèle.

c. Interactions polaires défavorables

Une interaction polaire dite défavorable est caractérisée par plusieurs conditions. Dans un premier temps, lorsque deux éléments de même polarité se font face, le cas est énergétiquement défavorable. Ces cas spécifiques impliquent à la fois des donneurs de liaisons hydrogènes en face de la région σ (cas 4 de la Figure 62) mais aussi des accepteurs de liaisons hydrogènes au niveau de la ceinture électronégative ou orienté vers un fluor (cas 7 et 9 de la Figure 62). Dans certaines études, ces interactions seront considérées comme répulsives. Toutefois aucune valeur énergétique n'a été mesurée pour ce type d'interaction. Dans un second temps, et spécifiquement pour les halogènes lourds, la zone électrostatique entre 110° et 140° n'est jamais décrite (cas 2 de la Figure 62). Elle est potentiellement neutre par la dispersion du nuage électronique, aussi les atomes polaires faisant face à cette région ont été analysés dans notre jeu de données.

d. Contacts hydrophobes

Un contact est considéré comme hydrophobe lorsqu'un halogène fait face à la surface accessible d'un groupement carboné. Une distinction est effectuée entre les différents carbones, regroupés en 5 catégories différentes : C_α du squelette peptidique, carbone aromatique, groupement méthyle terminale -RCH_3, -R_2CH_2 et R_3CH. L'ensemble de la surface accessible des halogènes est considéré dans cette étude, seule l'orientation du groupement hydrophobe est prise en compte en fonction du nombre d'hydrogènes implicites.

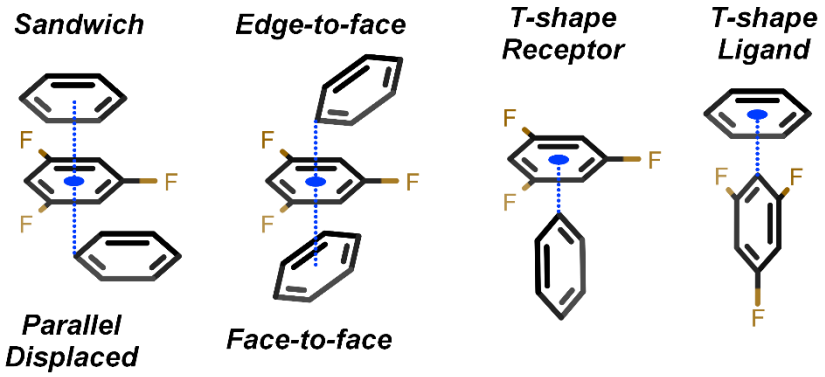
e. Description aromatique

La caractérisation des angles halogènes et angle récepteur est identique à celle décrite dans la Partie 2.2.C.b pour les angles *ligand* et *récepteur*.

L'arrangement des aromatiques comportant une région électronégative soit le tryptophane, la tyrosine et la phénylalanine, et un *π-hole* a été analysé. De même, à titre de comparaison, les interactions π - π ont été analysées. Les différentes configurations entre aromatiques ont été catégorisées selon les critères de Aravinda et collaborateurs [132] et sont illustrés sur la

Figure 63. Les configurations dites *sandwich* et *face-to-face* décrivent deux aromatiques dont les centres de masse sont alignés, en cas de décalage les conformations sont appelées *edge-to-face* ou *parallel-displaced*. *T-shape* correspond à une géométrie en forme de 'T', la nature de l'aromatique formant le tronc du T détermine la sous-catégorie.

Pour les configurations *T-shape* et *edge-to-face*, une inspection a été effectuée pour s'assurer qu'aucun atome lourd lié de façon covalente à un aromatique n'interférait entre les deux aromatiques.



Configuration	Angle orientation	Conditions
Sandwich	> 160°	Angle ligand/récepteur > 160°
Parallèle-décalé	> 160°	135° < Angle ligand/récepteur < 160°
Face à face	105° < x < 160°	Angle ligand/récepteur > 160°
Bord vers face	105° < x < 160°	135° < Angle ligand/récepteur < 160°
***T-shape* Récepteur**	75° < x < 105°	Angle ligand > 150°
***T-shape* Halogène**	75° < x < 105°	Angle récepteur > 150°

Figure 63 Paramètres utilisés pour décrire les différentes conformations aromatiques.

Chapitre 3 : Résultats

A. Analyse du jeu de données

Un total de 136 318 entrées PDB a été traité dans la base de données relationnelle 3decision® à la date du 15 Mars 2018. Parmi 24 733 ligands uniques, un quart (soit 5 950) contenait au moins un atome halogène, chaque ligand pouvant appartenir à plusieurs complexes. La moitié d'entre eux ne contenant qu'un seul atome halogène, 24% des ligands sont composés de plus de 3 halogènes, essentiellement des atomes de fluor. Un maximum de 23 halogènes sur un même ligand a été détecté sur le *tricosakis* (fluorenyl-dodecanoylamino acide propanoique) (voir Figure 64A).

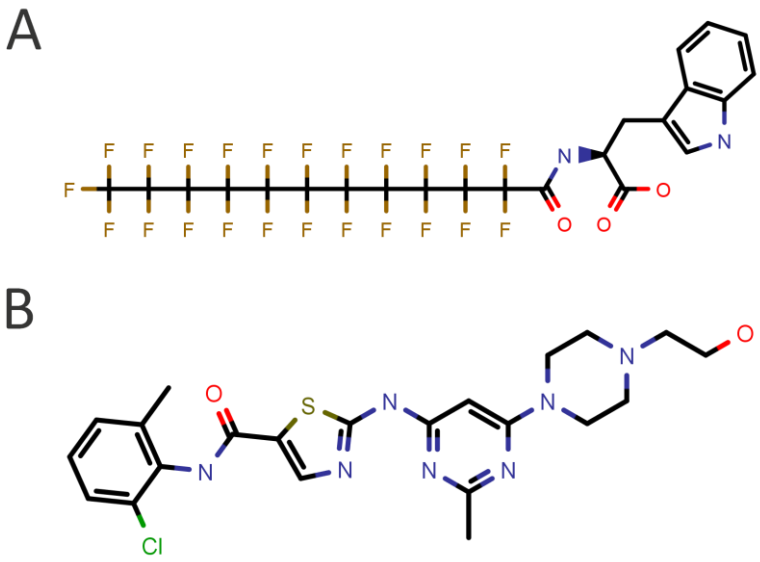


Figure 64 Représentation schématique 2D des ligands A. Tricosakis et B. Dasatinib.

Comparé au jeu de données utilisé par Sirimulla et collaborateurs en 2013, le nombre de contacts observés dans la PDB a augmenté d'un facteur 3 dans des conditions de détection similaires. Les halogènes interagissant toujours malgré une augmentation de distance de 1Å, 139 850 contacts ont été détectés en utilisant ce critère de distance. Ne prendre qu'un seul représentant par complexe pour limiter la redondance aboutit à un chiffre 3 281 complexes, soit 26,9% du jeu de données initial. Cette réduction de données ne modifie pas la distribution du nombre d'halogènes par ligand.

Certains ligands sont fréquemment présents tel que le *Dasatinib* (Figure 64B) retrouvés avec 11 complexes kinases distincts ou l'*Indomethacin* détecté avec 9 protéines différentes. 192 ligands sur 3 001 présentent plus d'une conformation dans notre jeu de données. De manière similaire, certaines protéines sont surreprésentées telles que la β-sécrétase 1 (BACE 1), l'anhydrase carbonique et des protéines de capsides identifiées avec respectivement 146, 74 et 64 ligands distincts.

La distribution des halogènes n'est pas uniforme dans le jeu de données car 60,4% des halogènes sont des fluors et 31,8% des chlores. Le brome et l'iode sont largement sous-représentés avec 383 (6,1%) et 115 (1,8%) occurrences respectivement. Cependant, ces pourcentages sont en accord avec les tendances (et contraintes actuelles) observées en *drug design*, le brome et l'iode par leur taille imposante sont plus difficiles à implémenter dans des petites molécules. De plus, les pourcentages observés dans la base de données de petites molécules ChEMBL23 indiquent des pourcentages relativement similaires avec 55,1% de fluor, 35,9% de chlore, 7,6% de brome et 1,4% d'iode.

Halogène	Nombre d' atomes	Liaison halogène	Hydrogène (amide)	Contact Hydrophobique	Interaction Défavorable
Fluor Aromatique	1606	/	278 (97)	988	468
Fluor Aliphatique	2201	/	349 (128)	1382	607
Chlore	2000	413	279 (105)	1416	445
Brome	383	75	62 (31)	239	94
Iode	115	43	20 (8)	70	19
Total	**6305**	**531**	**988**	**4095**	**1638**

Tableau 3 Récapitulatif du nombre d'halogènes présents dans le jeu de données ainsi que leur implication dans chaque interaction.

B. Diversité des fragments

L'implication des atomes d'halogènes dans les interactions moléculaires est en partie dépendante des groupements chimiques auxquels ils sont rattachés. Ainsi, plus un groupement, ou fragment, est électroattracteur, plus la région sera chargée positivement et la liaison halogène forte par exemple. La nature et la composition du fragment auquel est rattaché l'halogène vont donc influer sur sa direction et son moment polaire.

4 078 fragments halogénés impliquant au moins une interaction ont été détecté dans notre jeu de données. Regrouper ces fragments par descripteur moléculaires SMILES résulte en 273 fragments uniques (et distincts), certains étant représentés plusieurs fois. 175 de ces fragments sont aromatiques, et se retrouvent associés à 2 ligands sur 3. Le chlore, l'iode et le brome sont majoritairement intégrés sur des fragments aromatiques puisqu'ils apparaissent sur 95% des fragments présents dans notre jeu de données. Le fluor est un cas particulier puisque 42% de ces halogènes sont attachés à des groupements ou carbones aliphatiques, les plus fréquents étant le trifluorométhyle -RCF_3 et le -RCF_2H.

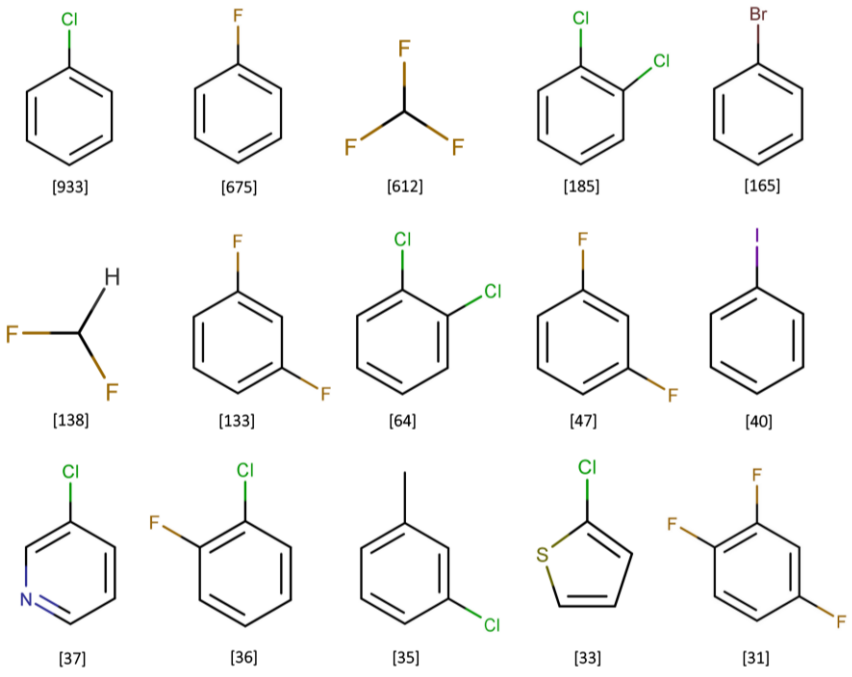


Figure 65 Représentation 2D des 15 fragments halogénés les plus fréquents dans le jeu de données ainsi que leur fréquence entre parenthèses.

Individuellement, le chlorobenzène est le fragment le plus représenté dans notre jeu de données avec 933 occurrences soit 23%, illustré sur la Figure 65. Cette surreprésentation est due notamment au *scaffold* commun utilisé dans l'élaboration des inhibiteurs de facteurs Xa (27 occurrences) et inhibiteurs de thrombines (15 complexes), jouant un rôle critique dans la liaison. Le fluorobenzène (15%) et trifluorométhyle (14%) sont aussi largement représentés. La surreprésentation de ces fragments dans les ligands est notamment induite par la réduction du pK_a modifiant ainsi son profil lipophile [133].

C. Interactions autour des halogènes lourds

Pendant longtemps, seules les liaisons halogènes étaient considérés. Depuis 2017, le rôle d'accepteurs de liaisons hydrogènes est aussi étudié à plus grande échelle. Cependant, toutes ces études ne considèrent qu'un partenaire interactif pour l'atome d'halogène alors qu'un nombre moyen de 6,5 atomes dans l'environnement direct de l'halogène soit détecté dans le jeu de données. Seul 4,7% des halogènes sont dans une situation où un seul atome du récepteur est détecté comme en interaction. Dès lors, un recensement des interactions déjà connues telles que la liaison halogène ou la liaison hydrogène a été effectué, mais aussi des interactions et contacts peu, voire pas caractérisés.

a. Région électropositive σ

19 143 contacts impliquant des halogènes lourds ont été détectés, dont 3 091 impliquant une base de Lewis (oxygène, sulfure ou azote de l'histidine). Après filtrage des contacts à l'aide des contraintes géométriques, 531 liaisons halogènes ont été détectées, soit 1 complexe sur 4 contenants uns atome lourd. 37% des atomes d'iode, 19,6% des bromes et 20,7% des chlores sont impliqués dans ce type d'interaction (voir Tableau 3).

L'oxygène du squelette peptidique est identifié comme étant l'accepteur de liaison halogène le plus fréquent avec un total de 152 occurrences non redondantes, quelques soit l'halogène impliqué illustré sur Figure 66. Ces résultats sont similaires à ceux obtenus par Wilcken et collaborateurs [95]. Les groupements aromatiques sont régulièrement en interaction avec la région sigma des halogènes. 19,5% des interactions observées sont réalisés entre le groupement aromatique d'une tyrosine et un atome de chlore. De plus, 19 complexes sont impliqués dans une interaction chlore - phénylalanine et 11 de type brome - phénylalanine.

Les groupements hydroxyles de la tyrosine et sérine ainsi que le soufre de la méthionine sont aussi des cas où plus de 10 occurrences sont recensées.

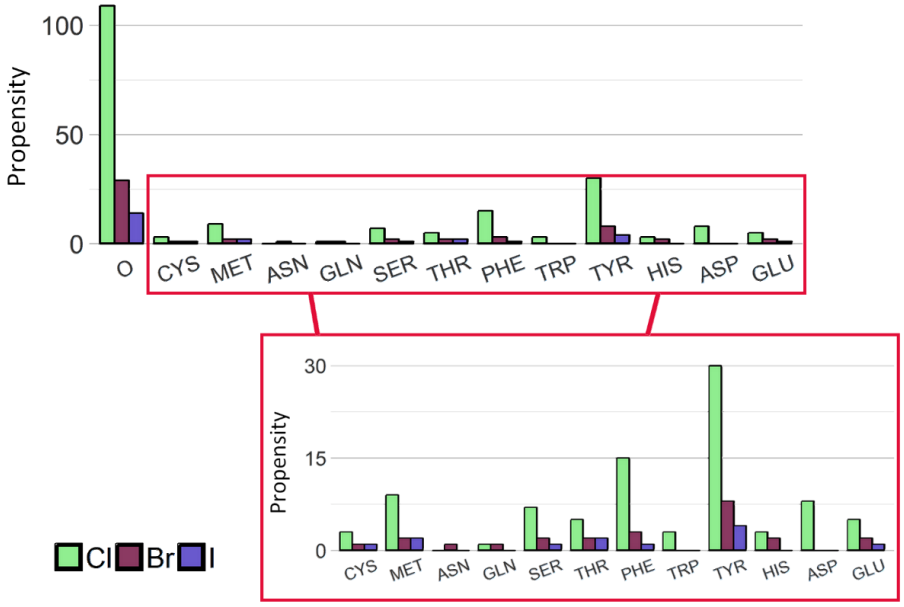


Figure 66 Distribution des liaisons halogènes par type d'halogènes et d'accepteurs de liaisons halogènes.

La préférence des halogènes pour des acides aminés spécifiques est un paramètre étudié par Sirimulla et collaborateurs [129]. Néanmoins, ces résultats doivent être considérés avec précaution étant donné qu'un halogène appartenant à un *scaffold* commun d'un inhibiteur protéique sera potentiellement recensé plusieurs fois. Les inhibiteurs de Factor Xa en sont un parfait exemple, 52 d'entre eux présente une liaison halogène avec la même tyrosine 228, pouvant laisser suggérer une préférence du chlore envers la tyrosine. De même, un nombre important d'interactions, 27 exemples de chlore – histidine ont été détectés dans l'ubiquitine protéine ligase Mdm2. En ne considérant qu'un représentant par fragment et protéine communs, une diminution de 106 à 17 interactions chlore – tyrosine est vue tandis que la liaison halogène chlore - phénylalanine ne diminue que de 4 occurrences.

De par sa taille réduite, il est rare de retrouver deux donneurs de liaisons halogènes en face de la région σ. Malgré des paramètres géométriques restreints, certains cas sont identifiables. Parmi les 11 complexes ayant cette configuration spécifique, deux bioisostères d'inhibiteurs de la protéine Aequorine ont été cristallisés avec pour substitution un brome et un iode. Les paramètres géométriques identifient deux donneurs de liaisons halogènes. Néanmoins, comme le montre la Figure 67, seule la chaine latérale de la tyrosine 166 réalise une liaison

halogène, l'oxygène de la valine 162 étant déjà impliqué dans une liaison hydrogène maintenant l'intégrité structurale de l'hélice α.

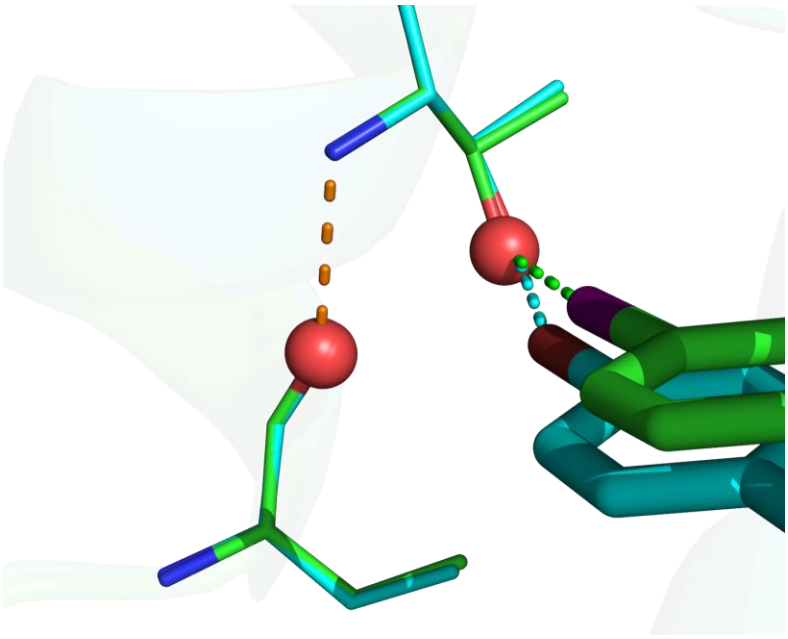


Figure 67 Représentation 3D d'une liaison halogène sur des bioisostères inhibiteurs d'aequorine (1uhi en vert et 1uhj en cyan). Une interaction hydrogène propre à l'hélice α est représentée en orange.

La géométrie linéaire des liaisons halogènes n'est que très rarement observé comme le montre la densité de distribution géométrique (voir Figure 68). Deux pics de distribution sont observés pour l'angle halogène du chlore à 140° et 165° tandis que les liaisons halogènes impliquant le brome sont majoritairement situés aux alentours de 160°. A l'inverse, la distribution de distance est plus en accord avec les observations quantiques mesurés, le pic de distribution se situant à 3,7Å pour le chlore (contre environ 3,45Å en théorie).

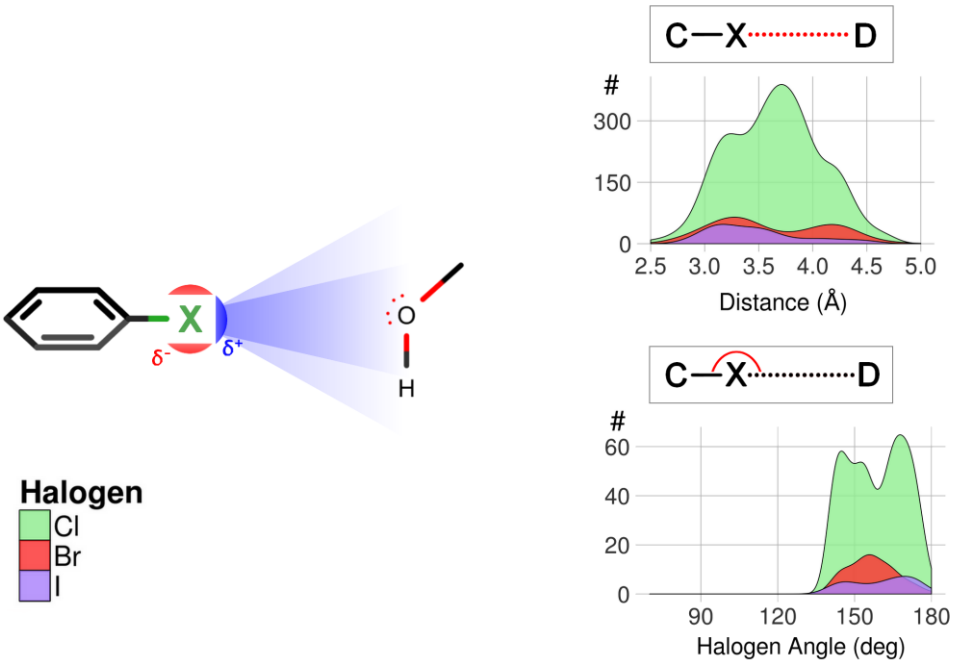


Figure 68 Distribution des densités des distances et angles observés pour les liaisons halogènes.

Un des éléments confirmant l'existence des régions σ et leur charge partielle positive relève du faible nombre de donneurs de liaisons hydrogènes se situant en face de cette région. Ce type d'interactions, répulsif ou défavorable d'un point de vue électrostatique, n'est relevé que dans 80 complexes dans le jeu de données (voir Figure 69). 56 cas correspondent à un groupement hydroxyle où la présence d'un hydrogène ou du doublet non liant est difficilement discernable en l'absence d'hydrogénation. Les 34 cas restant étant essentiellement constitué d'arginine et d'asparagine.

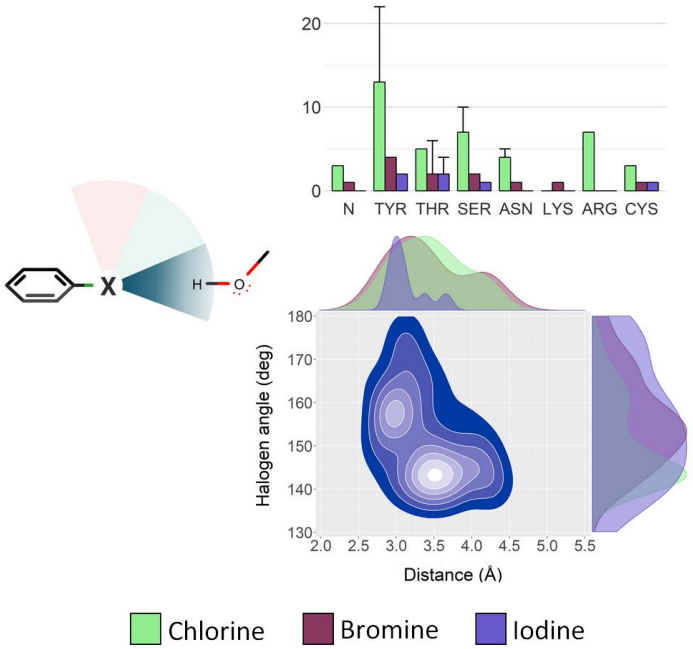


Figure 69 Propension des donneurs de liaisons hydrogènes en face du σ-hole (résultats redondant affichés en barres) ainsi que la distribution de densité de leur descripteur géométrique.

b. Anisotropie périphérique des électrons

La distribution anisotrope des électrons sur les halogènes lourds contribue à la formation d'une ceinture électronégative, localisée dans le prolongement de sa liaison covalente. Bien que le rôle d'accepteur de liaison hydrogène ait été évoqué en 1979 [113], la première étude exhaustive sur ce mécanisme d'interaction n'a été réalisé qu'en 2017 par Lin et collaborateurs [50] et suggère un rôle attractif plus important que la liaison halogène.

En plus des donneurs de liaisons hydrogènes, tout groupement fonctionnel présentant un moment polaire chargé positivement peut théoriquement interagir avec cette région. Sur les 2 498 halogènes lourds présents dans le jeu de données, 8,7% des atomes de chlores, 8,1% des bromes et 10,4% des atomes d'iodes sont dans ce type de configuration interactive (voir Tableau 3). Parmi ces éléments électropositifs, l'azote du squelette peptidique constitue le groupement le plus souvent en contact avec cette région. Le groupement hydroxyle de la tyrosine est un élément observé plusieurs fois en face de cette région riche en électrons, avec respectivement 17 et 23 occurrences pour le brome et le chlore. Le groupement guanidine de l'arginine fait partie des éléments ayant plus de 10 occurrences en interaction avec la région électronégative.

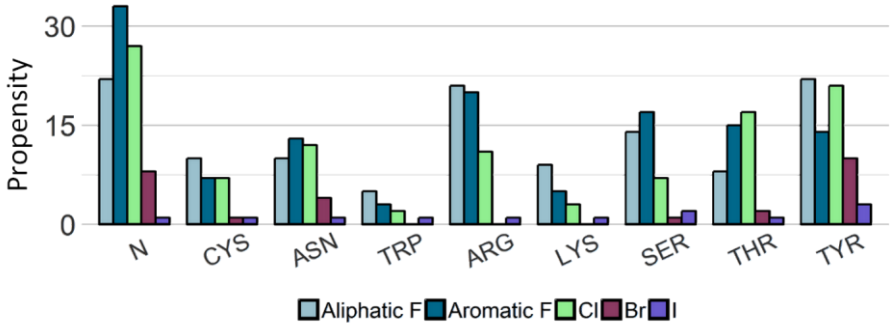


Figure 70 Distribution des liaisons hydrogènes par type d'halogènes et d'accepteurs de liaisons hydrogènes.

Le pourcentage modéré d'halogènes impliqués dans ce type d'interactions peut s'expliquer par la surface étroite accessible formée par le nuage électronique. Certains groupements tels que le carboxyle de la thréonine et de l'asparagine, présents plus de 10 fois au niveau de cette région, peuvent être ambigus. En effet, la détermination des atomes lors de la résolution par cristallographie peut amener à une inversion de l'assignation de la fonction carbonyle et hydroxyle de ces résidus. La résultante de cette inversion provoque un changement radical dans la nature de l'interaction détectée, favorable dans le cas de l'hydroxyle et défavorable pour le carbonyle.

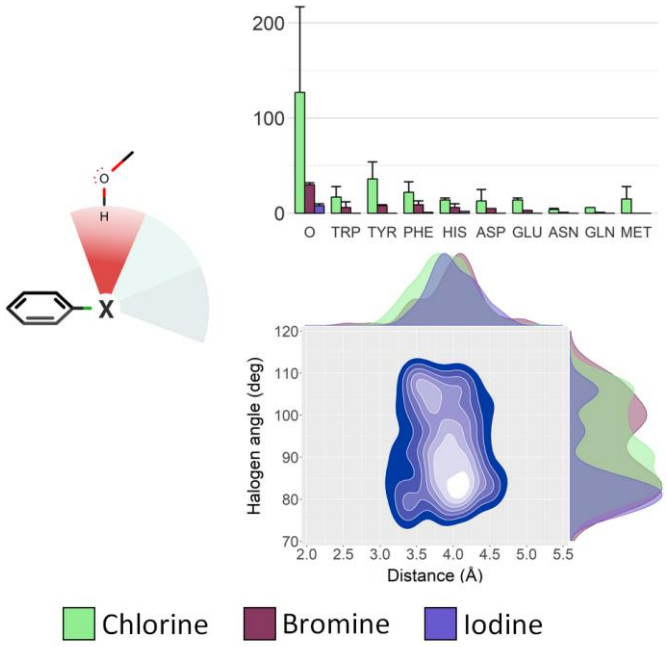


Figure 71 Propension des donneurs de liaisons hydrogènes en face du nuage électronique (résultats redondant affichés en barres) ainsi que la distribution de densité des descripteurs géométriques (distance et angle halogène).

Néanmoins, un nombre important d'éléments électrostatiques défavorables se situe au niveau de cette région. En effet, 527 interactions impliquant la circonférence électronégative des halogènes lourds et un moment de même polarité ont été observé dans le jeu de données. 470 halogènes lourds sont impliqués dont 18,9% de chlore, 21,1% de brome et 10,4% d'iode. Le carbonyle du squelette peptidique représente 37,3% des interactions détectés, suivi du système π des chaines latérales aromatiques avec 27,2% (voir Figure 71).

Les inhibiteurs de facteur anticoagulant Xa représentent une part significative de ces interactions avec 32 cas d'interactions orthogonales chlore – oxygène peptidique détectés. L'étude en détail de ces cas répétés montrent une certaine ambiguïté par l'implication de l'oxygène dans deux liaisons hydrogènes dans un feuillet β, représenté dans la Figure 72.

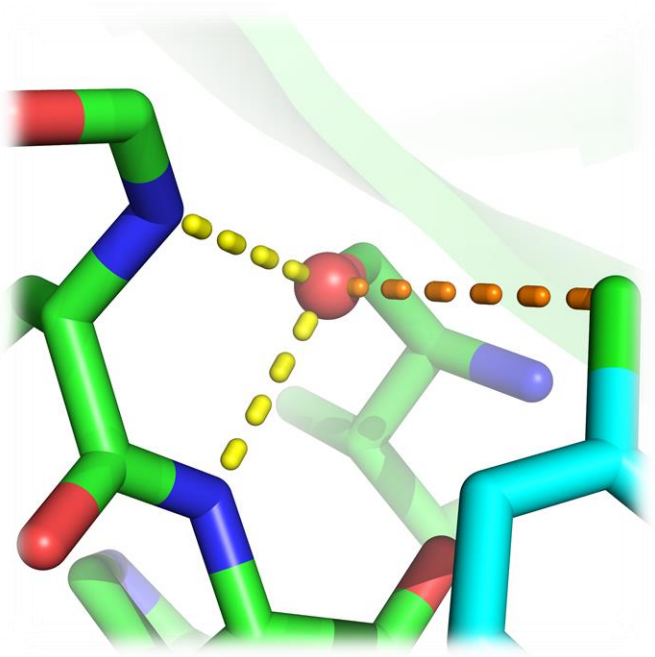


Figure 72 Illustration 3D d'une potentielle interaction défavorable (orange) et de deux liaisons hydrogènes (jaune) sur l'oxygène du squelette peptidique de l'Isoleucine 227 (code PDB 2pr3).

De même, la prothrombine, la β-sécrétase 1 et la méthionyle ARNt synthétase font partis des protéines dont les ligands présentent souvent une interaction défavorable impliquant l'oxygène peptidique. Les inhibiteurs chloriques interagissent de manière récurrente avec le cycle aromatique de la phénylalanine de la protéine HSP90α. Enfin, la ceinture électronégative du brome est engagée dans de nombreuses interactions, notamment le tryptophane 528 de la polyprotéine génomique et la phénylalanine 59 de la déhydratase FabZ.

c. Région apolaire

La région décrite comme neutre en terme électrostatique se situe dans des valeurs d'angle halogène autour de 125°. Cette région, à l'interface de la région et de ceinture polaire négative est décrite dans certaines études comme pouvant être à l'origine de répulsion. Néanmoins, les représentations 3D de la surface potentielle électrostatique laisse suggérer une zone relativement neutre.

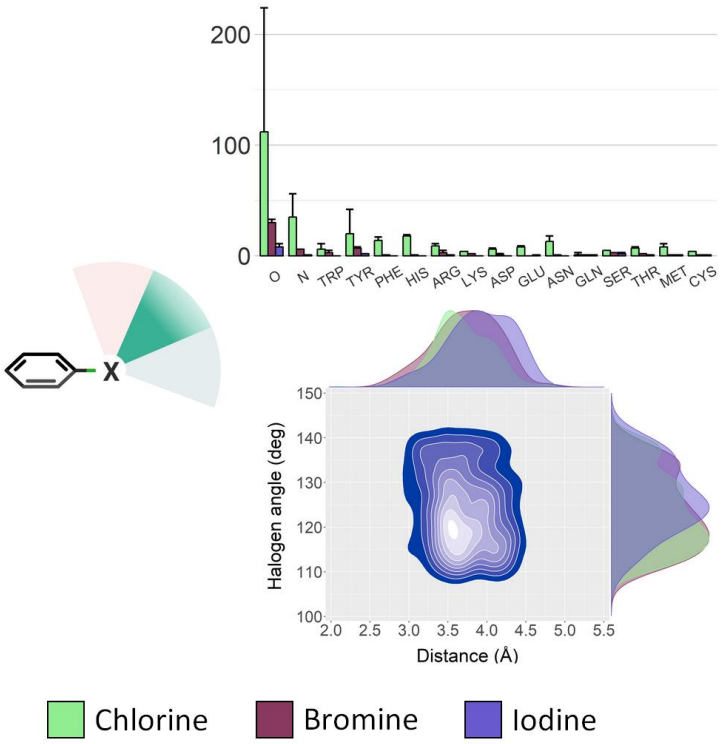


Figure 73 Propension des éléments polaires dans un angle halogène oscillant autour de 125° (résultats redondant affichés en barres) ainsi que la distribution de densité des valeurs géométriques d'angles et de distances.

543 éléments polaires ont été identifiés comme faisant face à ce type de région, soit un nombre plus important que les liaisons halogènes. Ce type d'interaction affecte de la même manière les trois types d'halogènes lourds avec des fréquences allant 18,3% des bromes impliqués jusqu'à 19,4% des chlores. De manière identique aux observations sur les éléments répulsifs, le groupement carbonyle commun à tous les résidus est fréquemment détecté puisque 224 interactions impliquant un chlore, 33 couplés au brome et 11 à l'iode ont été relevé (voir Figure 73). Les groupements amides du squelette et les cycles aromatiques sont

modestement situés dans cette région, spécialement pour le chlore avec des pourcentages respectifs de 14,5% et 16,8%.

Un nombre répété de cas, 57 occurrences impliquant le chlore proviennent des inhibiteurs ciblant la famille protéique des facteurs anticoagulants (facteurs Xa, XI, prothrombine…). La tyrosine 100 de l'ubiquitine ligase MDM2 est aussi une interaction souvent observée avec le chlore. La présence d'un nombre important d'éléments polaires dans cette région, aussi récurrent que les liaisons halogènes, peut suggérer la présence d'une interaction dans cette région d'un point de vue statistique.

D. Fluor

Considérer le fluor comme un accepteur faible de liaisons hydrogènes ne s'est fait que relativement récemment et reste controversé aujourd'hui. Toutefois, l'introduction d'atome de fluor dans certains bioisostères a contribué à l'augmentation de l'affinité dans l'optimisation de molécules inhibitrices de l'aldose réductase [122]. J'ai distingué ici deux catégories pour les fluors selon la nature du fragment auxquels ils sont rattachés : (i) fluor aliphatique pour des groupements de type trifluorométhyle -CF_3, et (ii) aromatique pour des fluors attachés directement à un cycle aromatique. Un total de 21 423 contacts impliquant un atome de fluor a ainsi été détecté dans le jeu de données.

12,1% des fluors aromatiques ont été identifiés comme étant impliqués dans une interaction de type liaison hydrogène, et ce malgré une surface d'interaction relativement large (voir Tableau 3). L'azote du squelette peptidique étant le donneur de liaison hydrogène le plus répandu dans une protéine, il est donc logique de le retrouver comme étant le partenaire interactif le plus souvent détecté pour les atomes fluoriques (28,7% des fluors en interaction).

Les groupements hydroxyle sont des partenaires récurrents des fluors aromatiques. 223 interactions ont été détecté dans ce type de configuration, soit 19,7%, 12,6% et 8,1% des fluors en interaction avec la thréonine, sérine et tyrosine respectivement.

Considérer une seule représentation de complexe ne modifie pas la préférence du fluor par rapport à l'azote du squelette peptidique (26,4% des fluors). Un nombre conséquent d'interaction fluor – tyrosine sont cependant issues de complexes redondants, une diminution de 44 à 15 occurrences est observée. Le groupement guanidine de l'arginine est peu détecté comme donneur de liaison hydrogène malgré le nombre important de donneur de liaison

hydrogène présent. De même, la sérine n'est en interaction que 13 fois dans notre jeu de données.

Zhou et collaborateurs ont suggéré en 2009 que l'azote de l'amide était l'élément le moins fréquemment en interaction avec le fluor. Les résultats obtenus ici sont en contradiction avec ce résultat, et montrent l'importance de mettre à jour de manière régulière nos connaissances, surtout dans le cadre d'événements peu fréquents [134].

De manière surprenante et dans un nombre de cas limité, un atome d'halogène impliquant de multiples interactions hydrogènes a été observé, et ceci malgré la présence d'une région accepteur de liaisons hydrogènes large. Parmi ces cas, les inhibiteurs d'anhydrase carbonique 2 se distinguent par la présence d'un fluor aromatique en interaction avec l'hydrogène du squelette peptidique thréonine 198 et potentiellement avec le même groupement de la thréonine 199, illustré sur la Figure 74. Ces interactions doubles peuvent renforcer l'attractivité de ces fluors spécifiques.

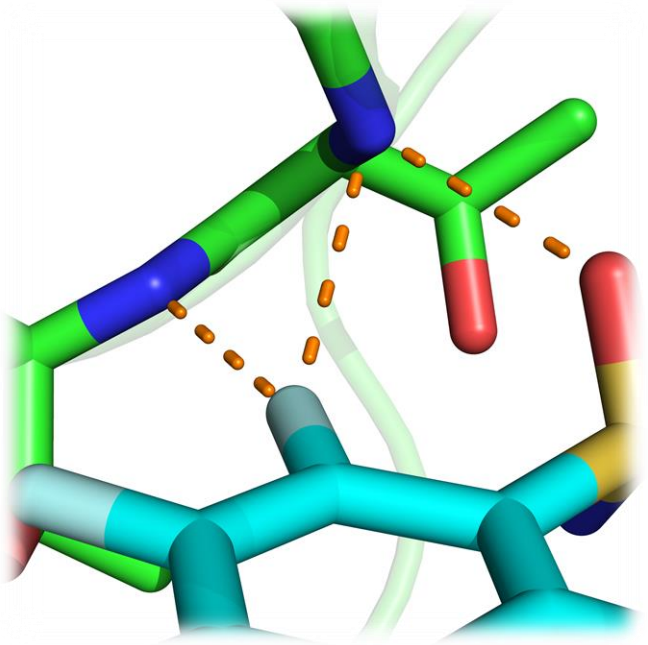


Figure 74 Représentation 3D d'un fluor potentiellement en interaction avec deux donneurs de liaisons hydrogènes (code PDB 4pzh).

Les propensions relevées dans la Tableau 3 indiquent un pourcentage de fluor aliphatique en interaction avec un donneur de liaison hydrogène similaire aux fluors aromatiques, soit 10,3%. L'azote du groupement amide du squelette peptidique est de nouveau le partenaire interactif le plus fréquemment observé à proximité d'un fluor aliphatique (21,8% des interactions

observés). Le groupement hydroxyle de la tyrosine ainsi que l'hydrogène porté par l'atome de sulfure de la cystéine sont aussi très souvent impliqués, respectivement 17,9% et 20,5%. 46 interactions engageant le groupement guanidine de l'arginine et un fluor aliphatique ont été détectés, les azotes de ces groupements n'ayant pas été distingués.

La présence répétée de certaines protéines indique que certains inhibiteurs sont composés du même *scaffold* de fluor aliphatique. Les inhibiteurs de facteur Xa (10 occurrences), chaine lourde de la myosine (8 occurrences) et chymotrypsine élastase (7 occurrences) disposent notamment d'un groupement aliphatique fluor conservé structuralement. Il résulte ainsi dans la sur-observation des interactions impliquant la tyrosine, la sérine, l'arginine ainsi que l'azote de la chaine peptidique. Ces résultats sont donc à modérer lors de l'interprétation.

Autre point d'intérêt, le carbone central du groupement amide du squelette peptidique dispose d'une charge partielle électropositive induite par le moment dipolaire de l'oxygène et de l'azote. Rarement considéré comme élément interactif, les interactions impliquant cette charge positive et les atomes de fluor ont été calculés. 116 cas de fluor aliphatique et 130 cas de fluor aromatique ont été considérées comme interagissant avec la surface électropositive du carbone. L'orientation du fluor vers le carbone se fait de manière inclinée dans la majorité des fluors aliphatiques, 91 occurrences. Au contraire, les fluors aromatiques sont orientés de manière à la fois inclinée (51,2%), mais aussi parallèle (45,7%), le reste correspondant au fluor dirigé linéairement vers ce carbone (*T-shape*).

De manière intéressante, certains atomes de fluors semblent être en interaction avec de multiples amides, en configuration parallèle et inclinée. 17 fluors aromatiques se retrouvent dans cette configuration, dont 9 cas concernant des ligands de la MAP kinase p38. Ces occurrences impliquent une interaction avec le carbonyle de la leucine 104 et la valine 105. De même, 3 complexes du Facteur VIIa impliquant des glycines (positions 216 et 219) sont en interaction avec un atome de fluor.

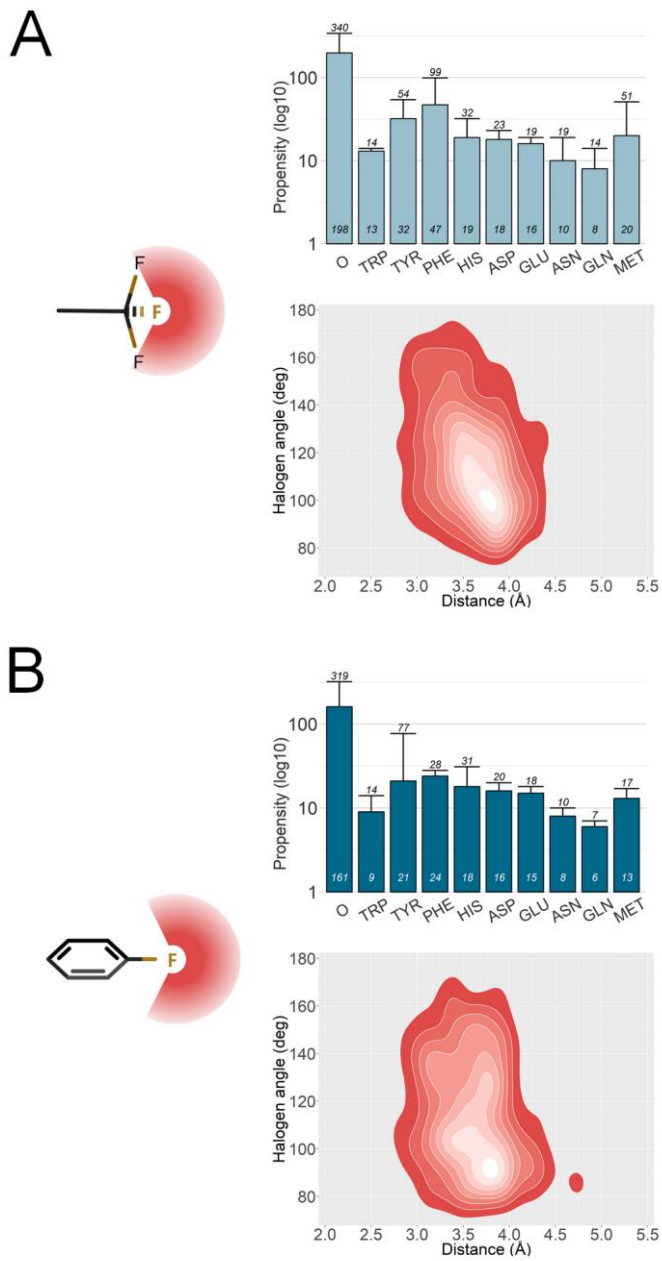


Figure 75 Propension des accepteurs de liaisons hydrogènes autour (A) des fluors aliphatiques et (B) fluors aromatiques (valeurs redondant affichés en barres) ainsi que la distribution de densité de leur paramètre géométrique.

Un nombre significatif d'interactions dites défavorables sont aussi observées pour les atomes de fluor. En effet, la surface polaire chargée négativement faisant face à un fluor est potentiellement responsable d'une répulsion électrostatique provoquée par les deux charges de même polarité. 1 075 interactions impliquant 29,1% des fluors aromatiques et 27,6% des fluors aliphatiques sont concernées par cet arrangement atomique peu décrit. L'interaction entre l'oxygène du squelette peptidique et les fluors aliphatiques et aromatiques représentent respectivement, 59,0% et 51,1% de ces arrangements défavorables. Les cycles

aromatiques sont aussi largement impliqués puisqu'ils comptent pour 25% des interactions répulsives des fluors aromatiques et 22% des fluors aliphatiques

Une redondance non négligeable dans le jeu de données est aussi observée pour ce type d'interaction avec notamment 30 occurrences entre la phénylalanine 182 et l'inhibiteur de la tyrosine phosphatase de type 1. De même, l'interaction défavorable entre la tyrosine 71 de la β-secrétase 1 et le fluor aliphatique est observée dans 28 cas distincts.

L'importante surface électronégative considérée pour le fluor permet d'identifier les cas fréquents où de multiples interactions favorables sont identifiées. La redondance de ces environnements défavorables valide leur présence et ne résulte donc pas de biais lors de la résolution structurale. Ainsi, 10 inhibiteurs de β-secrétase 1 présentent deux interactions défavorables avec trois résidus (la tyrosine en position 71, glutamine 73 et lysine 107). L'anhydrase carbonique est un cas similaire où 8 complexes protéines-ligands interagissent de manière défavorable avec la thréonine 200 et la proline 201. Les cas sont plus rares pour les fluors aliphatiques, par exemple 4 cas de ligands en interaction répulsive avec l'oxygène peptidique de la leucine 505 et le cycle aromatique de la phénylalanine 506 du récepteur nucléaire ROR-γ. La plupart de ces interactions se font dans une configuration géométrique orthogonale, soit avec un angle halogène proche de 90° (voir Figure 75).

E. Caractères hydrophobes

Deux tiers des 40 836 contacts halogénés basés sur un critère de distance ont été détectés avec un atome de carbone côté récepteur dans le jeu de données étudié. La prise en compte de la surface accessible des deux éléments en interaction ainsi que leur orientation réduit le nombre total de contacts hydrophobes à 7 016. Ces contacts sont plus récurrents que les interactions électrostatiques puisqu'elles affectent 4 095 des 6 305 halogènes de notre jeu de données.

Le Tableau 4 résume la proportion d'atomes effectuant au moins un contact hydrophobe. Ces proportions sont similaires pour le fluor, le brome et l'iode avec 62,2%, 62,4% et 60,8% respectivement. Le fluor est cependant plus propice à se situer dans un environnement hydrophobe, 70,8% des atomes détecté dans cette étude. Le fragment auxquels sont rattachés les fluors ne semblent pas avoir d'impact direct sur leurs propensions à effectuer des contacts hydrophobes (62,7% et 61,5%).

Halogène	Partenaire Hydrophobique	Valeur brute	Halogène impliqué (%)
Fluor aliphatique	Aromatique	163	7.3
	C_α	201	9.0
	CH	33	1.5
	CH_2	342	14.4
	CH_3	1528	47.6
Fluor aromatique	Aromatique	170	10.2
	C_α	171	10.6
	CH	48	2.9
	CH_2	190	11.1
	CH_3	1124	47.4
Chlore	Aromatique	181	8.4
	C_α	282	11.8
	CH	41	2.1
	CH_2	410	18.4
	CH_3	1618	53.4
Brome	Aromatique	39	9.4
	C_α	36	9.1
	CH	11	2.9
	CH_2	80	18.3
	CH_3	251	44.6
Iode	Aromatique	11	9.6
	C_α	19	15.7
	CH	6	5.2
	CH_2	34	25.2
	CH_3	67	41.7

Tableau 4 Propension et pourcentage d'halogènes impliqués dans des contacts hydrophobes par halogène

Sur les contacts 7 016 précédemment cités, le groupement méthyle -RCH_3 est le groupement hydrophobe le plus récurrent avec une fréquence d'halogènes en contact supérieure à 41% pour l'ensemble des halogènes. La grande surface de contact considérée pour le méthyle, par

son hybridation, peut être une des causes responsables de ces fortes proportions. Le méthylène -R_2CH_2 est modérément représenté, des pourcentages d'halogènes impliqués oscillant autour de 16% pour le brome, chlore et fluor. Cependant, un pourcentage plus important d'iode impliqué de 24,8% est à noter.

La prise en compte des carbones aromatiques dans les contacts hydrophobes s'effectue pour des contacts ayant lieu avec la surface aromatique pauvre en électrons. Les interactions impliquant les carbones α et carbones dits aromatiques se font de manière équivalente/équilibrée au vue de nos données (voir Tableau 4). Les contacts hydrophobes effectués par les atomes d'halogènes se font ainsi préférentiellement avec des groupements terminaux comme le méthyle.

Le nombre de groupements hydrophobes dans l'environnement interactif chaque type d'halogène présente des tendances similaires pour chaque type d'halogène (voir Figure 76). Plus de 35% des halogènes interagissent avec un seul groupement hydrophobe, à l'exception de l'iode qui interagit moins (27%). Les environnements riches en groupement hydrophobe, 4 à 5 contacts, concernent 5,2% des iodes jusqu'à un maximum de 8,5% de chlore. Certains environnements riches sont redondants et donc spécifiques de certaines protéines tel que les 17 complexes inhibiteurs chloriques ciblant l'ubiquitine ligase MDM2. 15 ligands différents de l'anhydrase carbonique 2 confirment l'environnement riche en groupements hydrophobes du site de liaison associé.

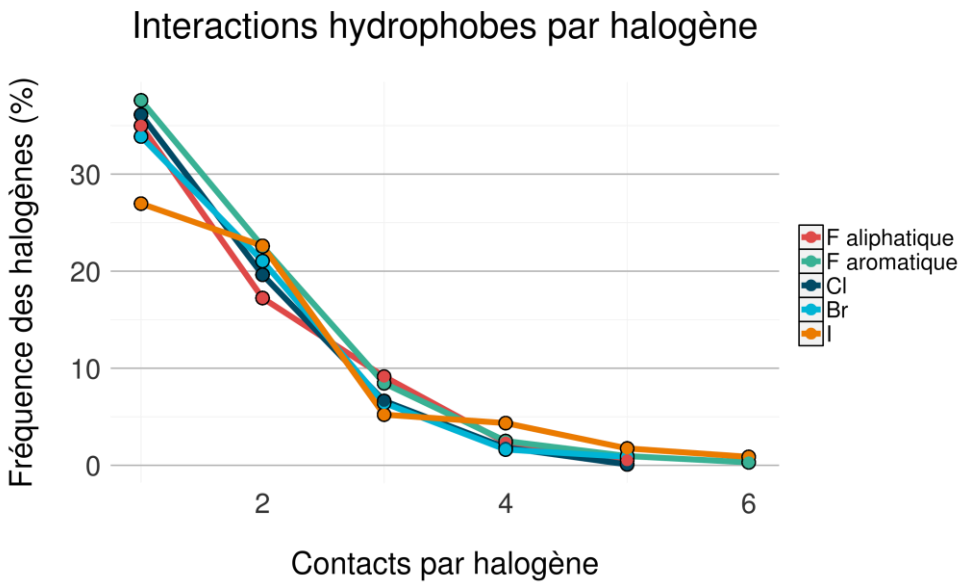


Figure 76 Distribution du nombre de partenaires hydrophobes par halogène.

La position des groupements hydrophobes autour des halogènes diffère selon leur nature. Ces groupements se situent majoritairement dans la ceinture électronégative neutre des atomes de chlore et brome (angle halogène proche de 90°). Concernant l'iode et le fluor, la distribution des composés carbonés est concentrée dans des valeurs d'angle halogène proche de 120°. La distribution maximale observée sur la distance séparant les deux éléments montre que la plupart des contacts se font à une distance supérieure au simple rayon de Van der Waals, par exemple pour le fluor et l'iode elle est supérieure de 0,5Å.

F. Environnement d'interactions des halogènes

Sur l'ensemble des 6 305 halogènes en contact, seuls 756 présentent des interactions décrites dans la littérature, ces halogènes sont essentiellement décrits pour leur aspect attractif. Or l'environnement d'un halogène pour décrire l'interaction protéine - ligand est plus complexe [135].

Un exemple frappant est le cas des 1 066 atomes d'halogènes identifiés comme ayant au moins un partenaire polaire distinct dans leur environnement proche. De plus, un nombre conséquent d'halogènes, 4 204 atomes, sont considérés comme effectuant un contact hydrophobe conduisant à un total de 5 270 halogènes sur 6 305 en interaction électrostatique et/ou contact hydrophobe. Les halogènes lourds peuvent potentiellement être impliqués dans des interactions de natures différentes dues à l'anisotropie de leur densité électronique. Afin de décrire plus précisément ces schémas d'interaction, la considération de l'intégralité de l'environnement interactif est donc plus pertinente.

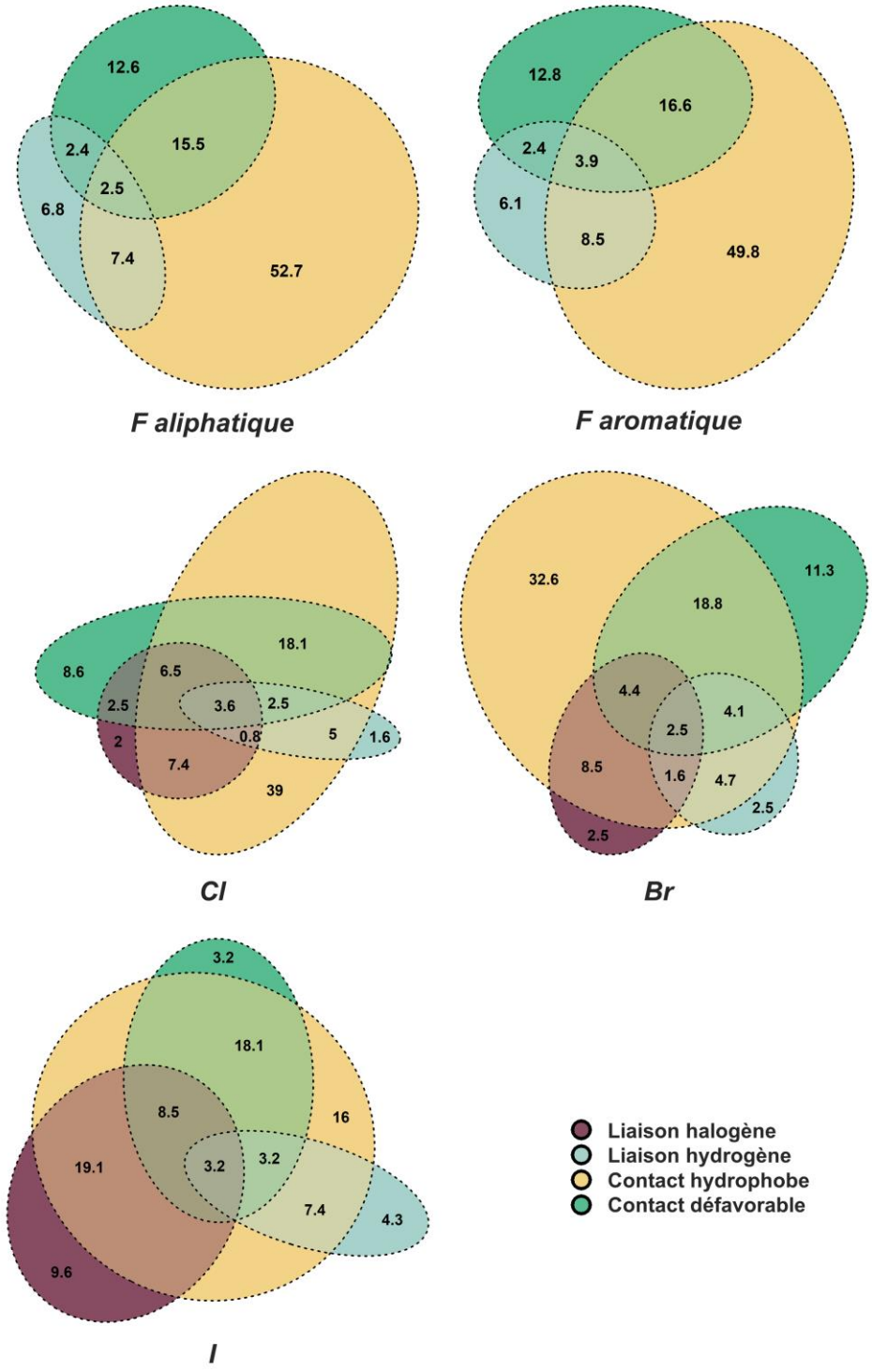


Figure 77 Diagrammes de Venn représentant l'environnement d'interaction de chaque halogène, les 4 types d'interactions sont décrites.

Le diagramme de Venn illustré sur la Figure 77 montre la co-occurrence entre les différents types d'interaction pour un atome spécifique dans un complexe.

Les résultats obtenus pour les halogènes lourds indiquent que les motifs d'interactions sont plus complexes. 48,8% du chlore, 51,1% du brome et 66,9% d'iode sont impliqués dans des interactions de natures différentes simultanément. Le chlore et le brome ont une distribution semblable de réseau d'interaction dont 35% sont des constitués exclusivement de contacts hydrophobes. La liaison halogène seule est un phénomène rare pour ces deux atomes puisque seuls 2,5% de ces atomes sont dans ce type de configuration contrairement aux idées préconçues.

Au contraire, l'iode affiche un profil d'interaction différent avec une proportion plus élevée de liaison halogène exclusive (9,6%). La co-occurrence de liaisons halogènes avec des contacts hydrophobes est un cas relativement fréquent, 1 fois sur 5. Les contacts uniquement hydrophobes sont notamment moins représentés sur l'iode, seulement 16% en comparaison des 32% du chlore.

35 complexes chloriques ont été identifiés comme étant en interaction avec l'ensemble des interactions décrites précédemment. 27 de ces cas se produisent sur la même famille protéique, observées à travers toute cette étude à savoir les facteurs anticoagulants (facteur XI, facteur X et prothrombine).

Le fragment auxquels sont rattachés les fluors n'impactent à première vue par les proportions de motifs d'interactions observés. La glycoprotéine gp160 et la β-sécrétase constituent des cas où le fluor aliphatique présent sur le ligand est dans un environnement interactif hétérogènes (hydrophobes, défavorables et donneurs de liaisons hydrogènes) avec 9 et 8 complexes respectivement. Des cas similaires pour les fluors aromatiques sont notamment identifiés sur les complexes de facteur Xa avec 11 occurrences.

Enfin, l'environnement d'interaction de la molécule *Dasatinib* contenant un chlore a été analysé plus en détail. Différents motifs d'interaction autour de ce chlore ont été observés en fonction de la protéine ciblée. Le récepteur éphrine de type A, impliqué dans les tyrosines kinase, par exemple présente une liaison halogène avec la sérine 756 (codes PDB 5i9y et 2y6o). Néanmoins, cette liaison halogène n'est pas détectée chez les autres tyrosine kinases, mettant en avant des schémas d'interaction très diversifiés pour un même atome. 3 complexes ciblant des protéines différentes affichent un motif interactif identique avec une liaison hydrogène et un contact hydrophobe : Serine/Thréonine protéine kinase 24, Tyrosine protéine kinase BMX et Tyrosine protéine kinase ABL2 (code PDB 4qms, 3sxr et 4xli respectivement).

G. Apport des halogènes sur l'interaction aromatique

Les cycles aromatiques présentent des propriétés d'interaction spécifiques. La distribution du nuage électronique créée de part et d'autre du plan des cycles une région électronégative appelée système π. La périphérie du cycle aromatique, appauvrie en électrons, représente donc un moment polaire chargé positivement. L'ajout de groupement électroattracteurs sur l'aromatique change sa polarité créant une région électropositive de part et d'autre du plan de l'aromatique appelé *σ-hole*.

Hunter et collaborateurs [136] ont analysé les configurations préférentielles adoptées par les cycles aromatiques de phénylalanine au sein des protéines en 1991. Les conformations *T-shape*, *edge-to-face* et parallèles décalés y étaient décrits comme étant les conformations les plus favorables [137]. Cependant, ces résultats ne s'appliquent qu'aux conformations impliquant des benzènes. Ici, l'ensemble des aromatiques carbonés ainsi que les cycles halogénés ont été étudiés de manière distincte. L'histidine n'a pas été considérée dans cette partie.

Dans la base de données relationnelle 3decision®, 10% de l'ensemble des interactions aromatiques sont en réalité une interaction *π-hole* (voir Figure 78). La majorité de ces cycles aromatiques ne contenant qu'un seul halogène, soit 1 207 aromatiques, il est important de distinguer les 315 cas contenant plusieurs halogènes.

L'analyse effectuée sur les aromatiques non-halogénés met en avant une forte représentation de la configuration *edge-to-face* avec 44,3% (voir Figure 78). Bien que 18,5% des aromatiques adoptent une conformation favorable *parallèle-décalée*, 27,2% soit 3 790 arrangements peuvent être décrits comme défavorables, *sandwich* ou *face-to-face*.

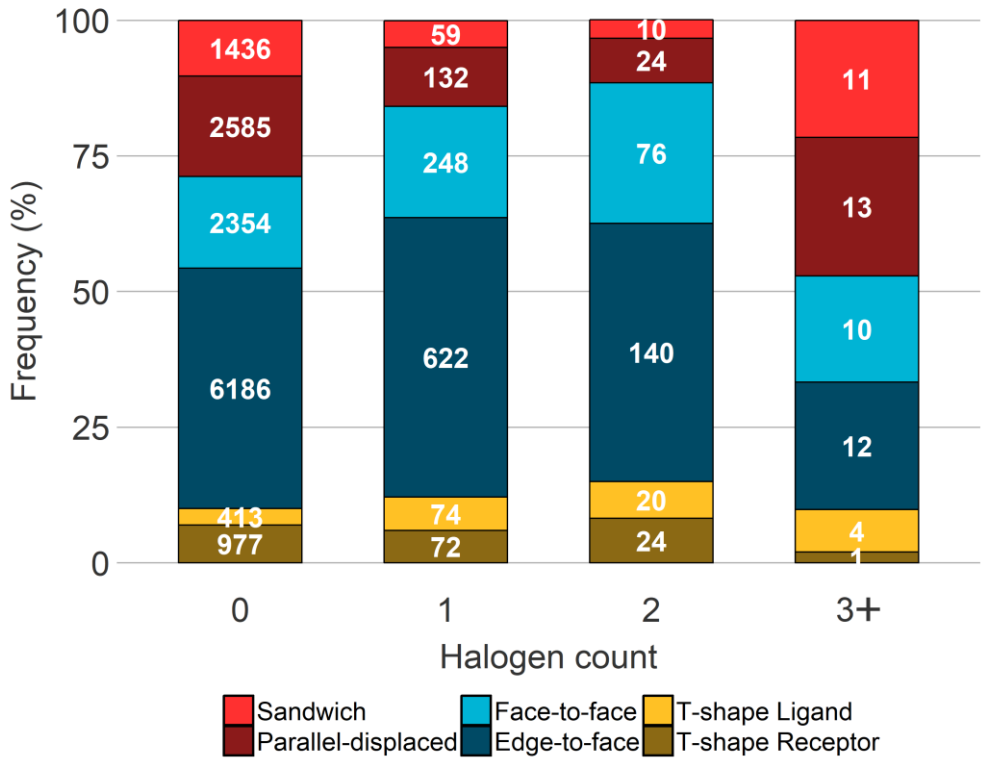


Figure 78 Distributions des conformations d'aromatiques en fonction du nombre d'halogènes présents.

Les configurations favorables pour les *π-holes* sont étonnamment sous-représentées quelques soit le nombre d'halogènes impliqués avec des propensions respectives de 21,6% et 4,6% pour les conformations *face-to-face* et *sandwich*. L'arrangement défavorable *edge-to-face* sont largement défavorables dans le jeu de données, avec 622 représentations pour les aromatiques à un halogène et 140 occurrences pour les *π-holes* à 2 halogènes.

Le test statistique Z entre deux proportions, appliquées entre les résultats obtenus pour les données aromatiques et *π-holes* (sauf 3 halogènes et plus) indique des différences significatives avec une *p-value* à 0,005. Les distributions observées pour les *π-holes* à 1 et 2 halogènes sont considérées comme étant similaires avec des *p-values* oscillant de 0,06 (*face-to-face*) à 0,35 (*sandwich*). La *p-value* maximale observée de 0,001 lors de la comparaison des aromatiques et des *π-holes* à un halogène confirme la différence de distribution. La confrontation entre les cycles aromatiques carbonés et les *π-holes* halogénés révèlent des distributions différentes à l'exception des proportions observées pour les arrangements *edge-to-face* avec une *p-value* calculée de 0,29.

L'augmentation de la fréquence des conformations *T-shape* observés pour les cycles aromatiques halogénés est surprenante de par leur aspect défavorable énergétiquement, deux charges polaires positives. En revanche, la propension supérieure d'arrangements *face-to-face* pour les aromatiques riches en halogènes est attendue grâce à la présence du moment électropositif faisant face au système π.

Les composés aromatiques contenant 3 halogènes ou plus sont rarement décelés dans la PDB (seulement 51 cas). Par conséquent, l'interprétation de ces résultats doit être considérée avec parcimonie en raison de la faible présence de ces cas. À l'exception des configurations *T-shape* (9,8% si combiné), toutes les autres configurations sont également représentées avec un minimum de 12 observations pour un arrangement *edge-to-face* jusqu'à un maximum de 13 occurrences pour les aromatiques *parallèles-décalés*.

Chapitre 4 : Interprétations

A. Pertinence du jeu de données et incertitudes

Depuis l'article publié par Clark et collaborateurs [138] notamment, une augmentation constante du nombre de ligands halogénés est observée dans la PDB, à l'exception de l'iode (voir Figure 79). Ce taux inférieur d'incorporation de l'iode dans le ligand est dû à de multiples facteurs tels que sa masse atomique élevée et inappropriée pour l'absorption de ligands ainsi que sa synthétisation coûteuse [125].

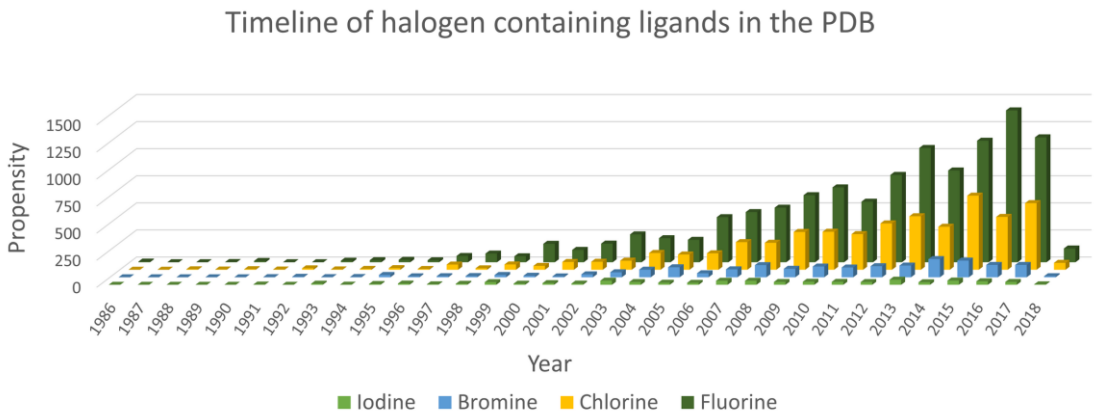


Figure 79 Histogramme de la propension de ligands halogénés ajoutés chaque année dans la PDB.

De multiples études ont analysés la propension et pertinence des halogènes dans le contexte des interactions protéine - ligand et leur intérêt dans la recherche pharmaceutique [Xu2014]. Cependant, certaines conclusions peuvent être entachées de différents biais d'analyse, le plus important étant une non-prise en compte de la redondance des données dans la PDB. De plus, se limiter à certaines interactions et leurs règles géométriques respectives restreint la description d'un halogène dans son environnement protéique.

La redondance dans la PDB est présente à plusieurs niveaux et ne doit pas être négligée. Il est ainsi possible de retrouver des complexes protéine - ligands très similaires dans une ou plusieurs entrées PDB ; ne pas considérer ces redondances contribue à la mauvaise interprétation statistique de ces données. Un exemple important est le grand nombre d'inhibiteurs du facteur X contenant un chlore conservé. Le nombre élevé de liaisons halogènes chlore – aromatique observées peut nous laisser imaginer une forte préférence à ce type d'interaction. Or, dans la majorité des cas, l'interaction observée est toujours la même.

Cette surestimation peut également être observée lors de la définition de l'interaction. Par exemple, l'utilisation des règles établies par Sirimulla et collaborateurs [129] conduit à un dénombrement individuel de chaque contact effectué par un atome aromatique comme une interaction halogène – π distincte. Dans la majorité de la littérature, une seule interaction est référencée par groupement aromatique. Ainsi, Sirimulla et collaborateurs énumèrent un nombre important d'interactions impliquant halogènes et phénylalanine, tyrosine ou tryptophane. L'utilisation de nos règles et de la détection par groupement aromatique diminue de moitié ce nombre de liaisons halogènes impliquant un cycle aromatique (voir Tableau 5).

Acide aminé	Contacts individuels	Contacts groupements aromatiques
Tryptophane	944	357
Tyrosine	1665	655
Phénylalanine	1876	906

Tableau 5 Enumération des contacts entre halogène et atomes aromatiques (méthode utilisée par Sirimulla [129]) et groupement aromatique.

Malgré le fait de ne considérer que les complexes non redondant, le biais lors de l'interprétation de tendance n'est réduit que partiellement. La représentation d'une interaction comme deux éléments interactifs est pratique d'un point de vue théorique. Néanmoins, dans le contexte protéine - ligand, les atomes sont généralement en contact, voire en interaction avec plusieurs atomes simultanément. Décrire l'environnement électrostatique de l'halogène dans son ensemble semble donc être une description plus appropriée.

B. Partenaires de l'interaction

Les travaux de Sirimulla et collaborateurs, n'utilisent qu'un critère de distance pour la détection d'interaction (somme des rayons de Van der Waals). Or la prise en compte d'angles dans la caractérisation des interactions permet de s'assurer de la directionnalité de chaque atome. Dans un second temps, ces critères permettent d'éviter la considération d'interactions aberrantes, toutes interactions ayant un angle directionnel inférieur à 70° par exemple. Le nombre important de partenaires interactifs impliqués rend l'analyse rapidement ardue, notamment par le caractère dipolaire des halogènes lourds. Pour clarifier l'analyse de ces interactions, une classification est préférable en fonction de la nature électrostatique des deux composantes.

Les interactions les plus répertoriées impliquant des halogènes, les liaisons halogènes, restent relativement rares dans la PDB avec une occurrence observé de 542 complexes. Ces observations sont en corrélation avec les résultats observés par Wilcken et collaborateurs où un nombre restreint a été relevé, illustré sur la Figure 80 [95]. Le pourcentage modéré, inférieur à 30% d'halogène engagé dans ce type d'interaction est singulier et n'apparait que dans un contexte favorable présent sur des protéines spécifiques. Cependant, une indication appuyant l'existence du *σ-hole* repose sur l'absence de moments électropositifs répulsifs en face de cette région.

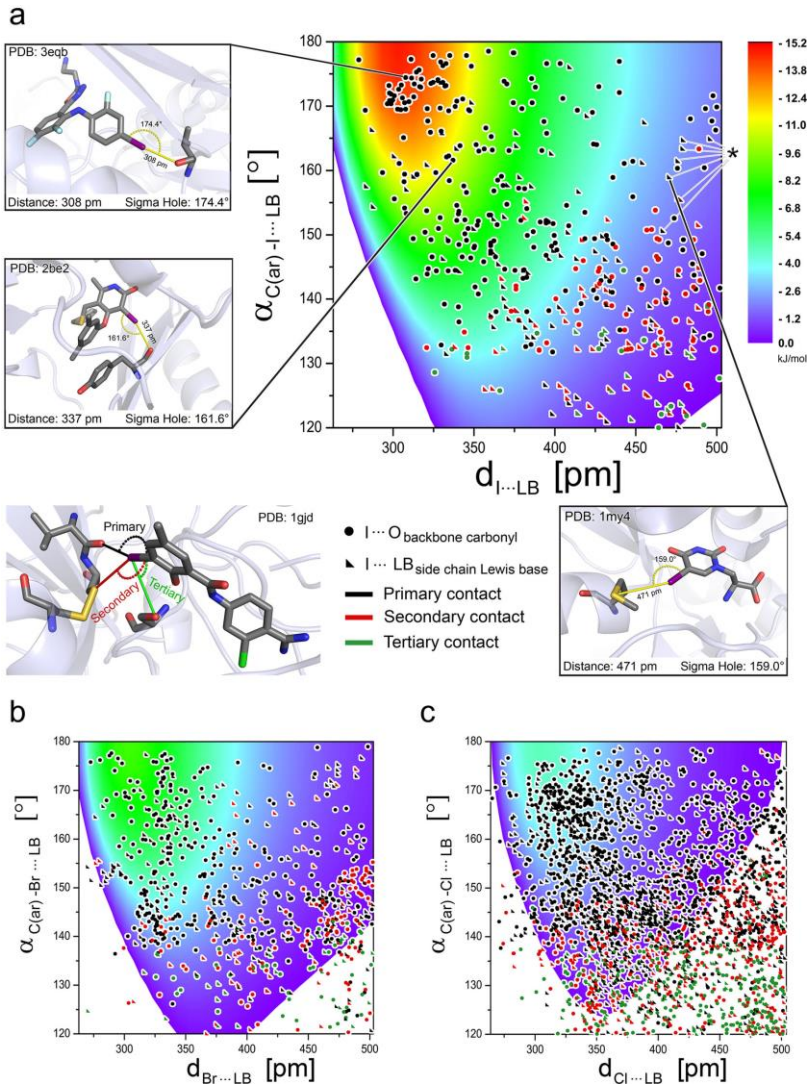


Figure 80 Distribution des liaisons halogène observés dans la PDB en fonction de la distance et de l'angle halogène pour A. l'iode B. le brome et C. le chlore. Approximations énergétiques estimées par modèle quantique [95].

En plus d'analyser les liaisons halogènes par le *σ-hole* comme Wilcken et collaborateurs [95], les liaisons hydrogènes et interactions défavorables ont aussi été étudiés. Ces dernières interactions sont négligées dans la majorité des études liées aux interactions moléculaires, pourtant leur propension est plus élevée (26,0% des atomes halogène impliqués) par rapport à la liaison halogène (8,4% des halogènes impliqués). Ces interactions pourraient également suggérer que la force de répulsion induite peut être aisément surmonté par d'autres interactions favorables dans le voisinage ou sur le reste de la molécule.

La région dénommée région neutre, caractérisée par un angle halogène autour de 125°, est en contact avec de nombreux atomes polaires. Toutefois, la corrélation de cette observation

avec le potentiel de surface électrostatique canonique des halogènes lourds [38] implique que seuls des interactions non-électrostatiques sont effectués.

La région électronégative de l'halogène amène un questionnement intéressant. Cette région est théoriquement favorable aux donneurs de liaisons hydrogène. Cependant, une propension supérieure de partenaires interactifs défavorables, accepteurs de liaisons hydrogènes, de partenaires plus défavorables (accepteurs de liaisons hydrogène) a été observés dans le voisinage du moment polaire du brome et chlore. Environ 10% des halogènes sont situés dans un environnement orthogonal défavorable par rapport aux interactions favorables. Cette observation pourrait remettre en question le rôle des halogènes lourds en tant qu'accepteurs de liaison hydrogène dans le mécanisme de fixation des ligands protéiques et donc l'interprétation de son anisotropie électronique.

Le fluor est généralement incorporé dans la conception médicamenteuse pour améliorer l'absorption, et non pas à des fins de sélectivité. Pourtant, certaines études expérimentales montrent clairement une préférence lors de l'introduction de ce type d'atomes. Néanmoins, les observations obtenues ici montrent que les interactions défavorables aux atomes de fluor sont plus fréquentes que les interactions favorables dans la PDB.

Les fluors aliphatiques et aromatiques semblent partager des propriétés analogues, des études expérimentales telles que les bioisostères de la cathepsine humaine et la MEK1 kinase [139] mettent en évidence des différences significatives entre les affinités de liaison entre -F et -CF_3. En effet, l'inhibiteur au groupement RCF_3 montre une meilleure affinité que les groupements -RF et -RCH_3 malgré l'environnement théoriquement défavorable entourant l'halogène.

Le rôle des contacts hydrophobes est complexe dans le contexte protéine - ligand. Souvent négligée en raison de l'absence de toute composant électrostatique, cette propriété peut permettre de décrire l'encombrement d'un atome interactif. Dans le jeu de données analysé, 70% des halogènes sont impliqués dans au moins un contact hydrophobe. De plus, 30% à 40% des halogènes sont entourés d'un minimum de 2 groupements hydrophobes dirigés. Ces groupements hydrophobes contribuent en grande partie à la nature électrostatique de l'environnement d'interaction.

L'introduction d'halogène dans l'optimisation pharmaceutique de cibles populaires est encore rare dans les structures cristallographiques telles que les kinases par exemple. Les principales

familles protéiques concernées par ces optimisations sont notamment les inhibiteurs de facteurs anticoagulants dont l'introduction du chlore (et même du brome) qui se retrouvent en interaction avec la tyrosine 228 aromatique augmente significativement l'affinité [127].

C. Autres rôles d'interaction

À partir du jeu de données initial contenant plus de 40 000 contacts, seuls 2 158 sont potentiellement considérées par des logiciels de détection d'interactions tels que *PLIP* ou *Arpeggio*. Dans cette étude, les autres types d'interaction n'ont pas été négligés, telles les interactions dites défavorables ou les contacts hydrophobes. La présence des halogènes dans des sous-poches interactives riches en éléments hydrophobes peut influencer grandement la liaison du ligand sur sa protéine. Cependant la quantification énergétique ainsi que le mécanisme derrière ces propriétés hydrophobes sont encore inconnues.

Les interactions entre aromatiques sont souvent considérées de manière identique peu importe la nature de l'aromatique. Or l'apport de groupements électroattracteurs modifie grandement les propriétés électrostatiques de ces derniers déplaçant le nuage électronique sur la périphérie de l'aromatique. Il existe donc des interactions de nature différentes selon la composition des aromatiques. Les arrangements géométriques entre cycles aromatiques ont été comparés en fonction de leur composition en halogène. Les résultats obtenus montrent une différence statistiquement significative dans la fréquence des arrangements conformationnels des aromatiques en fonction de la présence d'halogène sur un aromatique. Malgré un arrangement électrostatique favorable, la conformation *sandwich* est bien moins représenté en présence d'halogènes (10,3% sans halogènes, 5,1% avec) au contraire de la conformation *edge-to-face* dont la fréquence la plus forte (49,9% pour l'ensemble des aromatiques halogénés, voir Figure 78). De même, la conformation *parallel-displaced*, favorable aux interactions π-π, est plus représentée sur des cycles type benzène par rapport aux *π-holes* (18,1% et 10,5%). En revanche, l'interprétation de ces résultats est à relativiser. Le changement d'orientation d'un cycle aromatique nécessite d'une part un volume spatial dans la poche permettant cette rotation. D'autre part, des interactions effectuées par les atomes portés par l'aromatique, tels le sulfure sur certains inhibiteurs de facteurs Xa, stabilisent cette conformation et limite le changement d'orientation.

Enfin, il est important de noter que dans le mécanisme de liaison protéine - ligand, certains atomes du ligand sont plus importants dans la liaison que d'autres. Ceux-ci peuvent être observés par exemple par couplage de données expérimentales avec la structure cristallographique. Une autre solution pour identifier ces interactions primordiales et conservées réside dans la récurrence de ces interactions dans des ligands ayant des *scaffolds* communs.

D. Limites

Dans le mécanisme de liaison protéine - ligand, certaines interactions vont être plus importante que d'autres. Ces dernières peuvent être identifiées en association de données expérimentales permettant de mesurer l'affinité réelle ou bien par la multiplication des données structurales comportant le même *scaffold*.

L'absence de protonation dans la majorité des structures PDB rend l'analyse et l'interprétation des résultats potentiellement imprécises. Par exemple, le rôle dual du groupement hydroxyle en tant que donneur de liaison d'hydrogène et d'accepteur hydrogène peut générer dans certaines conditions une liaison halogène ou une interaction défavorable en fonction de l'orientation de l'hydrogène.

En outre, il est important de ne pas oublier l'environnement moléculaire lors de l'interaction protéine - ligand. D'une part, la focalisation sur des interactions très localisés dans le complexe ne permet en aucun cas d'expliquer les raisons permettant la liaison globale de la molécule sur sa cible. D'autre part, l'exemple des inhibiteurs d'Aequorine (voir Figure 67) et la prise en compte des interactions intra-protéiques, de type liaisons hydrogènes dans une hélice α, permet d'identifier les partenaires interactifs réels des halogènes.

L'importante quantité de données pour le fluor et le chlore permet d'aboutir à des conclusions fiables. Néanmoins, le nombre limité de ligands ayant du brome ou de l'iode rend l'interprétation des résultats plus ardue. Il sera intéressant de reproduire une étude similaire à l'avenir avec un jeu de données plus important.

Enfin, les structures cristallographiques représentent une image à un instant *t,* loin d'un processus dynamique. La forme du ligand et de la protéine ainsi que leurs déplacements l'un par rapport à l'autre sont des paramètres importants à prendre en compte dans l'analyse et peuvent expliquer l'absence de liaisons halogènes dues à la temporalité de la structure. Le

développement de paramètres spécifiques aux halogènes dans le champ de force, tel CHARMM [140] mais qui en est encore à ses balbutiements, laisse présager de nouvelles perspectives enthousiasmantes.

Chapitre 5 : Conclusion et perspectives

Le rôle interactif des halogènes a longtemps été restreint à la liaison halogène par sa région σ. Pourtant, nos résultats montrent que leur présence est relativement modérée au sein des complexes protéines-ligands de la PDB et reste spécifique à certains complexes. Ces résultats, déjà évoqués en 2013, permettent de mettre en valeur de nouvelles perspectives d'interaction pour ce type d'élément. Ainsi, un nombre non négligeable de liaisons hydrogènes en interaction sur le nuage électronique de l'halogène ont été mis en évidence. De plus, une quantité conséquente d'interactions considérées comme défavorables a été répertorié au sein de notre jeu de données. Ces interactions et contacts permettent de mieux appréhender la description de l'environnement d'interactions et démontrent la complexité dans la considération des interactions. Alors que ces interactions nouvellement référencées, telles les interactions défavorables, doivent encore être évaluées d'une perspective quantique, leur abondance indique un rôle non négligeable dans le processus de liaison. Ces résultats couplés aux mesures d'affinité obtenues dans les études expérimentales de type bioisostères pourront guider l'ajout d'halogènes dans l'élaboration médicamenteuse. L'approche environnementale dans la description des interactions pourra être appliqué à l'avenir à d'autres types d'atomes.

La soumission de cet article dans le *Journal of Medicinal Chemistry* a récemment fait l'objet de retours positifs de la part des évaluateurs nous incitant à effectuer certaines corrections avant réévaluation. Ces résultats ont aussi été présentés sous la forme d'un poster à la 11ème conférence de l'ICCS, *International Conference on Chemistry Structures*. Lors de cette conférence, l'importance des interactions peu décrites autour des atomes halogènes fut mis en avant par B. Kuhn lors d'une présentation orale confortant ainsi la méthode employée dans notre étude. Cette approche environnementale fait actuellement l'objet d'expérimentation en collaboration entre Discngine et un laboratoire de recherche.

Partie 4 : Etude de la diversité et redondance des complexes protéine - ligand dans la PDB

L'étude précédemment réalisée sur les interactions impliquant les halogènes a mis en évidence de manière claire le problème de la redondance importante dans la PDB, et plus spécifiquement dans l'analyse des complexes protéine - ligand. Cette redondance, connue mais peu quantifiée jusqu'à présent, peut introduire des biais dans l'analyse à grande échelle par exemple des interactions moléculaires. Afin de pallier à cette problématique, une analyse sur l'ensemble de la PDB fut réalisée à l'aide des données présentes dans 3decision® par l'intermédiaire d'alignement structuraux. Cette étude a pour but de quantifier la redondance des complexes protéine - ligand présents dans la PDB et d'identifier les complexes disposant de plusieurs modes de liaisons. De plus, une analyse sur la surreprésentation quantitative de certaines protéines et ligands a été réalisé. La génération d'un jeu de données non-redondant fut obtenu à l'issue de cette analyse et est accessible à l'ensemble de la communauté scientifique. Cette étude, présentée dans les chapitres suivants, fait l'objet d'une soumission dans l'*International Journal of Biological Macromolecules.*

Chapitre 1 : Contexte

Les structures protéiques sont le support des principales fonctions biologiques. Les multiples techniques de résolution ont permis la détermination d'un nombre important de structures tridimensionnelles ces dernières années. La PDB, comportant aujourd'hui plus de 140 000 structures, disposent notamment d'un grand nombre de complexes protéine - ligand. Ces complexes ont plusieurs utilités notamment dans l'étude des interactions dans le développement pharmaceutique [141, 142]. Ces structures protéiques sont aussi nécessaires dans l'évaluation des méthodes de modélisation moléculaire [143].

Une difficulté majeure dans les analyses et méthodes d'évaluation, *benchmarking* en anglais, repose sur l'assurance de disposer d'un jeu de données de qualité. Cette qualité est assurée par la non-redondance du jeu de données, garantissant un caractère non-biaisé.

Il existe plusieurs approches pour évaluer et générer ces ensembles non-redondants. Les méthodes les plus courantes utilisent la similarité de séquence d'acides aminés telles *PDBSelect* [144] ou *PISCES* [145]. Les approches considérant la redondance structurelle telle que PAPIA sont rares [146]. Des heuristiques sont également disponibles et faciles d'utilisation pour de grands ensembles de données comme BlastClust [86] ou CD-HIT [147]. Aujourd'hui, seuls des outils utilisant l'information de séquence permettent de récupérer de tels jeux de données dans la PDB. La redondance structurelle de cette base de données est largement reconnue par la communauté scientifique, mais peu explorée. Les seules études liées à ce sujet sont principalement axées sur les structures RMN, soit 8,5% de la PDB [148, 149], pour lesquels les différents modèles d'une protéine sont déjà répétés. Le *webserver* PAPIA [146] proposait une méthode permettant l'obtention d'un jeu de données basé sur des critères de similarité structurale. Toutefois, ce service n'est plus disponible depuis de nombreuses années.

Ces approches précédemment citées se concentrent uniquement sur la redondance protéique. Les protéines liées à l'ADN, ARN, petites molécules ou comportant des acides aminés avec des modifications post-traduction (PTM) sont généralement difficiles à analyser avec de telles méthodes. Par exemple les complexes protéine – ADN peuvent facilement atteindre des milliers d'acides aminés alors qu'un ADN est rarement composé de plus de 15 paires de bases [150]. Une situation similaire est observable pour les complexes protéine - ligand et leurs analyses.

Aujourd'hui, il existe quelques outils pour recueillir des ensembles de données de complexes protéine - ligand. *Binding MOAD* (*Mother of All Database*) [151] comprend 25 769 structures décrites comme de « qualité supérieure ». La qualité de ces complexes est jugée par différents facteurs tels que la résolution (inférieure à 2.5Å) ou la taille du ligand. Cependant, cette base de données n'a pas l'air d'être mise à jour depuis 2014. *PDBBind* [152] fournit chaque année de nouvelles versions et contient actuellement 17 900 complexes biomoléculaires dans la version 2017. La présence d'un complexe dépend entre autres de la nécessité de disposer de données expérimentales (K_d, K_i, IC_{50}). *scPDB* [70], une base de données contenant des annotations et structures de sites de liaisons, comporte 4 782 protéines et 6 326 ligands dans sa version 2017. Une absence de redondance est mentionnée dans la publication originale [153], cependant aucune métrique de similarité ne semble être utilisée pour définir cette

absence dans la construction des données. Bien que ces bases de données proposent des structures de qualité, généralement définies par résolution, aucune d'entre elles n'explorent la diversité et redondance structurale des modes de liaisons.

Wallach et Lilien en 2009 [154] se sont déjà partiellement intéressés à cette problématique. Un jeu de données non-redondant fut obtenu à partir de similarité de séquence par BLASTp [86] et des similarités de *fingerprints* moléculaires. Néanmoins, les cas où des ligands identiques se lient à différents sites de liaison sur la même protéine ne semblent pas être pris en compte. De plus, aucune évaluation structurale n'a été réalisée dans cette étude. Enfin, la dernière mise à jour des données semble avoir été effectuée en 2013. Drwal et collaborateurs ont récemment publié une étude portant sur 2 911 complexes de la PDB, 1 079 fragments et 1 832 petites molécules [155]. Les résultats indiquent que 74% de l'ensemble des données disposent de mode de liaison conservés. La conservation du mode de liaisons des fragments semble être indépendant des substitutions effectuées sur ce fragment.

Nous proposons ici pour la première fois une évaluation quantitative de la redondance et diversité structurale observée dans la PDB sur les complexes protéine - ligand. La surreprésentation de certaines protéines et ligands vont être illustrés. Les diverses conformations de chaque complexe vont être regroupées afin de définir le nombre exact de représentations uniques du complexe protéine - ligand. Un jeu de données non-redondant résultant de ce regroupement est proposer pour de futures analyses d'interaction moléculaire ou projets de *Virtual Screening*. Enfin, nous discuterons et illustrerons certaines découvertes inattendues.

Chapitre 2 : Données et méthodes

A. Jeu de données

Le jeu de données initial repose sur les structures présentes dans la PDB à la date du 22 août 2018. 128 843 structures protéiques obtenues par diffraction des rayons X, RMN et résolution cryo-EM ont été traitées et analysées au sein de 3decision®. Afin de faciliter l'identification d'un résidu à travers plusieurs entrées, les structures ont été annotées à l'aide de plusieurs sources telles que UniProtKB [156], ChEMBL [157], PFAM [158], et PROSITE [159]. Un ensemble de filtres a été appliqué sur les ligands afin d'éviter des molécules type lipides par

exemple. Ainsi, seules les molécules ayant au moins un cycle, un poids moléculaire entre 250Da et 850Da et une proportion de liaisons simples inférieure à 90% ont été conservées. Les chaînes protéiques interagissant avec la molécule ont été caractérisées en utilisant leur identifiant UniProtKB.

Les résidus considérés comme appartenant au site de liaison ont été définis à l'aide des contacts intermoléculaires présents dans 3decision®. Un site de liaison doit comporter un minimum de 3 acides aminés en contact avec le ligand pour être considéré.

Les complexes protéine - ligand sont analysés sous leur forme monomérique, un complexe sera composé d'une seule chaîne protéique et d'un ligand. Ainsi, un ligand en contact avec plusieurs chaines distinctes se verra séparer en plusieurs monomères. De même, il est possible d'avoir plusieurs représentations du même ligand sur la même protéine mais sur des sites de liaisons différents.

La comparaison de deux conformations distinctes est alors réalisée lorsque (i) deux chaines protéiques sont identiques, soit même annotation UniProtKB (ii) deux ligands sont identiques, même *fingerprint* SMILES (iii) un minimum de 3 résidus effectuant des interactions en commun afin de garantir une superposition.

B. Superposition et évaluation

La superposition des complexes a été effectuée sur les résidus environnants du ligand. Au moins un contact entre un atome du ligand et un atome de l'acide aminé est requis pour la prise en compte de ce résidu.

Les annotations UniProtKB assignées à chaque structure nous permettent d'identifier des résidus identiques entre différentes structures, indépendamment de la numérotation dans la structure. Par conséquent, sur la base de cette annotation de séquence-structure, l'alignement structural a été effectué sur les carbones α des résidus voisins entre deux complexes. Le calcul de la matrice de rotation permettant l'alignement a été réalisé selon l'algorithme de Kabsch [160] :

$$H = P^T Q \ (matrice\ de\ covariance)$$
$$H = USV^T \ (décomposition\ en\ valeurs\ singulières)$$
$$d = \det(VU^T) \ (correction)$$
$$R = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} U^T \ (matrice\ de\ rotation)$$

Où *P* et *Q* sont deux jeux de coordonnées centrés sur l'origine du système. Cet algorithme itératif pondère plus favorablement à chaque étape les résidus dont la correspondance spatiale est favorable à la minimisation de l'écart quadratique moyen, *Root-mean-square deviation* (RMSD) en anglais.

Le RMSD est une mesure permettant de quantifier la distance moyenne séparant deux entités, dans notre cas les conformations de deux ligands. Communément, le RMSD est calculé par la racine carré de la somme des distances séparant chaque atome :

$$RMSD = \sqrt{\sum_{i=1}^{N} \delta_i^2}$$

avec *N* nombre d'atomes et δ_i la distance entre les atomes. Cependant, dans notre cas il est nécessaire d'établir au préalable la correspondance entre chaque atome dont la distance va être mesurée. Or, il existe dans la PDB un nombre de cas non négligeables où la correspondance entre atomes est ambiguë, l'exemple le plus connu portant sur les aromatiques et substituants d'aromatiques. Dès lors, un RMSD, nommé *wRMSD$_f$* pour *weighted RMSD fragment,* calculé sur le centre de masse de chaque fragment composant le ligand et pondéré par la taille du dit fragment a été calculé suivant la formule :

$$wRMSD_f = \sqrt{\sum_{j=1}^{K} \frac{k_{atoms\_j}}{k_{atoms\_lig}} \delta_j^2}$$

avec *K* le nombre de fragments, k_{atoms_j} le nombre d'atomes dans le fragment *j*, k_{atoms_lig} le nombre total d'atomes dans la molécule et δ_j la distance entre le centre du fragment de masses. L'approche de fragmentation moléculaire est identique à celle effectuée dans le protocole de détection de contacts.

Les distances séparant les centres de masse de chaque fragment ont été conservés ainsi que le RMSD calculé sur les résidus alignés lors de la superposition afin d'évaluer (i) la qualité de la superposition et (ii) la diversité structurale du récepteur.

C. Seuils et méthodes de regroupements

a. Seuils considérés

Afin de définir si deux conformations sont identiques ou non, des valeurs seuils doivent être déterminées au préalable. Deux conformations d'un même complexe sont considérées comme similaires lorsque (i) la valeur de $wRMSD_f$ est inférieure à 1,0Å et (ii) aucun centre de masse n'est distant de plus de 1,5Å. Les conformations d'un même complexe ont été regroupés ensemble et comparé 1 à 1.

b. Regroupement hiérarchique

Pour des complexes comportant des valeurs $wRMSD_f$ au-dessus de 1,0Å, un regroupement hiérarchique, en anglais *hierarchical clustering*, a été réalisé. Cette méthode nécessite dans un premier temps une matrice de distances, la distance ici étant une métrique mesurant la dissimilarité entre deux conformations. Le $wRMSD_f$ a été utilisé comme métrique de distance. Dans un second temps, l'algorithme va regrouper les éléments de faible dissimilarité ensemble (étape 1 de la Figure 81). Puis, itérativement par valeur de dissimilarité croissante (étape 2 et 3 de la Figure 81), les éléments vont être regroupés par des liens jusqu'à ce que l'ensemble des éléments soit regroupés (étape 4 de la Figure 81). La résultante de cette agrégation est représentée sous la forme d'un dendrogramme. Enfin, les groupes sont déterminés soit par le nombre de groupes désiré par l'utilisateur, soit par la hauteur à laquelle le dendrogramme sera coupé.

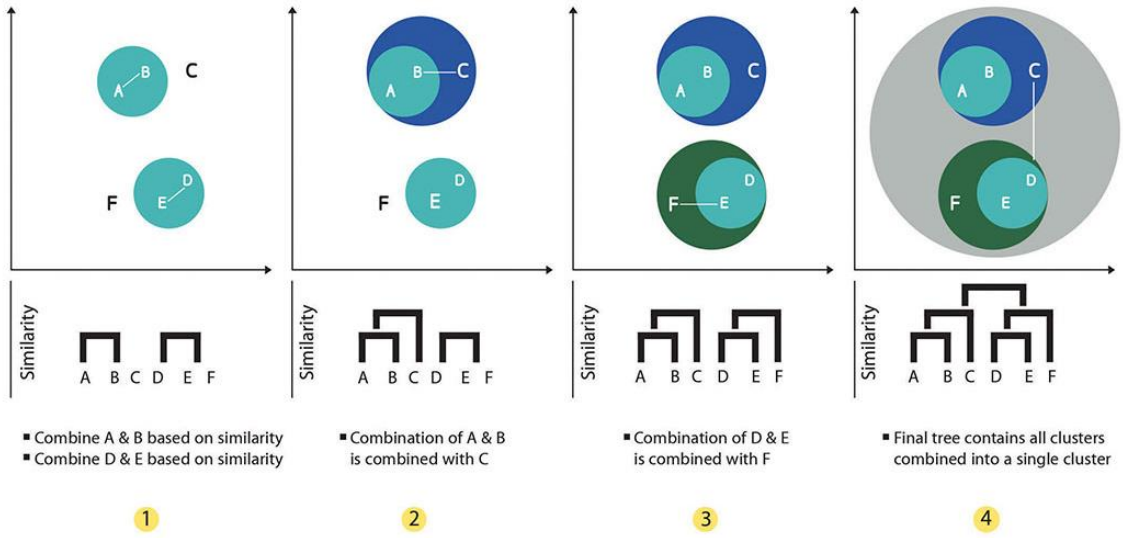


Figure 81 Etapes d'un regroupement hiérarchique et construction du dendrogramme correspondant. 1. Les couples A, B et E, D ont une distance de dissimilarité faible induisant un premier regroupement. 2. A un seuil de dissimilarité croissant, l'élément C est rapproché du sous-groupe A, B à une échelle supérieure du dendrogramme. 3. Le processus est répété itérativement jusqu'au regroupement de F. 4. Le dendrogramme est enraciné.

La méthode d'agglomération choisi ici est la méthode du diamètre, *complete-linkage* en anglais. Le regroupement a été réalisé par l'intermédiaire d'un script R (version 3.4.4 [161]).

c. Classification des complexes

Les complexes protéine - ligand sont ensuite catégorisés en fonction des comparaisons effectuées entre conformations. Les complexes n'étant présent qu'en un exemplaire dans la PDB sont caractérisés comme complexes *uniques*. Les complexes dits *homogènes* sont les complexes dont les conformations ont été identifiés comme toutes identiques. Enfin, les complexes *hétérogènes* disposent de plusieurs conformations définis comme distinctes.

D. Facteurs B

Les facteurs B ont été récupérés pour les structures PDB et normalisés selon la méthode utilisée par Bornot et collaborateurs [162]. Seuls les facteurs B des atomes du ligand et du récepteur ont été considérés dans notre étude. Les valeurs atomiques ont ensuite été moyennées et considérées séparément pour le ligand, le squelette peptidique et la chaîne latérale de chaque acide aminé.

Chapitre 3 : Résultats

A. Composition du jeu de données initial

a. Statistiques

À la date du 22 août 2018, la requête initiale a permis de récupérer 110 735 complexes interactifs à partir de la base de données 3decision®. Le filtrage des ligands réduit ce jeu de données à 92 475 complexes répartis à sur 39 411 entrées PDB (voir Figure 82). La distinction des complexes multimériques en conformations monomériques résulte dans la génération de 104 777 conformations monomériques.

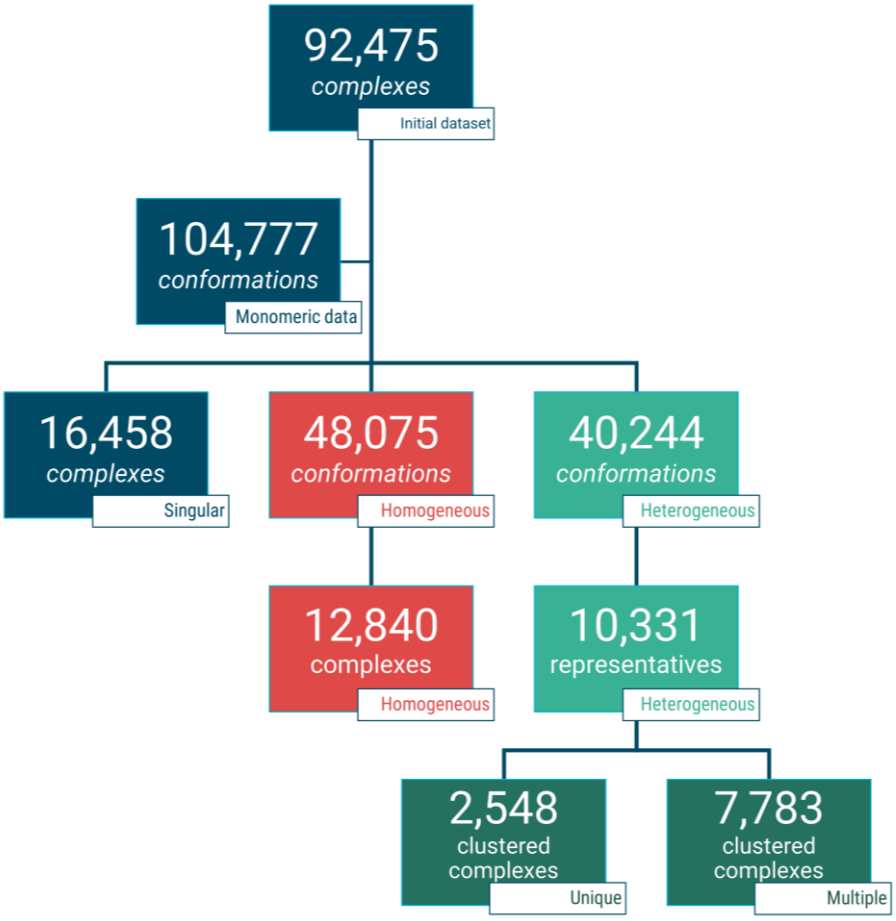


Figure 82 Organigramme récapitulatif des données initiales et des résultats issus du regroupement.

b. Diversité

Les distributions des ligands, protéines et complexes ont été étudiées ensuite. De manière attendu, le ligand le plus représenté dans la PDB est l'hème avec 9 088 conformations distinctes soit 7,9% de l'ensemble des ligands. Parmi les autres ligands les plus représentés, les dérivés de bases azotés tel que l'adénosine-triphosphate (ATP) ou la flavine (FAD) sont aussi fréquemment observés. Ainsi 17 135 ligands uniques sont retrouvés dans notre jeu de données après regroupement *fingerprint* SMILES.

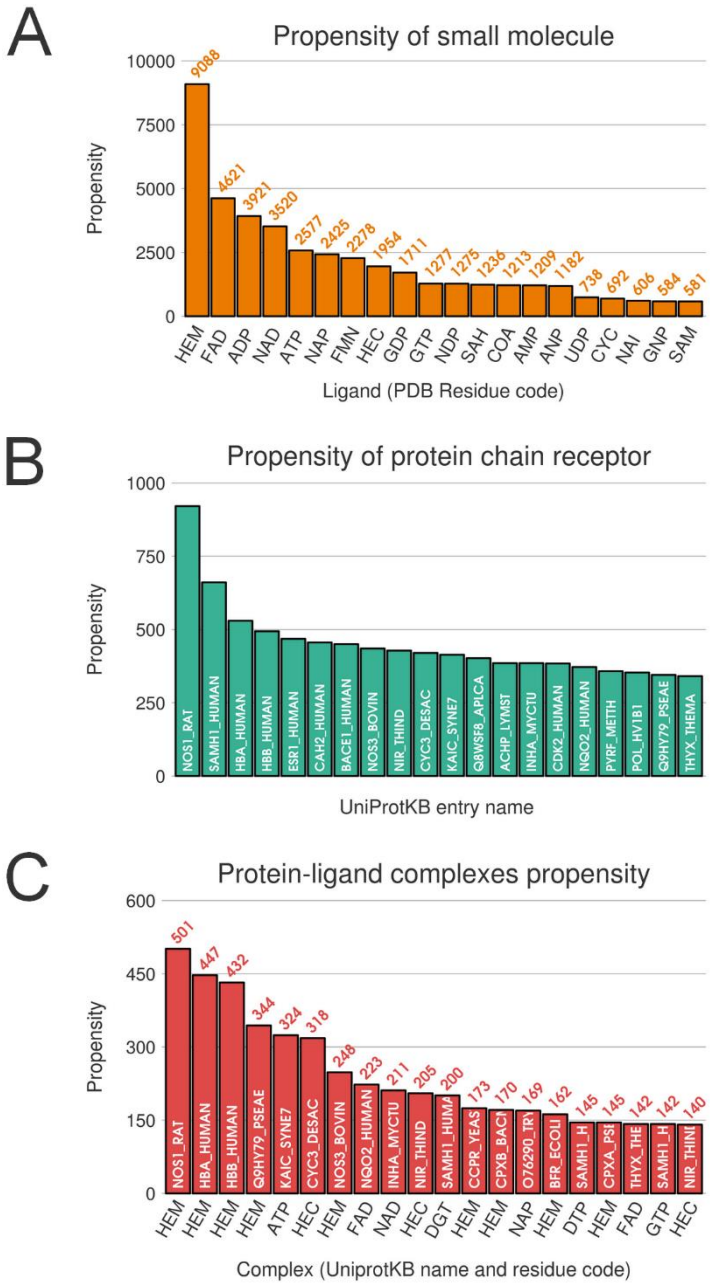


Figure 83 Distribution des 20 éléments les plus récurrent du jeu de données monomérique. A. distribution des ligands, B. distribution des chaînes protéiques et C. distribution des complexes protéine - ligand.

Ce déséquilibre observé dans la propension des ligands est aussi largement présent sur les protéines. Bien que 9 997 protéines distinctes soient identifiées dans notre jeu de données, l'oxyde nitrique synthase (code UniProtKB NOS1_RAT) présente dans le cerveau est représenté par 921 occurrences. La surreprésentation de l'hème est corrélée à la forte présente de protéines sanguine, telles que les sous-unités α (code UniProtKB HBA_HUMAN) et β (code UniProtKB HBB_HUMAN) de l'hémoglobine avec respectivement 530 et 494 cas (voir Figure 83B). Ces deux histogrammes indiquent déjà l'importance de la redondance dans la PDB et exposent les biais apparents lors de certaines analyses.

La redondance observée sur les protéines et ligands est par conséquent retrouvé dans les complexes. L'analyse de la redondance des complexes se fait conventionnellement par la considération de la similarité de séquence et de ligand. L'application de cette approche sur notre jeu de données, par l'intermédiaire du *fingerprint* SMILES et de l'annotation UniProtKB, résulte en 30 873 complexes distincts. Parmi ces complexes récurrents, l'hème en interaction avec l'oxide nitrique synthase est le plus fréquent avec 501 occurrences.

Cependant, cette approche ne tient pas compte qu'un même ligand peut potentiellement se lier à plusieurs endroits de la même chaine. L'ajout de contraintes telles qu'un chevauchement de 3 résidus en commun pour considérer le même site de liaison, aboutit à 34 936 complexes uniques, soit une hausse de 13% en ne considérant que similarité de séquence et de ligand. La fréquence du complexe hème – nitrique oxyde synthase n'est pas modifié puisqu'un seul site de liaison est présent (Figure 83C).

B. Complexes protéine - ligand uniques

Les complexes protéine - ligand dits *uniques* sont caractérisés par la présence d'une seule occurrence dans la PDB. Elles correspondent à 15,7% de notre jeu de données, soit 16 458 complexes sur les 104 777 monomères initiaux (Figure 82).

Parmi ces complexes, une forte récurrence des ligands de type base azotée est observée avec 739 occurrences combinées pour l'adénosine mono- (code PDB *AMP*), di- (code PDB *ADP*), triphosphate (code PDB *ATP*) et la phosphoaminophosphonique (ANP). L'hème est en revanche moins représenté dans ce sous-groupe puisque seul 174 cas (code PDB *HEM*) ont été dénombré ainsi que 152 cas de protoporphyrine IX (code PDB *HEC*). Toutefois, la grande majorité des ligands, 84,7%, ne sont représentés qu'une seule fois.

Ces 16 458 complexes décrivent 5 239 protéines distinctes. L'étude approfondie des cibles récurrentes nous permet d'identifier et de mesurer la diversité des ligands cristallisés avec la protéine. Cette distribution, largement déséquilibrée, met en avant 259 complexes de l'anhydrase carbonique 2 impliquant différents ligands. La Figure 84 illustre le *scaffold* commun des inhibiteurs de l'anhydrase ainsi que le mode de liaison conservé au niveau de ce scaffold. Une redondance semblable est constatée pour la prothrombine mais aussi pour la β-secrétase 1 et la kinase dépendante de cycline de type 2 avec respectivement 204, 199 et 195 occurrences. Contrairement aux ligands, les chaines protéiques sont fréquemment représentées par plusieurs conformères puisque 16,6% de ces complexes uniques n'ont qu'une seule occurrence. Les chaines présentes à plus d'un exemplaire dans le jeu de données sont en moyenne représenté par 6,02 occurrences.

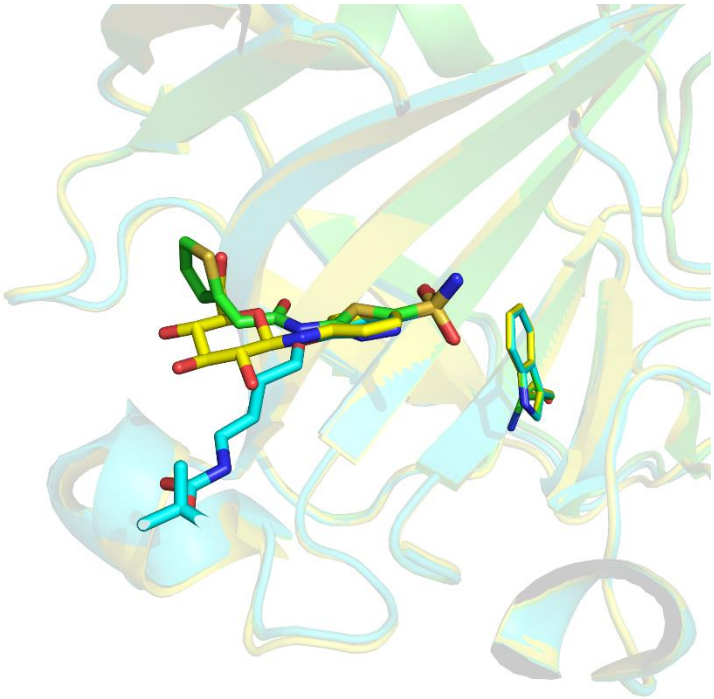


Figure 84 Représentation 3D de différents ligands liés au récepteur de l'anhydrase carbonique 2 (codes PDB 4iwz en vert, 2rfc en bleu, et 2hl4 en jaune).

Ces redondances spécifiques indiquent clairement des tendances générales vers quelques cibles protéiques d'intérêt dans le domaine de la recherche biologique. Cependant, la répétition de ces protéines couplées à des ligands différents nous donne aussi des indications sur les fragments fréquemment intégrés dans leurs conceptions, mais aussi l'importance de certains résidus dans le mécanisme de liaison. Par exemple, 95,1% des cas de prothrombine

impliquent les résidus tryptophane 215 et 94,1% l'alanine 190 dans le mécanisme de fixation. D'autres résidus sont impliqués de manière plus sporadiques tels que l'acide glutamine 217 présent dans 55% des complexes et la phénylalanine 227 présent dans 24,5% des complexes (voir Figure 85).

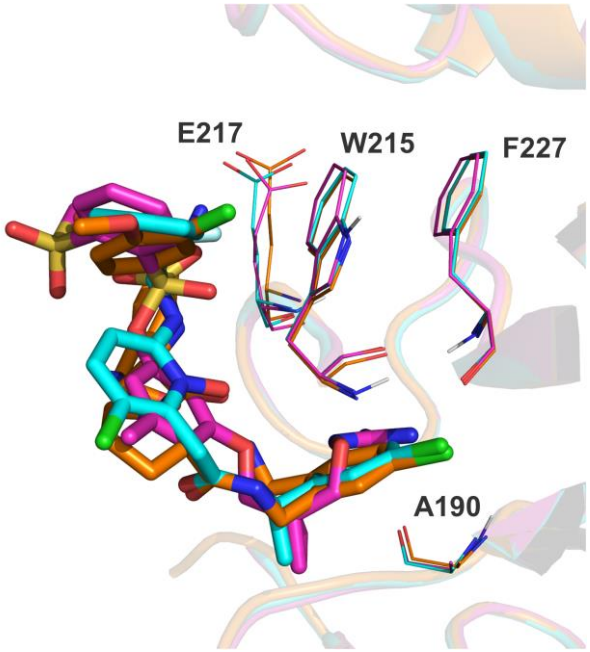


Figure 85 Conservation du site de liaison de la prothrombine et des résidus récurrents dans le mécanisme de liaison (codes PDB 3u9a, 1z71 et 1t4u).

C. Regroupement des conformations d'un complexe

Les conformations restantes, 88 319 cas (84,2% de notre ensemble de données), ont été réunies en groupe en tenant compte des ligands, des chaînes protéiques et des sites de liaison identiques (3 résidus de liaison communs), correspondant à diverses conformations du même complexe. Chaque conformation au sein d'un complexe ou groupe doit être comparé aux autres conformations. La propension du nombre de complexes par groupe montre une distribution déséquilibrée soulignant une redondance élevée vers des sites de liaison comportant 2 conformères du même ligand dans la PDB (9 542 groupes détectés) (voir Figure 86). 251 complexes comportant plus de 30 conformations distinctes, soit 1,4%, représentent en réalité 17,8% des conformations de cette catégorie.

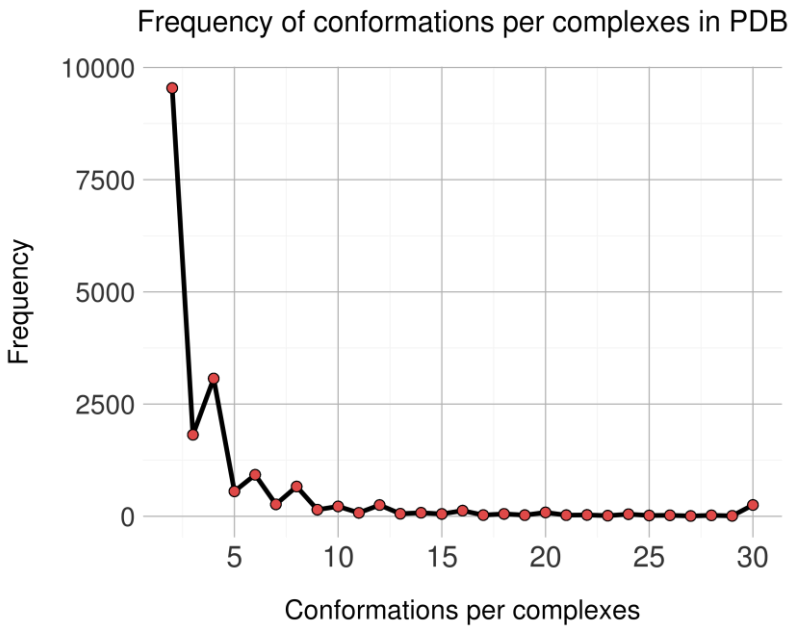


Figure 86 Distribution du nombre de conformations observés par complexe.

Cette première phase de regroupement aboutit à un total de 18 482 groupes ou complexes comportant entre 2 et 501 conformations de ligand. Deux conformations issues du même complexes sont superposées sur les résidus environnants en commun. La différence structurale des ligands, quantifiée par le $wRMSD_f$, est enregistrée dans une matrice de similarité de $\frac{n(n-1)}{2}$ comparaisons (n nombre de conformations dans le groupe). Ainsi, plus de 1 130 000 comparaisons avec calcul du $wRMSD_f$ ont été effectués. Les complexes ont ensuite été caractérisés selon l'homogénéité ou l'hétérogénéité des conformations.

a. Complexes homogènes

Des complexes sont dits homogènes si les multiples conformations du ligand dans un même site de liaison sont toutes identiques d'un point de vue structurale. Cette caractérisation se traduit par l'absence de valeurs de $wRMSD_f$ supérieure à 1,0Å et aucune distance entre fragments alignés supérieure à 1,5Å.

12 840 complexes ont été identifiés comme homogènes dans notre jeu de données, équivalent à 48 075 complexes (soit 45,9% de l'ensemble des monomères). Étant donné que les modes de liaison sont identiques dans chaque groupe, il est possible de considérer une réduction d'un facteur 3,75 pour ces complexes protéine - ligand, résultant donc de 12 840 représentants uniques.

Des résultats intéressants sont observés dans des complexes comportant un nombre important de conformations. Ainsi, 173 conformations de l'hème lié au cytochrome C provenant de 158 entrées PDB distinctes sont caractérisées comme étant toutes identiques malgré l'absence de structure RMN. Un exemple d'un mode de liaison conservé issu de ce groupe peut être visualisé sur le code PDB *1aen*.

La majeure partie de nos comparaisons implique un nombre identique de résidus en contact permettant la superposition de deux conformations. Cependant, pour 160 complexes homogènes (2 154 conformations), on retrouve des comparaisons dont le nombre de résidus en commun est inférieur aux trois quarts du nombre maximal de résidu en commun dans ce groupe. L'étude visuelle de certains de ces cas spécifiques révèlent les raisons de ces différences telles que (i) des résidus détectés comme en contact dans un seul des deux complexes, potentiellement flexibles (en pointillé sur la Figure 87), et (ii) des résidus manquants dans l'une des structures (représentés en rouge sur la Figure 87). Malgré ces quelques cas, le mode de liaison du ligand sur sa protéine reste homogène et n'impactent donc pas l'approche utilisée ici. Au contraire, elle peut mettre en évidence des fluctuations d'interaction pour des résidus flexibles dans certaines circonstances.

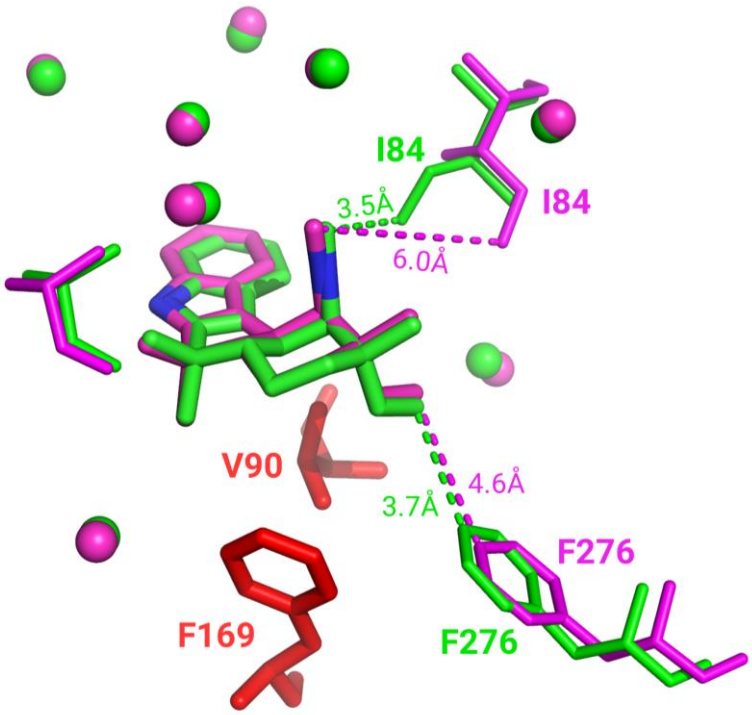


Figure 87 Représentation 3D de la protéine WelO5 en complexe avec le ligand 6C, représenté sous forme de battons (codes PDB 5iqv vert et 5iqu magenta). Les résidus non résolus dans le fichier PDB 5iqu sont colorés en rouge. Les distances divergentes entre les deux conformations sont représentées sous forme de tirets. Les résidus correctement superposés sont affichés en représentation sphérique.

L'utilisation des Cα dans la superposition des poches explicites permet d'éviter partiellement le problème de flexibilité des chaînes latérales. 97,1% de nos comparaisons homogènes affichent un RMSD de superposition de poche inférieur à 0,5Å, correspondant donc à des poches très conservées structuralement. De manière plus générale, le RMSD moyen calculé sur l'ensemble des poches est de l'ordre de 0,18Å avec un écart-type de 0,13Å.

Seules 615 comparaisons sur 233 417 affichent un RMSD du site de liaison supérieur à 1,0Å. Ces cas spécifiques sont en grande partie induits par des régions flexibles telles des boucles mise en avant sur la Figure 88. Ce cas spécifique de flavine sur son récepteur adénine dinucléotidique met en avant un nombre conséquent de résidus divergents structuralement. La Glycine 397 faire partie d'une boucle flexible menant à une distance mesurée de 11,7Å entre les deux conformations après alignement (en orange sur la Figure 88). Le Glutamate 49 et la Valine 395 affichent également une divergence importante avec des distances de séparation mesurées respectivement à 7,4Å et 3,0Å. Là encore, ces différences n'affectent pas la pose du ligand et révèlent de potentiels rôles interactifs différents pour certains résidus.

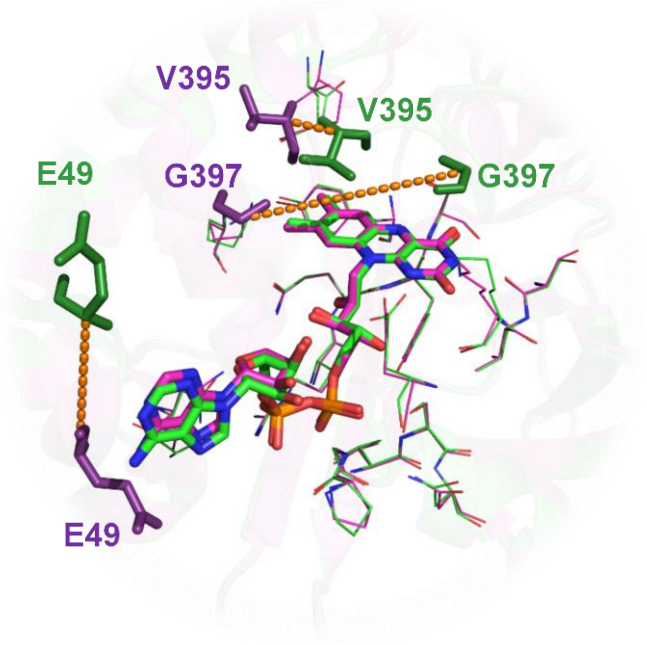


Figure 88 Représentation 3D des sites de liaisons de la flavine (codes PDB 1cqx en vert et 3ozv en violet). Les résidus structurellement conservés et correctement alignés sont affichés dans la représentation type 'ligne'. Les résidus flexibles sont affichés en représentations 'sticks' et leurs déviations respectives en pointillés.

b. Groupes hétérogènes

Les 5 638 autres complexes, soit 40 244 conformations (38,4% du jeu de données monomériques), présentent tous au moins un $wRMSD_f$ supérieur à 1,0Å ou une distance entre deux fragments alignés supérieures à 1,5Å. Si toutes les comparaisons au sein d'un groupe sont supérieures à 2.0Å, l'ensemble des conformations sont alors considérées comme distinctes dans le groupe. Dans les groupes où les valeurs de $wRMSD_f$ étaient variables, une étape de regroupement hiérarchique a été réalisée afin d'identifier les conformations uniques de chaque complexe.

Les résultats mettent en évidence 10 331 modes d'interactions distincts pour ces 40 244 conformations, soit une réduction potentielle d'un facteur 3,89. 2 360 conformations proviennent de groupes où seules des conformations distinctes étaient présentes. 2 548 complexes ne disposent que d'une conformation représentative après l'étape de regroupement. Les 5 423 conformations représentatives restantes ont été récupéré par la méthode de regroupement hiérarchique dont la majorité est composé de 5 conformations identiques (voir Figure 89).

La Figure 89 souligne le nombre de conformations initiales par complexes en fonction du nombre de représentants obtenus via regroupement par complexes. En dépit d'avoir initialement 22 complexes avec plus de 100 conformations initiales distinctes, aucun des complexes regroupés n'est décrit par plus de 30 conformations, seuls 4 en ont plus de 20.

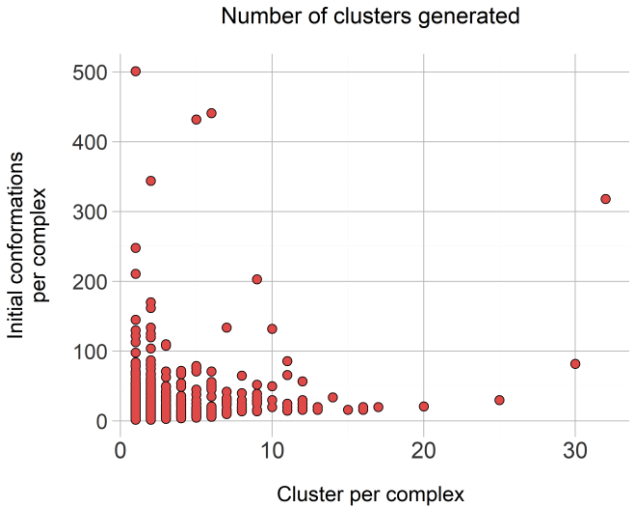


Figure 89 Répartition du nombre de conformations représentatives générés par regroupement par rapport au nombre de conformations initiales dans les complexes hétérogènes.

Les 501 conformations du complexe le plus récurrent dans notre jeu de données, l'hème lié à l'oxyde nitrique synthase, ont été regroupés en une conformation représentative avec un $wRMSD_f$ moyen de 0.37Å (écart-type de 0.18), illustré sur la Figure 90.

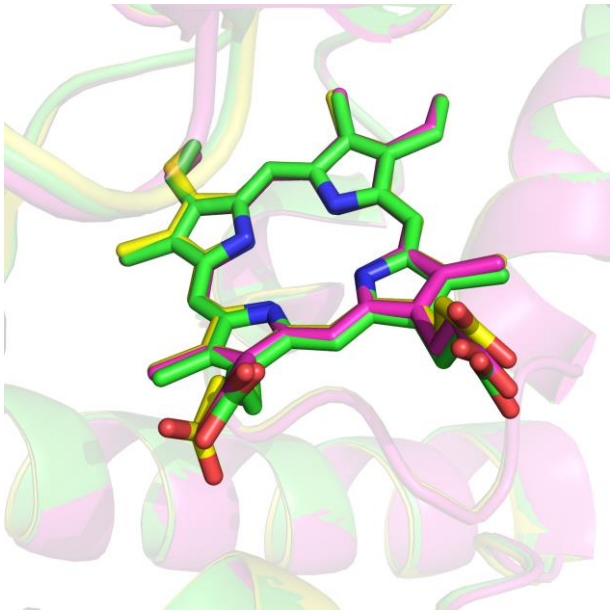


Figure 90 Conformation représentative du complexe hème - oxyde nitrique synthase (codes PDB 3n5z, 3ufq et 3n5w).

Des résultats intéressants sont aussi observés sur certaines structures RMN mis en avant par la Figure 91 par exemple. Pour la protéine xylose isomérase (code PDB 1lxf), 30 modèles initiaux ont été identifiés au sein de la structure. L'approche de regroupement a permis l'identification de 12 modes de liaison distincts, le plus grand groupe contenant 6 conformations. La Figure 91A illustre 3 conformations issues de ce groupe dont le $wRMSD_f$ moyen est de 0,73Å, alors qu'un alignement sur les autres clusters sur la Figure 91B montre une déviation structurale importante.

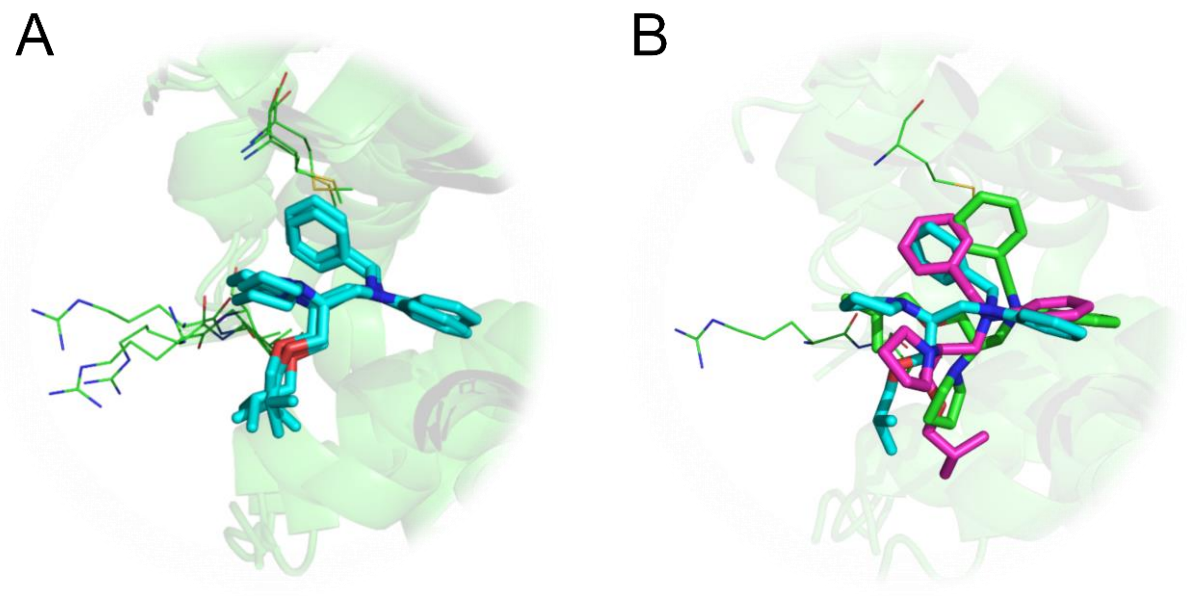


Figure 91 Illustration 3D de la troponine cardiaque en liaison avec le Bépridil (code PDB 1lxf). Le ligand est représenté sous forme de 'sticks' et les résidus permettant la superposition sont affichés en 'lignes'. A. Un groupe homogène de 3 conformations identiques (moyenne wRMSDf 0.67Å). B. 3 conformations jugées distinctes de la même structure RMN (moyenne wRMSDf 2.9Å).

Les complexes regroupés comportant 3 conformations ou plus ont été considérées comme homogènes. En effet, la répétition des conformations observées limite les incertitudes lors de la résolution de la structure. Pour les autres structures obtenues par cristallographie par rayons X et d'une résolution supérieure à 1,5Å, soit 7 642 conformations, les facteurs B ont été calculés puis analysés.

1 122 conformations affichent des valeurs normalisées de facteurs B supérieures à 2,0 pour le ligand, seuil correspondant à une certaine flexibilité comme décrit par Bornot et collaborateurs [162]. De manière intéressante, ces ligands se situent dans des environnements relativement rigides puisque seuls 4 cas de squelette peptidique (illustré en rouge sur la Figure 92) et 77 cas de chaînes latérales sont considérées en moyenne comme flexibles. Cette observation indique clairement une rigidité totale du site de liaison en opposition à la flexibilité importante du ligand. Ainsi, seul 2,7% des conformations hétérogènes peuvent être considérées comme incertaines par la flexibilité importante pouvant amener des erreurs lors de la résolution structurale.

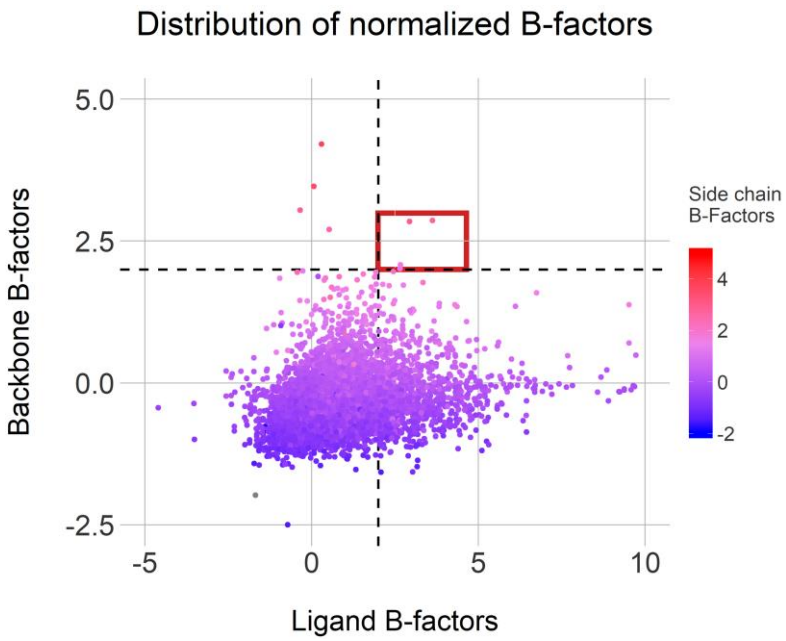


Figure 92 Distributions des facteurs B pour des complexes possédant des conformations distinctes (valeurs RMSD élevées) et complexes avec moins de 3 conformations groupées. Les valeurs des facteurs B sont normalisées. Les cas avec une grande flexibilité dans le ligand et la poche sont mis en évidence en rouge.

D. Génération d'un jeu de données non-redondant

Les résultats obtenus à travers cette étude peuvent aboutir à la génération d'un jeu de données protéine - ligand non-redondant. Les critères utilisés pour sélectionner le représentant de chaque conformation repose sur un agrégat entre la résolution de la structure, le nombre maximal de résidus en contact et la moyenne de $wRMSD_f$ pour chaque conformation. Le jeu de données final comporte 39 629 complexes, une réduction d'un ordre de grandeur 2,64 par rapport aux données initiales. Les données disponibles sous la forme d'un fichier texte à l'adresse suivante https://github.com/Discngine comporte les codes PDB, informations sur les résidus et le ligand ainsi que des informations complémentaires telles que la conformation alternative considérée.

L'utilisation d'une approche plus classique pour définir un jeu de données identiques auraient abouti à 9 997 complexes en ne considérant que la similarité protéique. L'ajout d'une contrainte de similarité de ligand auraient proposé un nombre de 30 873 complexes, soit une diminution de 22,8% par rapport à notre jeu de données final. Cette différence vient de plusieurs facteurs différents notamment la prise en compte de plusieurs sites de liaisons

distincts sur une même chaine protéique ainsi que la considération de plusieurs conformations possibles.

La composition de notre jeu de données final est ainsi déséquilibrée, 31 846 complexes, soit 80,3%, ne sont représentés que par une seule conformation. 4 555 complexes hétérogènes disposent de plusieurs représentations, 95,0% d'entre eux sont symbolisés par moins de 5 modes de liaisons différents. L'anhydrase carbonique 2, la β-sécrétase 1 et la cycline dépendante kinase 2 sont identifiés comme les complexes les plus représentés avec respectivement 332, 297 et 272 occurrences. Ces valeurs importantes peuvent être expliquées par le nombre élevé de ligands distincts cristallisés avec une protéine ainsi que le nombre élevé de conformations pouvant être adoptés.

16 771 ligands distincts sont présents, l'hème étant toujours le ligand le plus récurrent. Cependant, par rapport aux 9 088 cas d'hème observés initialement, le regroupement a permis la réduction par un facteur 8,4 du nombre de conformations d'hème distincts (voir Figure 93). De même, la flavine a vu son nombre d'occurrences diminué par un facteur 6,4 indiquant des redondances significatives pour ces molécules.

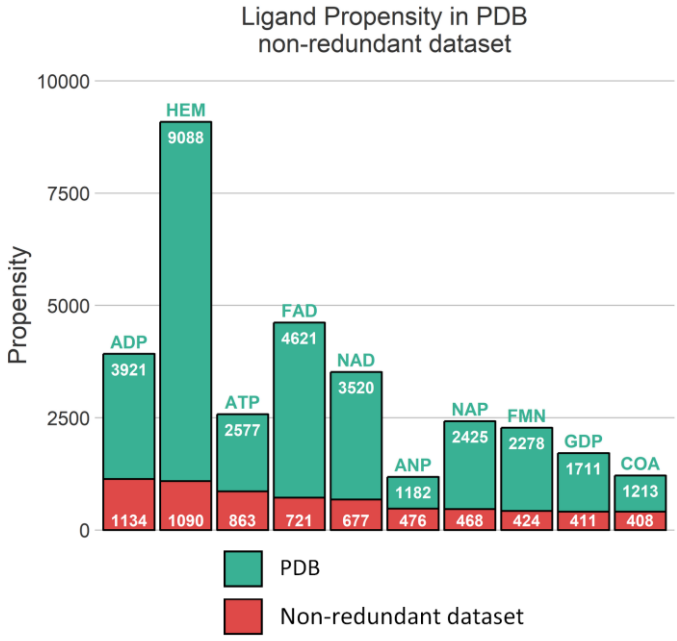


Figure 93 Distribution des 10 ligands les plus fréquemment observés dans la PDB (vert) et dans le jeu de données final non-redondant (rouge).

Chapitre 4 : Discussion

A travers cette étude, nous avons mis en évidence la conservation et diversité des modes de liaison dans des complexes protéine - ligand identiques. Ainsi, le nombre de complexes distincts dans la PDB peut être réduit à 39 629 complexes, une diminution d'un facteur 2,64 par rapport aux données initiales.

Des jeux de données protéine - ligand ont été généré dans le passé tels que scPDB [153], ils se concentrent uniquement sur des critères spécifiques tels que la résolution, ligand de type *drug design*… La génération d'un jeu de données avec des contraintes similaires sur l'ensemble de la PDB par PISCES résulte dans l'obtention de 130 000 chaînes protéiques [145]. Cette surreprésentation est notamment due à l'incapacité de regrouper des séquences à travers plusieurs entrées PDB.

L'analyse structurale effectuée ainsi que la génération d'un jeu de données peuvent être utilisé à diverses fins. Ainsi, il est possible d'identifier des résidus caractérisés comme rigides ou flexibles dans le mécanisme de liaison (voir Figure 88). La diversité de ligand peut aussi être analysé pour une protéine spécifique (voir Figure 85) tout comme l'identification de multiples modes de liaisons, illustré sur la Figure 91, utile pour des études telles que l'*Ensemble Docking*.

Nous avons illustré cette redondance multiple observée à travers la PDB dans le contexte protéine - ligand. Certaines protéines spécifiques sont largement surreprésentées et ont été analysés soigneusement dans notre analyse. Ainsi, le complexe le plus récurrent, l'hème lié à l'oxyde nitrique synthase, peut être représenté par 3 conformations différentes, comparés aux 501 présentes dans la PDB par exemple. L'ensemble des complexes impliquant l'hème est réduit de 9 088 occurrences à 1 090 représentations uniques.

De plus, l'approche utilisée ici est –nous l'espérons- non biaisé par des changements structuraux dans la chaîne protéique : les résidus manquants dans une structure ou l'insertion d'un peptide chimérique conduisent au regroupement de modes de liaisons similaires. Ces résultats permettent de valider certaines conformations obtenues à des résolutions élevés mais fréquemment observés dans la PDB.

Une comparaison entre les métriques de similarité peut être effectué. L'utilisation générale du RMSD calculé sur les atomes nécessite un *matching* prédéfinis de ces mêmes atomes. Or, il est fréquent dans la PDB d'observer des inversions de nom d'atomes (ou *label*), notamment

sur les aromatiques, parfois dans une même structure. La Figure 94 illustre un exemple typique où après superposition, les deux ligands parfaitement alignés affichent un RMSD atomique de 2,9Å et $wRMSD_f$ de 0,8Å dus à l'inversion des carbones 11 et 20 par exemple. Le coefficient de corrélation de Pearson a été calculé sur les valeurs de RMSD atomique et $wRMSD_f$ et indique une corrélation modérée (0,64) entre les deux métriques. Néanmoins, le $wRMSD_f$ a ses propres limites. La rotation spatiale du plan d'un aromatique n'est pas caractérisé par cette métrique, modifiant potentiellement la nature de l'interaction mis en jeu.

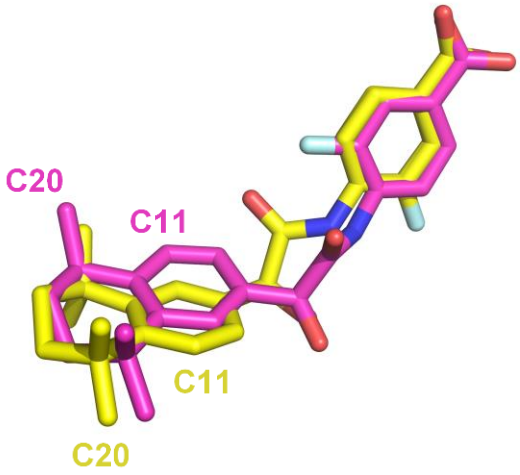


Figure 94 Illustration de l'inversion de label d'atome sur deux entrées PDB (codes PDB 4lbd en magenta et 1exx en jaune).

La valeur seuil pour déterminer la similarité entre deux complexes peut aussi être discutée. L'évaluation de méthodes tel que l'assemblage (en anglais *docking*) par exemple identifie une pose comme acceptable lorsque que le RMSD est inférieur à 2,0Å par rapport la conformation native. Toutefois les RMSD sont une mesure moyenne et peuvent être insensibles à des valeurs très divergentes. Un exemple est illustré sur la Figure 95 où une comparaison dont le RMSD est proche de 2,0Å peut être lié à (i) une déviation spatiale de la molécule entière comme illustrée dans Figure 95A, et (ii) une région spécifique du ligand conservée dans le mode de liaison en parallèle d'un fragment complètement décalé (voir Figure 95B).

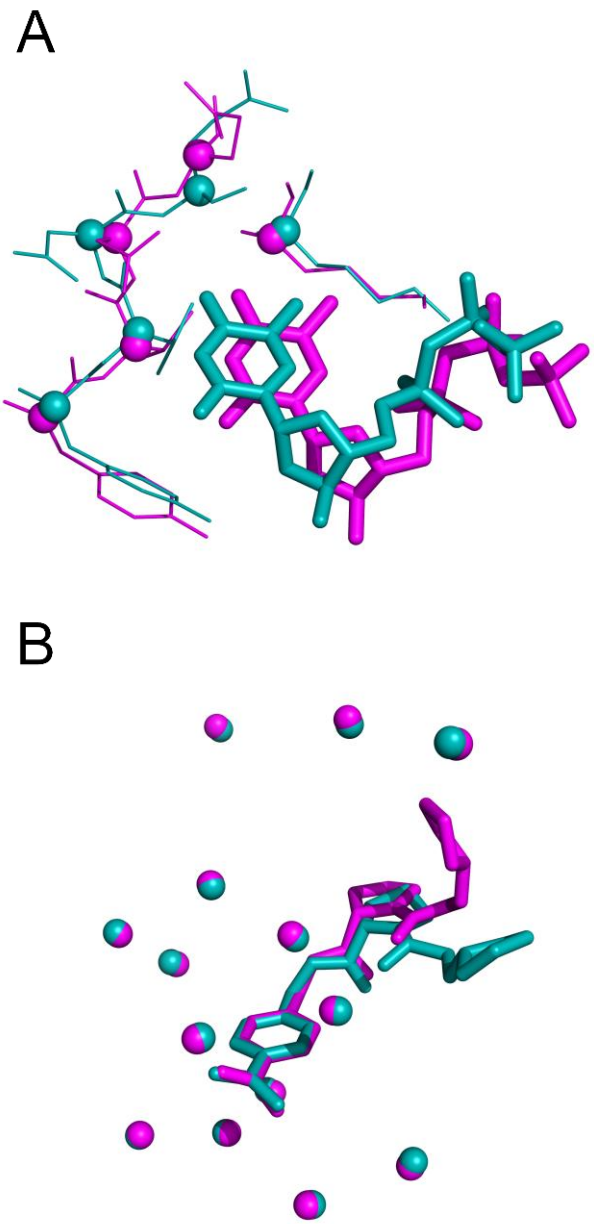


Figure 95 Illustration 3D de conformations dont le RMSD est proche de 2,0Å. A. Décalage entier de la molécule TTP par rapport à son récepteur ribonucléotide (codes PDB 3hnd en bleu et 3hnf en magenta, wRMSDf 2,08Å). B. Fragment décalé d'un inhibiteur de la poche de la trypsine S3 (codes PDB 3ljj dans 2zft bleu en magenta, wRMSDf 1,97Å).

Chapitre 5 : Conclusion et perspectives

Cette étude a permis de mettre en évidence la redondance présente dans la PDB dans le contexte protéine - ligand. Sur les 104 777 complexes monomériques initiaux, seuls 39 629 correspondent à des modes de liaisons distinct, soit une réduction d'un facteur 2,64. Ces résultats, quantifiées pour la première fois, ont permis l'élaboration d'un jeu de données non

redondants de complexes protéine - ligand. Ces données pourront être utilisées par le reste de la communauté scientifique pour des analyses globales d'interactions moléculaires par exemple. Cependant, cette redondance est aussi utile pour mettre en évidence la conservation importante de certains fragments dans le *design* des médicaments. De même, certains exemples ont montré à la fois une préservation de certains résidus ainsi qu'une mobilité importante d'autres acides aminés dans le site de liaison. Enfin, des complexes dits hétérogènes, présentant potentiellement plusieurs modes de liaisons, ont été caractérisés à travers cette étude. Ces complexes apportent des informations supplémentaires à prendre en compte pour de futures études de *docking*.

A moyen terme, ces résultats pourront être exploités dans 3decision® en (i) évaluant la qualité d'une structure par l'observation répétée de conformations identiques, (ii) proposant directement à l'utilisateur les conformations disponibles pour des complexes hétérogènes, ou (iii) à des fins prospectives dans lesquelles durant l'enregistrement d'une structure une comparaison structurale serait réalisée indiquant à l'utilisateur la diversité de son complexe. Enfin, la redondance des complexes peut être utilisé à des fins computationnelles puisqu'elle permet par exemple de limiter le nombre de comparaisons par la considération des conformations représentatives par exemple.

Partie 5 : Travaux complémentaires sur l'analyse structurale des protéines

Des travaux additionnels ont été réalisés dans le cadre des recherches menées spécifiquement par l'équipe Dynamique des Systèmes et Interactions des Macromolécules Biologiques (DSIMB) du laboratoire "Biologie Intégrée du Globule Rouge" (INSERM UMR_S1134). Ces analyses, portant sur l'étude de la structure des protéines, sont liées au sujet de ma thèse puisque le fonctionnement d'une protéine se fait par l'intermédiaire d'interactions et que dans le même temps, sa fonction est dépendante de sa structure. Les études dont la publication ont fait l'objet d'une contribution sont décrits dans les paragraphes suivants.

Chapitre 1 : Flexibilité, mobilité et déformation des différents types d'hélices

A. Types d'hélices

Une analyse du comportement dynamique des hélices a été réalisée avec le Dr Pierrick Craveur et le Dr Tarun J. Narwani menant à une publication dans le journal *Archives of Biological Sciences* (co-premier auteur) [163].

L'hypothèse de départ de cette étude repose sur l'idée (et des données de la littérature) que les structures secondaires hélicoïdales, considérées comme assez stables et rigides, puissent être soumises à des échanges conformationnels au cours du temps. Pour rappel, 3 types d'hélices sont classifiables en fonction des *patterns* de liaisons hydrogènes (et des angles ψ et ϕ). Ainsi, l'hélice α, la plus fréquente avec un résidu sur trois impliquée, présente un motif d'interaction tous les 5 résidus, 4 résidus pour les hélices 3_{10} et 6 résidus pour les hélices π (Figure 96).

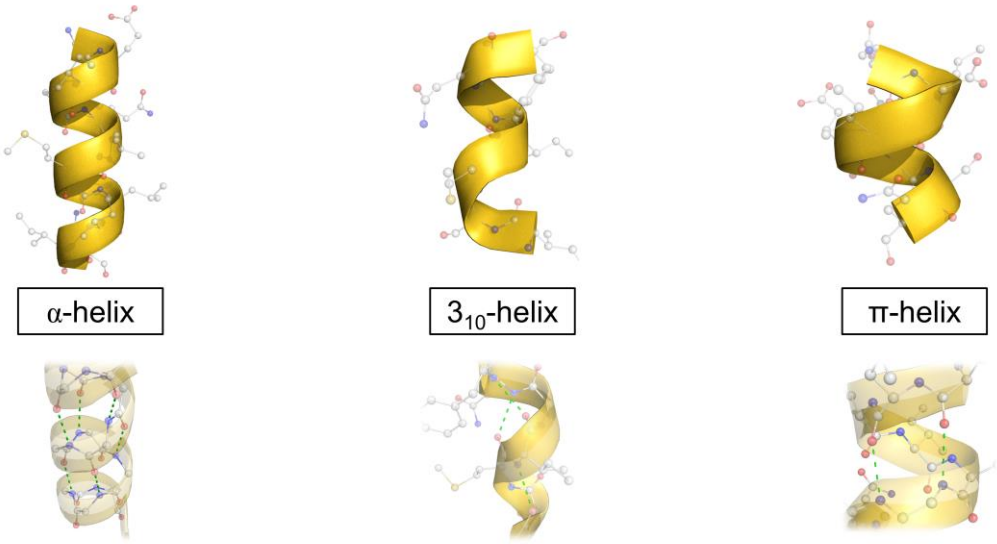


Figure 96 Représentation des 3 différents types d'hélices assignés par DSSP et leurs patterns de liaisons hydrogènes (codes PDB 1uyd, 1uyd et 2qd3, respectivement).

B. Jeu de données et métriques de flexibilité

Pour répondre à cette question, 169 structures de protéines globulaires ont été extraites de la PDB. Elles ont une résolution inférieure à 1,5Å, une taille comprise entre 50 et 250 résidus et aucun résidu n'est manquant. Les classes structurales de ce jeu de données sont équilibrées, puisque les 4 classes structurales majoritaires décrites par SCOP (α, β, α/β et α+β) sont présentes de manière équivalente. De plus, la proportion d'hélices retrouvée dans le jeu de données est identique à celle présente en moyenne dans les protéines, par exemple, un résidu sur trois est associé à une hélice α.

3 dynamiques moléculaires (voir Annexe 2) indépendantes de 50ns ont été réalisées pour chacune des 169 structures du jeu de données à l'aide du logiciel *Gromacs* 4.5.7 [164] avec le champ de force *Amber99sb* [165]. Des conformations ont été sauvegardées toutes les picosecondes. Les structures secondaires ont été assignées avec le logiciel DSSP version CMBI 2000 (*Define Secondary Structure of Proteins*) [166]. DSSP assigne 8 états distincts dont 3 associés aux hélices ('H' pour les hélices α, 'I' pour les hélices π et 'G' pour les hélices 3_{10}), deux états associés aux coudes β ('T' pour les coudes β ayant une liaison hydrogène et 'S' pour *bent*, des coudes sans liaisons stabilisatrices), 'E' pour le brin β, 'B' pour le pont β (en anglais *ß-bridge*) et les boucles ('C' ou *coil*).

Un calcul du RMSF (*Root-Mean-Square Fluctuation*) sur les carbones α a été aussi effectué sur chaque trajectoire. La valeur du RMSF a été normalisée. Elle permet la quantification de la mobilité moyenne d'un résidu par rapport à sa position initiale observée au cours de la dynamique. Les facteurs B normalisés ont été récupérés sur chaque structure initiale, quantifiant ainsi la mobilité observée expérimentalement due aux vibrations atomiques ou au désordre statique du cristal. L'assignation d'un alphabet structural de 16 lettres (*a* à *p*) appelé Blocs Protéiques (BP, voir Annexe 3), va décrire la conformation locale spécifique de chaque résidu [167]. Chaque BP est caractérisé par les valeurs d'angles ϕ et ψ de 5 résidus consécutifs centré sur son résidu central. Le calcul du N_{eq}, mesure entropique du nombre moyen de BP qu'un résidu adopte dans le temps, permet de quantifier sa déformation par la formule :

$$Neq = \exp(-\sum_{i=1}^{16} f_i \, lnf_i)$$

avec f_i la fréquence d'un BP *i* à une position spécifique. Une valeur de N_{eq} proche de 1 indique que le résidu ne change donc pas de conformation et donc est très rigide.

Chaque résidu est donc caractérisé par une assignation initiale d'une structure secondaire, ainsi qu'un vecteur de taille 8 décrivant chaque structure secondaire assigné par DSSP à chaque conformation au cours de la dynamique. Afin de caractériser les changements conformationnels dans le temps, un regroupement par la méthode des *k-moyens*, *k-means* en anglais, a été réalisé l'ensemble des vecteurs (*k=5,* après quelques tests).

C. Analyse du comportement dynamique des hélices

L'analyse des distributions des facteurs B normalisés (Figure 97A) et RMSF normalisés (Figure 97B) montrent à première vue une très forte similarité entre les deux distributions, indiquant des résidus rigides et peu mobiles. Néanmoins, la modeste corrélation mesurée (coefficient de Pearson avec un *r = 0,43*) indique une certaine disparité entre flexibilité et mobilité. Des régions hélicoïdales présentent des valeurs de RMSF et facteurs B proches de zéro, confirmant le caractère rigide de ces structures secondaires, les autres zones sont plus complexes. De forts RMSF normalisés avec des faibles facteurs B normalisés peuvent correspondre à des zones contraintes dans le cristal, tandis que le contraire peut provenir de limitations computationnelles ou du milieu expérimental qui a entrainé un certain désordre.

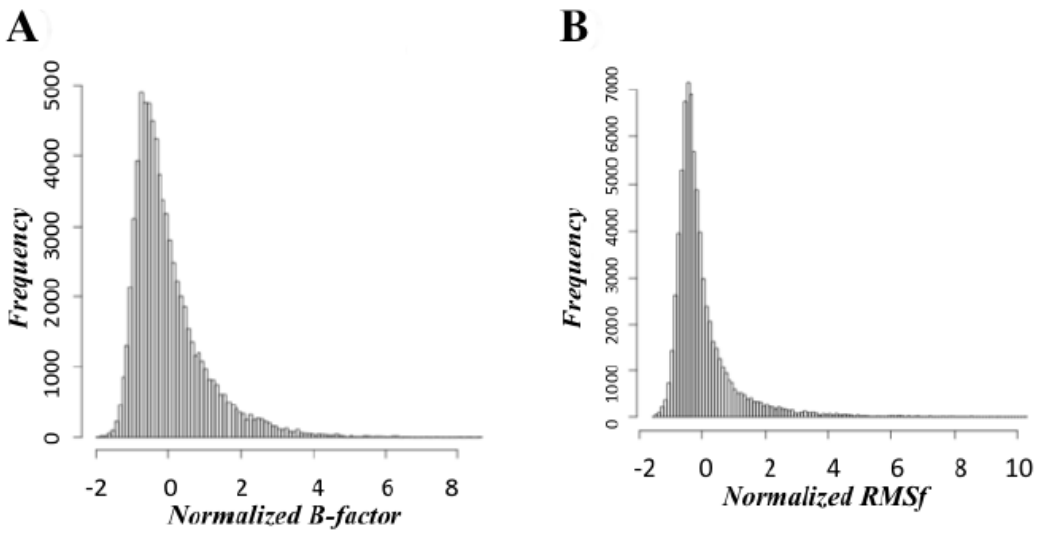


Figure 97 A. Distribution des facteurs B normalisés sur l'ensemble des résidus présents dans le jeu de données. B. Distribution des RMSF normalisés calculés sur les 507 dynamiques.

L'analyse statistique des conformations adoptées par les résidus assignés initialement en hélice α (code DSSP 'H') pendant la dynamique confirme le caractère rigide et stable de ces hélices. 91,4% de ces hélices α reste hélice α dans plus de 50% du temps alors que seul 3,9% reste moins de 25% du temps de la dynamique (Figure 98A). Ces résultats sont confirmés par le regroupement par *k-moyens* regroupement où le groupe α_1 confirme ces observations (76,4% des résidus) (Figure 98B). Cependant, des états transitoires vers une conformation coude *β* (*β-turn*s) sont modérément observées notamment dans les clusters α_{T1}, α_2 et α_{T2} (22,4% des résidus). Ces changements conformationnels sont reflétés par des valeurs supérieures de RMSF et N_{eq}.

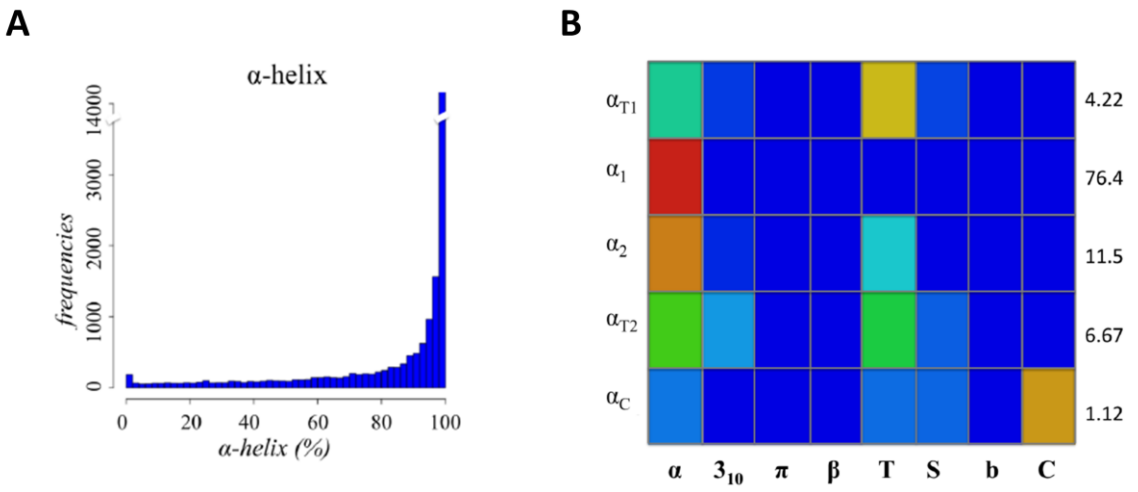


Figure 98 Comportement dynamique des hélices α. A. Distribution de la proportion d'assignation hélice α observée pour l'ensemble des résidus B. Carte (Heatmap) des différents profils d'assignation de structures secondaires pour les hélices α obtenus après regroupement. Sont notés les 5 groupes obtenus avec sur l'abscisse, les 8 classes assignées par DSSP, hélices α, 3_{10} et π, puis feuillet β, coudes β (T et S), pont β (b) et les boucles (C). La couleur va du bleu (0%) au rouge (100%), à droite est notée l'occurrence, en pourcent, du groupe proprement dit.

L'analyse des hélices 3_{10} indique une conservation modérée au cours du temps, seuls 40,5% de ces résidus restent réellement associés à une conformation locale 3_{10} (distribution sur la Figure 99A et groupe 3_{10} sur la Figure 99B). La distribution de la Figure 99A est très différente de celle de la Figure 98A pour l'hélice α. La transition vers un état *β-turn*, mise en évidence par les groupes 3_{10}^{T1} et 3_{10}^{T2}, est tout aussi fréquente avec 25,0% et 17,5% des résidus impliqués. Ce comportement peut s'expliquer par la forte similarité observée au niveau des angles ϕ et ψ, mise en avant notamment par Richardson [5], entre les hélices 3_{10} et les *β-turns* de type III et III' (éliminés de la catégorie des coudes du fait de cette ambiguïté). Néanmoins, ces deux groupes sont fréquemment associés à des valeurs de RMSF faibles. Le groupe 3_{10}^{C} est sujet à de nombreux changements structuraux confirmés par une forte déformabilité (N_{eq} élevé) et une forte mobilité (fort RMSF). De manière plus surprenante, seulement 10,5% des résidus adoptent préférentiellement la conformation d'une hélice α.

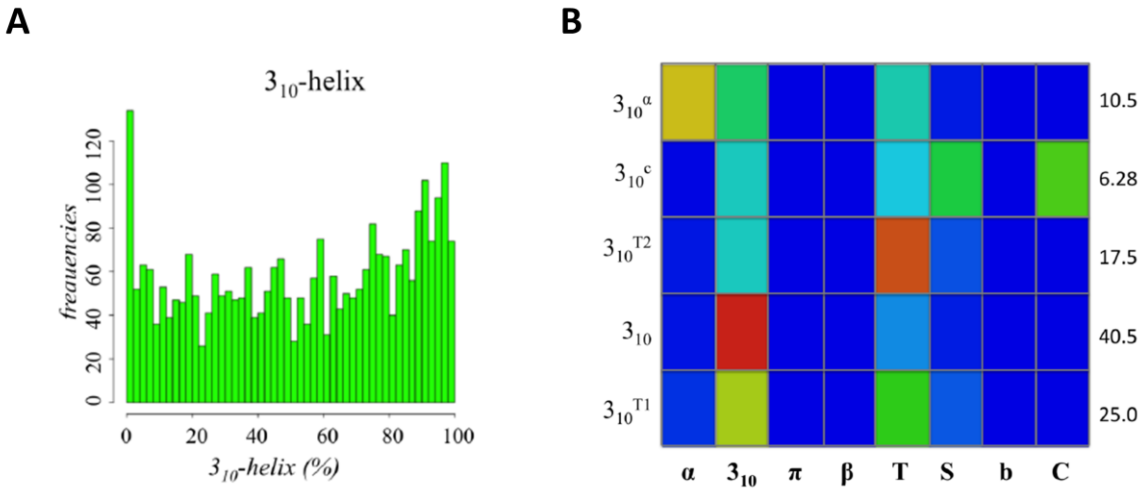


Figure 99 Comportement dynamique des hélices 3_{10}. A. Distribution de la proportion d'assignation hélice 3_{10} observé pour l'ensemble des résidus B. Carte (Heatmap) des différents profils d'assignation de structures secondaires pour les hélices 3_{10} obtenus après regroupement. La couleur va du bleu (0%) au rouge (100%).

Contrairement aux autres types d'hélices, les hélices π sont sensibles à la méthode d'assignation par DSSP. En effet, une mise à jour récente du logiciel modifie l'ordre d'assignation des hélices. La version la plus courante, CMBI de 2000, caractérisait dans un premier temps les hélices α, puis les hélices 3_{10} puis enfin les hélices π défavorisant la présence de ces dernières. Cet ordre est inversé dans la version 2.2.1 [168] favorisant les conformations les plus rares, aboutissant à une assignation des acides aminés 15 fois plus importante sur ce type d'hélice, de 0,02% à 0,32%. Le regroupement par k-moyens a été effectué sur les deux versions, représenté sur la Figure 100.

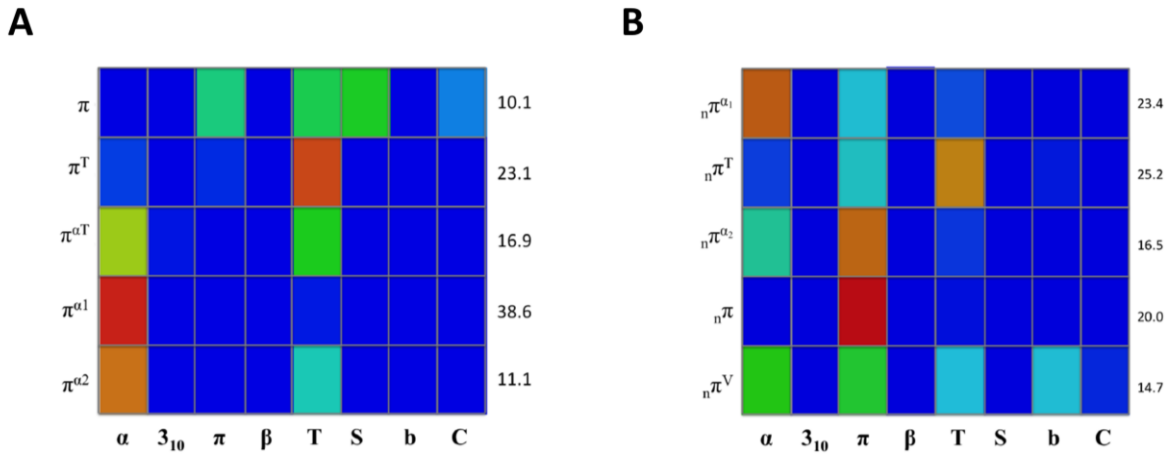


Figure 100 Cartes de densité des profils de transitions conformationnels des hélices π obtenus par la méthode de regroupement par k-moyens. A. DSSP version CMBI 2000 B. DSSP version 2015.

Les résultats obtenus par la version 2000 indiquent un comportement très instable des hélices π puisque seuls 10,1% des résidus restent ponctuellement assignés au même état (Figure 100A, groupe π). Cet unique groupe π est en plus très hétérogène puisque des assignations *turns*, *bends* et *coils* sont aussi relevées. La transition vers un état *β-turn* et hélice α est très fréquemment observée, 38,6% seulement vers l'hélice α (groupe $\pi^{\alpha 1}$), 23,1% des acides aminés transitent vers le *β-turn* (groupe π^{T}) et 28,0% vers ces deux structures secondaires (groupes $\pi^{\alpha T}$ et $\pi^{\alpha 2}$).

La version mise à jour de DSSP suggère un comportement plus stable que précédemment observé (Figure 100B). Bien que 39,9% des hélices π transitent vers une hélice α (groupes ${}_{n}\pi^{\alpha 1}$ et ${}_{n}\pi^{\alpha 2}$), 20% d'entre elles maintiennent un état hélicoïdal π (groupe ${}_{n}\pi$). Un autre groupe majeur émerge regroupant les hélices adoptant une conformation *β-turn* impliquant 25,5% des acides aminés (groupe ${}_{n}\pi^{T}$). Enfin, 14,7% des hélices π sont beaucoup plus variables et adoptent un nombre important de structures secondaires tels qu'un *β-turn* ou une hélice α (groupe ${}_{n}\pi^{V}$).

D. Conclusion

A travers cette étude, nous avons mis en évidence l'aspect dynamique du changement structural des hélices π et 3_{10} ainsi que leur préférences conformationnelles résumées sur la Figure 101. 60% des résidus assignés en tant qu'hélice 3_{10} alternent entre leur conformation hélicoïdale et une conformation *β-turn* lors de la dynamique. Cependant, 10% d'entre eux vont en plus s'assimiler à des hélices α, généralement associés à de fortes valeurs de flexibilité (voir Figure 101). Les hélices considérées comme α dans une structure PDB conserveront la même forme dans plus de 90% du temps. Néanmoins, les hélices α pouvant changer de conformation vers un *β-turn* sont caractérisées par des valeurs de RMSF et de facteurs B normalisées supérieures par rapport aux hélices conservées. L'interprétation des résultats pour les hélices π dépend de la méthode d'assignation de structure secondaire employée. Ainsi, l'ancienne version décrit des hélices π caractérisées comme instables puisque seuls 10% de ces résidus restent dans le même état hélicoïdal, et ce, temporairement. La nouvelle version, favorisant les conformations sous-représentées, donnent un profil de transition structurale proche de celui observé pour les hélices 3_{10}.

Des analyses sont en cours pour l'analyse de la dynamique des coudes et des feuillets pour compléter cette étude, ainsi qu'au travers des Blocs Protéiques.

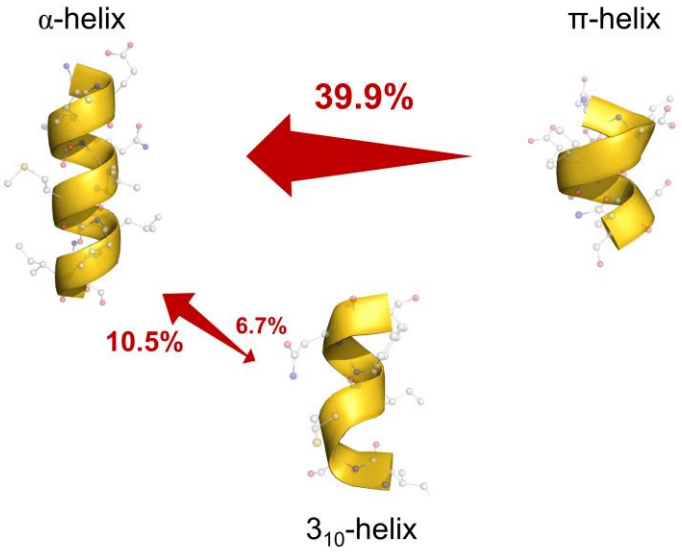


Figure 101 Résumé des changements conformationnels adoptés par les différents motifs hélicoïdaux.

Ces résultats, en plus de faire l'objet d'un article scientifique, ont aussi été exposés lors d'une présentation orale effectuée lors du congrès GGMM 2017 (Groupe de Graphisme et Modélisation Moléculaire) à Reims.

Chapitre 2 : Un point sur les PolyProline de type II

Une contribution avec une revue portant sur les hélices PolyProlines de type II (PPII) en tant que 3ème auteur a été effectuée [169]. Ces structures secondaires, plus fréquentes qu'on ne le pense souvent (environ 5% des résidus), présentent un *pattern* répété d'angles ψ et ϕ proches de 150° et -75° respectivement, les confondant souvent avec les angles des brins *β*. De plus, elles ne disposent pas de réseau de liaisons hydrogènes par leur constitution exclusive de prolines. La distinction entre hélices PolyProlines se fait notamment par leur provenance, on retrouve les hélices de type I dans des milieux riches en alcool tandis que les hélices de type II peuvent être détectées dans des structures protéiques. Toutefois, seul un nombre limité (3) de logiciel d'assignation de structures secondaires caractérise les PPII, notre laboratoire ayant développé une extension dédiée de DSSP, nommée DSSP-PPII, en 2011 pour les caractériser [170]. Cette absence de liaison hydrogène les rendent très accessibles au solvant favorisant

leur rôle dans les interactions protéine-protéine (domaine SH3) et protéine-ADN (facteurs de transcription). Enfin, ces structures secondaires ont récemment été explorées dans le développement de peptides permettant l'introduction de molécules dans la cellule [171].

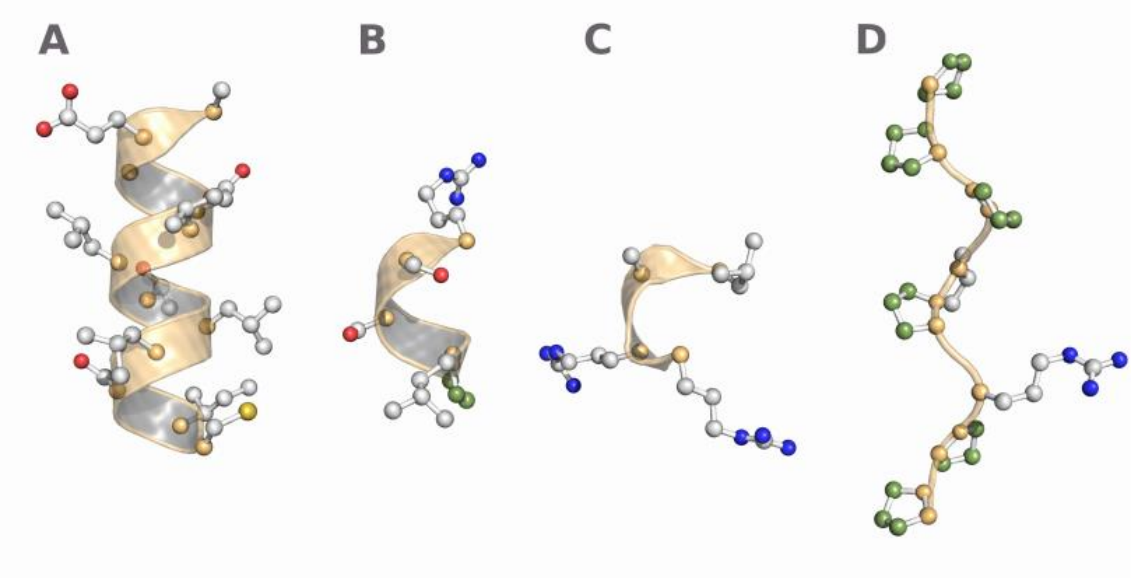


Figure 102 Représentations 3D des différentes structures secondaires hélicoïdales. A. Hélice α, B. hélice 310, C. hélice π, et D. hélice PolyProline type II.

Chapitre 3 : Flexibilité protéique et alphabet structural

A travers ces travaux de thèse, une participation à la revue de Craveur et collaborateurs [172] a pu être réalisé en tant que 6ème auteur. Ces travaux exhaustifs explorent les relations sous-jacentes entre les différentes métriques et notions caractérisant le concept de flexibilité protéique. De manière générale, le terme flexibilité est utilisé dans le domaine de la dynamique moléculaire pour désigner un élément mobile. Or cette notion de mobilité peut en réalité être caractérisée par plusieurs types de mouvements distincts tels qu'un mouvement translationnel ou une déformation. Cette distinction fut explorée à travers cette étude exhaustive. L'utilisation de dynamique moléculaire couplée au calcul du RMSF permettent la quantification d'un mouvement translationnel d'un résidu ou atome dans un environnement 3D, ou *mobilité*. L'étude des facteurs B normalisés indique une mesure expérimentale de la mobilité de chaque atome. Enfin, la déformation locale d'une structure a

été mesurée par l'assignation des BPs pendant la dynamique et quantifiée par la mesure du N_{eq}. Un exemple concret, représenté sur la Figure 103, sur une boucle du domaine *Calf-1* des intégrines souligne cette distinction. Bien que cette boucle soit considérée comme flexible, elle est en réalité très mobile (RMSF élevé) et rigides (N_{eq} faible).

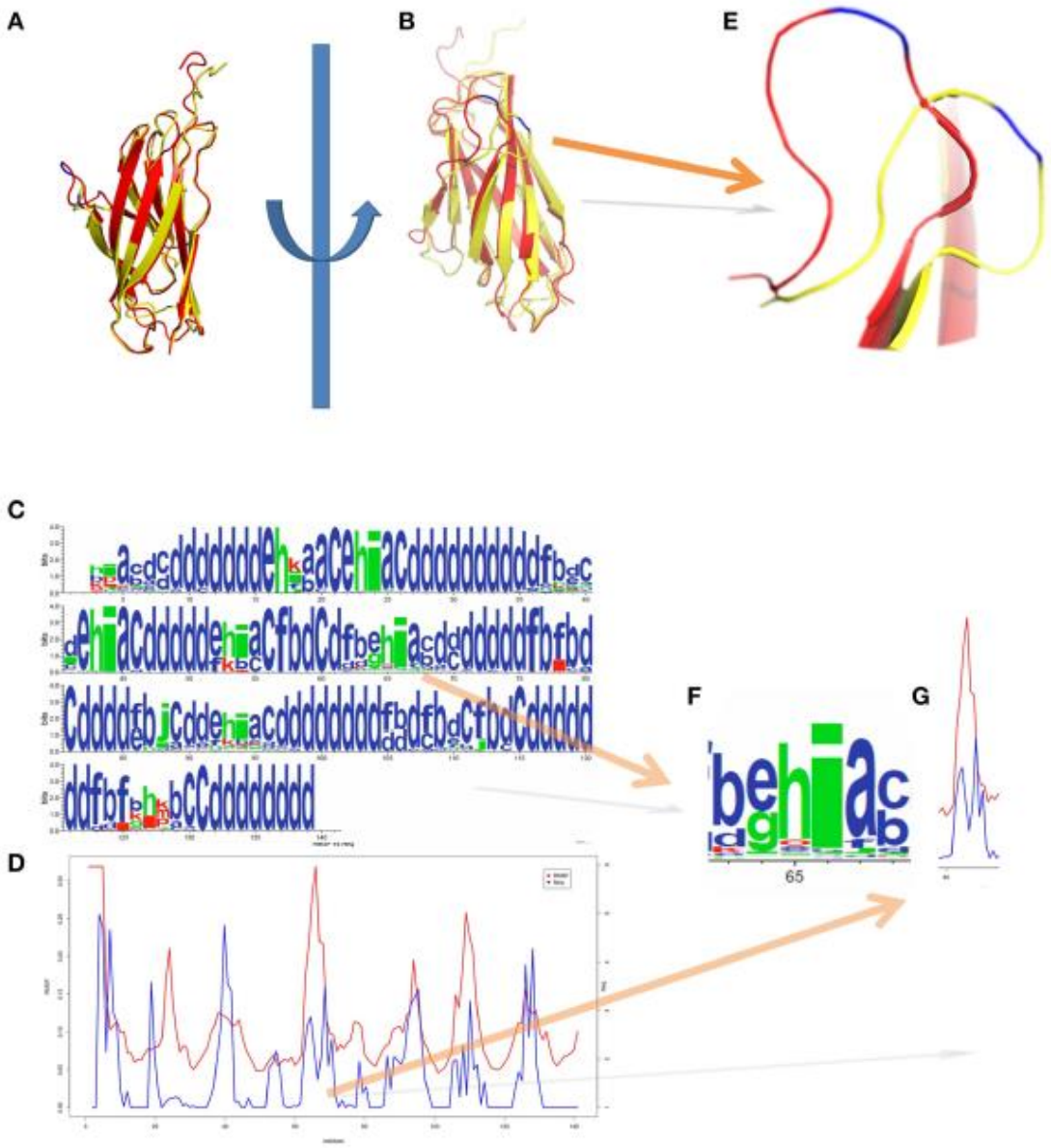


Figure 103 Comportement dynamique de la structure du domaine Calf-1. A et B. Superpositions 3D de la structure initiale et une représentation obtenue lors de la simulation. C. Distribution des PBs au cours du temps par la représentation WebLogo [173] D. Superposition des valeurs de Neq et RMSF E. Zoom sur une boucle du domaine présentant F. une faible déformation par les BPs et G. une forte mobilité caractérisée par un fort RMSF.

Chapitre 4 : Présentations sous forme de poster

En complément de ces travaux menant à des publications, 3 posters ont été présentés lors de cette thèse, ils portent principalement sur les interactions.

Le premier poster, présenté le 7 Novembre 2016 à la *German Cheminformatics Conference* (GCC) 2016 à Fulda, Allemagne exposait des travaux comparatifs en matière d'interactions à partir des données et des résultats fournis par Lenselink et collaborateurs [77]].

Le deuxième poster fut présenté à la *German Cheminformatics Conference* (GCC) 2017 à Mayence, Allemagne. Ce poster illustre l'analyse dans la base de données des contacts interatomiques observés dans 3decision® et la mise en évidence d'interactions peu caractérisés dans les approches de détection classiques tels *PLIP* [64].

Enfin, le dernier poster fut exposé lors de la 11ème *International Conference of Chemical Structures*, du 27 au 31 Mai 2018 à Noordwijkerhout, Pays-Bas. Ce poster met en avant les travaux effectués sur l'analyse fine des halogènes.

Conclusion générale

Le fonctionnement d'une protéine dans un processus biologique classique repose essentiellement sur des interactions réalisées avec d'autres molécules tels que l'ADN, l'ARN, d'autres protéines, substrats ou encore de petites molécules comme certains médicaments. Un des enjeux majeurs du domaine pharmaceutique réside dans la compréhension de ces interactions. Ainsi, l'appréhension des mécanismes régissant l'affinité intermoléculaire permet le développement de substituts, médicaments, agissant notamment sur le fonctionnement de la voie métabolique. Le projet 3decision® est né avec pour objectif d'assister les travaux de recherches sur le développement médicamenteux en optimisant l'utilisation des données de structures. Le cadre de cette thèse s'inscrit dans cette optique d'optimisation par l'étude des interactions effectuées entre une cible thérapeutique, protéine, et de molécules de petite taille composées de moins de 100 atomes.

Cette liaison du ligand sur sa protéine dépend notamment des interactions moléculaires dites non covalentes qui vont contribuer à l'affinité de la molécule pour son site de liaison. Ces interactions, considérées comme faibles en termes d'énergie, sont généralement regroupées en fonction de la nature électrostatique et de l'arrangement spatial des atomes impliqués (liaison hydrogène etc.). Bien que ces descriptions facilitent la compréhension du mécanisme de liaison protéine - ligand, leurs applications peuvent être limitées notamment dans le cadre de la comparaison de complexes [75, 77]. De plus, ces définitions limitent la détection à des interactions déjà connues, restreignant la mise en évidence de certains *patterns* récurrents mais peu voire non décrits.

Ainsi, le logiciel de détection d'interactions *PLIP* [64], dont la définition des interactions repose sur ces descriptions, affichait dans un certain cas un nombre relativement faible d'interactions, peu pertinent pour expliquer la liaison du ligand sur son récepteur.

Pour pallier ces limitations, mon travail a alors consisté à implémenter une solution informatique dans 3decision® permettant l'identification rapide des contacts présents entre une protéine et son ligand. Deux challenges furent rencontrés : l'utilisation du *framework* Pipeline Pilot ne disposant que de peu d'outils adaptés au traitement des structures 3D ainsi que le fonctionnement de ce programme pour de grands jeu de données venant du secteur

public et industriel. Ce programme a permis d'établir la liste des contacts dans un critère de distance prédéfini sur la totalité de la base de données. Cet outil est à l'heure actuelle utilisé à la fois dans l'enregistrement de nouvelles structures mais aussi l'analyse de poses de *docking*.

L'enregistrement de l'ensemble des contacts protéine - ligand dans une base de données a permis d'explorer de nouvelles descriptions de l'interaction moléculaire. Cette approche a été appliquée aux halogènes ici. Jusqu'à présent, l'implication des halogènes dans les petites molécules étaient essentiellement décrits par les liaisons halogènes et très récemment par leur capacité à former des liaisons hydrogènes. Or, moins de 30% de ces halogènes sont impliqués dans ces deux types d'interactions. Ces résultats ont amené une étude plus approfondie afin de caractériser les 70% restant de manière plus détaillée. Une analyse de l'environnement des interactions autour de ces atomes a pu être réalisée et révèle une description plus complète des interactions dans le contexte protéine - ligand. La présence de liaison halogène est par exemple très prépondérante des inhibiteurs de complexes impliquant des facteurs anti-coagulants tel le facteur X. D'autres récepteurs protéiques mettent en évidence la présence d'interactions potentiellement défavorables dans leur environnement et souligne la complexité de l'analyse des interactions dans ce contexte. Ces interprétations sont toutefois à relativiser par rapport au faible nombre d'atomes de brome et d'iode présent dans la PDB. De plus, la redondance des données de structures n'a été considérée que partiellement à travers cette étude. Des critères simples tels que la similarité de séquence et moléculaire ont été utilisés laissant un certain biais dans nos résultats.

Dans la continuité de cette démarche, une étude sur la redondance, connue mais peu explorée, des modes de liaisons des complexes issus de la PDB a été réalisé. La finalité de cette approche est double : établir un jeu de données non redondant en termes de structure et explorer la diversité des modes de liaisons de certains complexes. Notre base de données de contacts a permis d'identifier les résidus impliqués dans le mécanisme de liaison. Dès lors, un alignement structural des complexes impliquant des protéines et ligands identiques sur ces résidus a mis en évidence une redondance structurale importante dans la PDB. Les résultats indiquent que seule une proportion modérée de complexes, 4 555, présentent plusieurs modes de liaisons distincts, pouvant potentiellement être pris en compte dans l'analyse de résultats de *docking* par exemple. A contrario, la majorité des complexes comportent un mode

de liaison unique, jusqu'à 501 conformations ont pu être regroupées en un représentant. La quantification de cette redondance à l'échelle de la PDB a ainsi pu être analysé pour la premières fois. Ces résultats s'inscrivent dans plusieurs optiques. Tout d'abord, la réduction des données peut faciliter le travail d'analyse à grande échelle. Puis, la diversité des modes de liaisons sur certains complexes peut être révélateur de leur spécificité.

Dans le cadre des recherches du laboratoire DSIMB, j'ai pu aussi étudier la dynamique et la flexibilité des protéines avec une étude complémentaire sur le comportement dynamique des différentes structures hélicoïdales a pu être réalisée durant cette thèse. Les protéines, adoptant un ensemble conformationnel dans le temps, ce travail a permis de mettre en évidence la conservation de l'hélice la plus fréquente, l'hélice α. En parallèle, les hélices 3_{10} et π adoptent un comportement plus transitionnel, pouvant adopter des conformations locales proches du *β-turn* ou d'une hélice α. L'étude des structures secondaires nécessitent l'utilisation d'algorithme d'assignation de ces mêmes structures secondaires, or les différentes implémentations aboutissent à des interprétations différentes. Ainsi, la version datant de 2000 de DSSP sous-identifie les hélices π et leur attribue un comportement instable. A contrario, la version plus récente, 2.2.1, privilégie l'assignation des motifs hélicoïdaux plus rares qui se traduit par un profil dynamique proche de l'hélice 3_{10}. Toutefois, il est important de tenir compte de l'impact des champs de force sur la conservation des structures secondaires. Amber étant réputé pour stabiliser fortement les conformations hélicoïdales [174], il sera intéressant de reproduire ces résultats avec l'utilisation d'autres champs de force tels que CHARMM [175] ou GROMOS [165] par exemple. L'approche utilisée, couplant alphabet structural et regroupement par k-moyens, peut être utilisé à l'avenir pour l'étude conformationnelle d'autres structures secondaires, une étude au sein du laboratoire étant d'ailleurs en cours de publication.

En plus de ces travaux d'analyse destinés à la recherche publique, des outils portant sur les contacts à des fins applicatives ont été incorporés dans 3decision® dans l'optique de guider et faciliter le développement médicamenteux. Ainsi, la superposition des contacts sur un atome d'une protéine ou d'un ligand à partir de l'ensemble des données observées dans la base s'inscrit dans cette démarche. Cet alignement de contacts permet de créer des grilles de densité, où les éléments les plus fréquents représentent de potentielles suggestions

d'addition ou substitution moléculaires. Le module, disponible prochainement, montre déjà des résultats prometteurs illustrés à travers cette thèse.

En complément de ces travaux de superposition, diverses approches de comparaison d'interactions entre complexes ont été explorées en vue d'une intégration à 3decision®. De nouvelles descriptions d'interactions furent explorées telles qu'une approche centrée sur les fragments. Cette dernière ne permit pas l'obtention de résultats concluants dans un contexte de *Virtual Screening*. Toutefois, le développement d'une approche centrée sur les contacts interatomiques sans considération du type d'interaction montre des résultats concluants et pourra être implémentée à l'avenir après optimisation des performances de comparaison.

A travers ce travail de thèse, des études de recherche ainsi que applications concrètes portant sur les interactions moléculaires ont pu être intégrées à 3decision®. 3decision® est une solution en constante évolution, de nombreuses améliorations peuvent être considérées et implémentées à court et moyen terme. La définition des interactions, bien que relativement complète peut faire l'objet d'améliorations plus poussées. La prise en compte par exemple des interactions intramoléculaires telles que celles présentes dans les structures secondaires est un élément pouvant affiner la caractérisation en contrepartie d'un processus de détection plus lent. Des interactions jusque-là peu explorées dans le contexte protéine - ligand de la PDB telles que l'interaction oxygène - sulfure aromatique mis en avant par Zhang [58] sont donc intéressante à être considérées. Ces travaux soulignent l'importance de considérer les fragments environnant les partenaires de l'interaction analysé. Ils montrent de plus les limites d'une analyse statistique des interactions où la description du contact resterai exclusivement binaire.

La description des interactions à travers ces travaux de thèse s'inscrit dans le contexte protéine - ligand de la PDB. Dès lors, l'identification de ces interactions et contacts dépend des données à notre disposition. L'ensemble du protéome et des représentations chimiques ne sont pas présentes dans la PDB, d'où la nécessité d'explorer de nouvelles sources de données à l'avenir. Ainsi, les interactions protéine-protéine mais aussi les structures provenant de la CCDC (*Cambridge Crystallographic Data Centre*) devront être examinées afin de mieux appréhender les interactions moléculaires.

De même, un accent fort a été mis sur l'étude des interactions impliquant la protéine, cependant il ne faut pas négliger les interactions effectuées avec d'autres ligands ou

molécules d'eau qui sont tout aussi importantes dans le mécanisme de liaison. La notion de molécules d'eau dites conservées dans le site de liaison est récurrent dans la littérature [43].

Enfin, la comparaison des *patterns* d'interactions entre deux complexes peut faire l'objet d'expérimentations additionnelles. Une approche quantitative décrivant l'environnement favorable mais aussi défavorable d'un atome du ligand, comme décrit sur les halogènes peut être envisagée. Une collaboration entre Discngine et un laboratoire de recherche public va d'ailleurs dans ce sens. En outre, il sera intéressant d'explorer la relation entre conservation du mode de liaison d'un ligand et différences observées dans les *patterns* d'interactions et de contacts à travers nos résultats de redondance de complexes.

Un autre point d'amélioration dans la comparaison de contacts reste une approche dite topologique. Une métrique quantitative de similarité laisse peu de place à l'interprétation lors d'une comparaison et ne permet pas l'alignement des structures selon leur *pattern* de contacts en commun. Rares sont les méthodes innovantes, telles que *Grim* proposée par Desaphy et collaborateurs [70], proposant des approches topologiques de comparaison. L'utilisation de cette méthode est restreinte à un nombre limité d'interactions typées par atomes. La considération des contacts s'affranchit de telles définitions et considère potentiellement plusieurs fois un contact de même nature par atome du ligand par exemple. Ainsi, le nombre important de contacts recensés impacterait la détection de clique maximale. De nombreuses expérimentations ont été réalisés dans ce sens au sein de Discngine et feront l'objet d'études supplémentaires. L'utilisation d'un algorithme d'optimisation combinatoire similaire à l'algorithme dit Hongrois est envisagée [176].

Les avancées récentes dans la compréhension des interactions moléculaires permettent de mieux appréhender la conception médicamenteuse. Ces travaux vont dans ce sens et permettront de guider, par l'intermédiaire de l'utilisation du logiciel 3decision®, l'élaboration de ces molécules. En effet, de nombreuses recherches dans le domaine restent encore à effectuer pour améliorer au maximum l'efficacité de ces composés et de réduire les risques d'effets non désirés.

Bibliographie

[1] G. Faure, A. Bornot et A. G. Brevern, «Protein contacts, inter-residue interactions and side-chain modelling,» *Biochimie,* vol. 90, pp. 626-639, 4 2008.

[2] M. I. Sadowski et D. T. Jones, «The sequence–structure relationship and protein function prediction,» *Current Opinion in Structural Biology,* vol. 19, pp. 357-362, 6 2009.

[3] L. Pauling, R. B. Corey et H. R. Branson, «The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.,» *Proceedings of the National Academy of Sciences of the United States of America,* vol. 37, n° %14, pp. 205-211, 4 1951.

[4] G. N. Ramachandran, C. Ramakrishnan et V. Sasiekharan, «Stereochemistry of polypeptide chain configurations.,» *Journal of molecular biology,* vol. 7, pp. 95-99, 7 1963.

[5] J. S. Richardson, «The Anatomy and Taxonomy of Protein Structure,» chez *Advances in Protein Chemistry Volume 34*, Elsevier, 1981, pp. 167-339.

[6] S. A. Hollingsworth et P. A. Karplus, «A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins,» *BioMolecular Concepts,* vol. 1, 1 2010.

[7] R. S. Vieira-Pires et J. H. Morais-Cabral, «310helices in channels and other membrane proteins,» *The Journal of General Physiology,* vol. 136, pp. 585-592, 11 2010.

[8] J. C. KENDREW, B. O. D. O. G, H. M. DINTZIS, R. G. PARRISH, H. WYCKOFF et D. C. PHILLIPS, «A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.,» *Nature,* vol. 181, n° %14610, pp. 662-666, 3 1958.

[9] M. R. Lee et P. A. Kollman, «Free-Energy Calculations Highlight Differences in Accuracy between X-Ray and NMR Structures and Add Value to Protein Structure Prediction,» *Structure,* vol. 9, pp. 905-916, 10 2001.

[10] M. P. Williamson, T. F. Havel et K. Wüthrich, «Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry.,» *Journal of molecular biology,* vol. 182, n° %12, pp. 295-315, 3 1985.

[11] Y. Huang, P. A. Winkler, W. Sun, W. Lü et J. Du, «Architecture of the TRPM2 channel and its activation mechanism by ADP-ribose and calcium,» *Nature,* vol. 562, pp. 145-149, 9 2018.

[12] A. Tokuhisa, S. Jonic, F. Tama et O. Miyashita, «Hybrid approach for structural modeling of biological systems from X-ray free electron laser diffraction patterns,» *Journal of Structural Biology,* vol. 194, pp. 325-336, 6 2016.

[13] H. Tsuruta et T. C. Irving, «Experimental approaches for solution X-ray scattering and fiber diffraction,» *Current Opinion in Structural Biology,* vol. 18, pp. 601-608, 10 2008.

[14] H. M. Berman, «The Protein Data Bank,» *Nucleic Acids Research,* vol. 28, pp. 235-242, 1 2000.

[15] «Crystallography: Protein Data Bank,» *Nature New Biology,* vol. 233, pp. 223-223, 10 1971.

[16] «UniProt: the universal protein knowledgebase,» *Nucleic Acids Research,* vol. 45, pp. D158--D169, 11 2016.

[17] A. G. Murzin, S. E. Brenner, T. Hubbard et C. Chothia, «SCOP: A structural classification of proteins database for the investigation of sequences and structures,» *Journal of Molecular Biology,* vol. 247, pp. 536-540, 4 1995.

[18] A. Wlodawer, W. Minor, Z. Dauter et M. Jaskolski, «Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures,» *FEBS Journal,* vol. 275, pp. 1-21, 11 2007.

[19] R. A. Laskowski, M. W. MacArthur, D. S. Moss et J. M. Thornton, «PROCHECK: a program to check the stereochemical quality of protein structures,» *Journal of Applied Crystallography,* vol. 26, pp. 283-291, 4 1993.

[20] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson et D. C. Richardson, «MolProbity: all-atom contacts and structure validation for proteins and nucleic acids,» *Nucleic Acids Research,* vol. 35, pp. W375--W383, 5 2007.

[21] R. J. Read, P. D. Adams, W. B. Arendall, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend et P. H. Zwart, «A New Generation of Crystallographic Validation Tools for the Protein Data Bank,» *Structure,* vol. 19, pp. 1395-1412, 10 2011.

[22] C. Bissantz, B. Kuhn et M. Stahl, «A Medicinal Chemist's Guide to Molecular Interactions,» *Journal of Medicinal Chemistry,* vol. 53, pp. 5061-5084, 7 2010.

[23] R. F. Freitas et M. Schapira, «A systematic analysis of atomic protein–ligand interactions in the PDB,» *MedChemComm,* vol. 8, pp. 1970-1981, 2017.

[24] L. Pérez-Benito, H. Keränen, H. Vlijmen et G. Tresadern, «Predicting Binding Free Energies of PDE2 Inhibitors. The Difficulties of Protein Conformation,» *Scientific Reports,* vol. 8, 3 2018.

[25] F. Gräter, S. M. Schwarzl, A. Dejaegere, S. Fischer et J. C. Smith, «Protein/Ligand Binding Free Energies Calculated with Quantum Mechanics/Molecular Mechanics,» *The Journal of Physical Chemistry B,* vol. 109, pp. 10474-10483, 5 2005.

[26] J. Michel et J. W. Essex, «Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations,» *Journal of Computer-Aided Molecular Design,* vol. 24, pp. 639-658, 5 2010.

[27] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini et R. P. Mee, *Journal of Computer-Aided Molecular Design,* vol. 11, pp. 425-445, 1997.

[28] C. L. Perrin et J. B. Nielson, «STRONG HYDROGEN BONDS IN CHEMISTRY AND BIOLOGY,» *Annual Review of Physical Chemistry,* vol. 48, pp. 511-544, 10 1997.

[29] M. W. Feyereisen, D. Feller et D. A. Dixon, «Hydrogen Bond Energy of the Water Dimer,» *The Journal of Physical Chemistry,* vol. 100, pp. 2993-2997, 1 1996.

[30] M. J. Minch, «An Introduction to Hydrogen Bonding (Jeffrey, George A.),» *Journal of Chemical Education,* vol. 76, p. 759, 6 1999.

[31] X.-Z. Li, B. Walker et A. Michaelides, «Quantum nature of the hydrogen bond,» *Proceedings of the National Academy of Sciences,* vol. 108, pp. 6369-6373, 4 2011.

[32] M. Ceriotti, J. Cuny, M. Parrinello et D. E. Manolopoulos, «Nuclear quantum effects and hydrogen bond fluctuations in water,» *Proceedings of the National Academy of Sciences,* vol. 110, pp. 15591-15596, 9 2013.

[33] J. Perlstein, «The Weak Hydrogen Bond In Structural Chemistry and Biology (International Union of Crystallography, Monographs on Crystallography, 9) By Gautam R. Desiraju (University of Hyderabad) and Thomas Steiner (Freie Universität Berlin). Oxford University Press: Oxford and New York. 1999. xiv 507 pp. 150. ISBN 0-19-850252-4.,» *Journal of the American Chemical Society,* vol. 123, pp. 191-192, 1 2001.

[34] E. Nittinger, T. Inhester, S. Bietz, A. Meyder, K. T. Schomburg, G. Lange, R. Klein et M. Rarey, «Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein–Ligand Interfaces,» *Journal of Medicinal Chemistry,* vol. 60, pp. 4245-4257, 5 2017.

[35] E. H_ckel, «Quantentheoretische Beitr_ge zum Benzolproblem,» *Zeitschrift f_r Physik,* vol. 70, pp. 204-286, 3 1931.

[36] J. S. Murray, P. Lane, T. Clark, K. E. Riley et P. Politzer, «-Holes, -holes and electrostatically-driven interactions,» *Journal of Molecular Modeling,* vol. 18, pp. 541-548, 5 2011.

[37] A. Bauzá, T. J. Mooibroek et A. Frontera, «Towards design strategies for anion–interactions in crystal engineering,» *CrystEngComm,* vol. 18, pp. 10-23, 2016.

[38] M. H. Kolář et P. Hobza, «Computer Modeling of Halogen Bonds and Other -Hole Interactions,» *Chemical Reviews,* vol. 116, pp. 5155-5187, 2 2016.

[39] O. Hassel, J. Hvoslef, E. H. Vihovde et N. A. Sörensen, «The Structure of Bromine 1,4-Dioxanate.,» *Acta Chemica Scandinavica,* vol. 8, pp. 873-873, 1954.

[40] Y. Levy et J. N. Onuchic, «Water and proteins: A love-hate relationship,» *Proceedings of the National Academy of Sciences,* vol. 101, pp. 3325-3326, 3 2004.

[41] J. E. Ladbury, «Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design,» *Chemistry & Biology,* vol. 3, pp. 973-980, 12 1996.

[42] S. Roehrig, A. Straub, J. Pohlmann, T. Lampe, J. Pernerstorfer, K.-H. Schlemmer, P. Reinemer et E. Perzborn, «Discovery of the Novel Antithrombotic Agent 5-Chloro-N-((5S)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-ylmethyl)thiophene-2-carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor,» *Journal of Medicinal Chemistry,* vol. 48, pp. 5900-5908, 9 2005.

[43] M. Jukič, J. Konc, S. Gobec et D. Janežič, «Identification of Conserved Water Sites in Protein Structures for Drug Design,» *Journal of Chemical Information and Modeling,* vol. 57, pp. 3094-3103, 12 2017.

[44] C. R. Martinez et B. L. Iverson, «Rethinking the term pi-stacking,» *Chemical Science,* vol. 3, p. 2191, 2012.

[45] J. Novotný, S. Bazzi, R. Marek et J. Kozelka, «Lone-pair– interactions: analysis of the physical origin and biological implications,» *Physical Chemistry Chemical Physics,* vol. 18, pp. 19472-19481, 2016.

[46] M. Egli et S. Sarkhel, «Lone Pair-Aromatic Interactions: To Stabilize or Not to Stabilize,» *Accounts of Chemical Research,* vol. 40, pp. 197-205, 3 2007.

[47] A. Jain, V. Ramanathan et R. Sankararamakrishnan, «Lone pair interactions between water oxygens and aromatic residues: Quantum chemical studies based on high-resolution protein structures and model compounds,» *Protein Science,* pp. NA--NA, 2009.

[48] P. G. Wang, High-Throughput Analysis in the Pharmaceutical Industry, CRC PR INC, 2008.

[49] R. Loewenthal, J. Sancho, T. Reinikainen et A. R. Fersht, «Long-Range Surface Charge-Charge Interactions in Proteins,» *Journal of Molecular Biology,* vol. 232, pp. 574-583, 7 1993.

[50] F.-Y. Lin et A. D. MacKerell, «Do Halogen–Hydrogen Bond Donor Interactions Dominate the Favorable Contribution of Halogens to Ligand–Protein Binding?,» *The Journal of Physical Chemistry B,* vol. 121, pp. 6813-6821, 7 2017.

[51] J. M. Ontoria, E. H. Rydberg, S. D. Marco, L. Tomei, B. Attenni, S. Malancona, J. I. M. Hernando, N. Gennari, U. Koch, F. Narjes, M. Rowley, V. Summa, S. S. Carroll, D. B. Olsen, R. D. Francesco, S. Altamura, G. Migliaccio et A. Carfi{\`}, «Identification and Biological Evaluation of a Series of 1H-Benzo[de]isoquinoline-1,3(2H)-diones as Hepatitis C Virus NS5B Polymerase Inhibitors,» *Journal of Medicinal Chemistry,* vol. 52, pp. 5217-5227, 8 2009.

[52] A. Rahim, P. Saha, K. K. Jha, N. Sukumar et B. K. Sarma, «Reciprocal carbonyl–carbonyl interactions in small molecules and proteins,» *Nature Communications,* vol. 8, 7 2017.

[53] R. Paulini, K. Müller et F. Diederich, «Orthogonal Multipolar Interactions in Structural Chemistry and Biology,» *Angewandte Chemie International Edition,* vol. 44, pp. 1788-1805, 2 2005.

[54] Y. N. Imai, Y. Inoue, I. Nakanishi et K. Kitaura, «Amide- interactions between formamide and benzene,» *Journal of Computational Chemistry,* pp. NA--NA, 2009.

[55] F. R. Fischer, P. A. Wood, F. H. Allen et F. Diederich, «Orthogonal dipolar interactions between amide carbonyl groups,» *Proceedings of the National Academy of Sciences,* vol. 105, pp. 17290-17294, 11 2008.

[56] B. R. Beno, K.-S. Yeung, M. D. Bartberger, L. D. Pennington et N. A. Meanwell, «A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design,» *Journal of Medicinal Chemistry,* vol. 58, pp. 4383-4438, 3 2015.

[57] W. B. Motherwell, R. B. Moreno, I. Pavlakos, J. R. T. Arendorf, T. Arif, G. J. Tizzard, S. J. Coles et A. E. Aliev, «Noncovalent Interactions of Systems with Sulfur: The Atomic Chameleon of Molecular Recognition,» *Angewandte Chemie,* vol. 130, pp. 1207-1212, 12 2017.

[58] X. Zhang, Z. Gong, J. Li et T. Lu, «Intermolecular SulfurOxygen Interactions: Theoretical and Statistical Investigations,» *Journal of Chemical Information and Modeling,* vol. 55, pp. 2138-2153, 10 2015.

[59] D. Rogers et M. Hahn, «Extended-Connectivity Fingerprints,» *Journal of Chemical Information and Modeling,* vol. 50, pp. 742-754, 5 2010.

[60] A. C. Wallace, R. A. Laskowski et J. M. Thornton, «LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions,» *"Protein Engineering, Design and Selection",* vol. 8, pp. 127-134, 1995.

[61] I. K. McDonald et J. M. Thornton, «Satisfying Hydrogen Bonding Potential in Proteins,» *Journal of Molecular Biology,* vol. 238, pp. 777-793, 5 1994.

[62] R. A. Laskowski et M. B. Swindells, «LigPlot: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery,» *Journal of Chemical Information and Modeling,* vol. 51, pp. 2778-2786, 10 2011.

[63] A. Schreyer et T. Blundell, «CREDO: A Protein-Ligand Interaction Database for Drug Discovery,» *Chemical Biology & Drug Design,* vol. 73, pp. 157-167, 2 2009.

[64] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme et M. Schroeder, «PLIP: fully automated protein–ligand interaction profiler,» *Nucleic Acids Research,* vol. 43, pp. W443--W447, 4 2015.

[65] N. M. OBoyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch et G. R. Hutchison, «Open Babel: An open chemical toolbox,» *Journal of Cheminformatics,* vol. 3, p. 33, 2011.

[66] H. C. Jubb, A. P. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. B. Ascher et T. L. Blundell, «Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures,» *Journal of Molecular Biology,* vol. 429, pp. 365-371, 2 2017.

[67] Z. Deng, C. Chuaqui et J. Singh, «Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions,» *Journal of Medicinal Chemistry,* vol. 47, pp. 337-344, 1 2004.

[68] B. Lee et F. M. Richards, «The interpretation of protein structures: Estimation of static accessibility,» *Journal of Molecular Biology,* vol. 55, pp. 379--IN4, 2 1971.

[69] T. T. Tanimoto, «Elementary mathematical theory of classification and prediction,» 1958.

[70] J. Desaphy, E. Raimbaud, P. Ducrot et D. Rognan, «Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs,» *Journal of Chemical Information and Modeling,* vol. 53, pp. 623-637, 3 2013.

[71] C. Bron et J. Kerbosch, «Algorithm 457: finding all cliques of an undirected graph,» *Communications of the ACM,* vol. 16, pp. 575-577, 9 1973.

[72] D. L. Theobald, «Rapid calculation of RMSDs using a quaternion-based characteristic polynomial,» *Acta Crystallographica Section A Foundations of Crystallography,* vol. 61, pp. 478-480, 6 2005.

[73] M. Weisel, H.-M. Bitter, F. Diederich, W. V. So et R. Kondru, «PROLIX: Rapid Mining of Protein–Ligand Interactions in Large Crystal Structure Databases,» *Journal of Chemical Information and Modeling,* vol. 52, pp. 1450-1461, 6 2012.

[74] C. Da et D. Kireev, «Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study,» *Journal of Chemical Information and Modeling,* vol. 54, pp. 2555-2561, 8 2014.

[75] P. J. Ballester, A. Schreyer et T. L. Blundell, «Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity?,» *Journal of Chemical Information and Modeling,* vol. 54, pp. 944-955, 2 2014.

[76] R. W. Homer, J. Swanson, R. J. Jilek, T. Hurst et R. D. Clark, «SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries,» *Journal of Chemical Information and Modeling,* vol. 48, pp. 2294-2307, 12 2008.

[77] E. B. Lenselink, W. Jespers, H. W. T. Vlijmen, A. P. IJzerman et G. J. P. Westen, «Interacting with GPCRs: Using Interaction Fingerprints for Virtual Screening,» *Journal of Chemical Information and Modeling,* vol. 56, pp. 2053-2060, 9 2016.

[78] M. M. Mysinger, M. Carchia, J. J. Irwin et B. K. Shoichet, «Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking,» *Journal of Medicinal Chemistry,* vol. 55, pp. 6582-6594, 7 2012.

[79] J.-F. Truchon et C. I. Bayly, «Evaluating Virtual Screening Methods: Good and Bad Metrics for the Early Recognition Problem,» *Journal of Chemical Information and Modeling,* vol. 47, pp. 488-508, 3 2007.

[80] «Dassault Systemes BIOVIA, BIOVIA Pipeline Pilot, Release 2017, San Diego: Dassault Systemes, 2017.,» 2017.

[81] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit et A. R. Leach, «The ChEMBL database in 2017,» *Nucleic Acids Research,* vol. 45, pp. D945--D954, 11 2016.

[82] M. Hendlich, «Databases for Protein–Ligand Complexes,» *Acta Crystallographica Section D Biological Crystallography,* vol. 54, pp. 1178-1182, 11 1998.

[83] A. Bergner, J. Günther, M. Hendlich, G. Klebe et M. Verdonk, «Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects.,» *Biopolymers,* vol. 61, n° %12, pp. 99-110, 2001.

[84] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin et G. Sherlock, «Gene Ontology: tool for the unification of biology,» *Nature Genetics,* vol. 25, pp. 25-29, 5 2000.

[85] A. Bairoch et R. Apweiler, «The SWISS-PROT protein sequence database: its relevance to human molecular medical research.,» *Journal of molecular medicine (Berlin, Germany),* vol. 75, n° %15, pp. 312-316, 5 1997.

[86] S. F. Altschul, W. Gish, W. Miller, E. W. Myers et D. J. Lipman, «Basic local alignment search tool,» *Journal of Molecular Biology,* vol. 215, pp. 403-410, 10 1990.

[87] D. Weininger, «SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,» *Journal of Chemical Information and Modeling,* vol. 28, pp. 31-36, 2 1988.

[88] V. L. Guilloux, P. Schmidtke et P. Tuffery, «Fpocket: An open source platform for ligand pocket detection,» *BMC Bioinformatics,* vol. 10, p. 168, 2009.

[89] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard et S. D. Morley, «rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids,» *PLoS Computational Biology,* vol. 10, p. e1003571, 4 2014.

[90] X. Q. Lewell, D. B. Judd, S. P. Watson et M. M. Hann, «RECAPRetrosynthetic Combinatorial Analysis Procedure:0.167em A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry,» *Journal of Chemical Information and Computer Sciences,* vol. 38, pp. 511-522, 5 1998.

[91] J. Gasteiger et M. Marsili, «A new model for calculating atomic charges in molecules,» *Tetrahedron Letters,* vol. 19, pp. 3181-3184, 1 1978.

[92] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland et J. Laufer, «Description of several chemical structure file formats used by computer

programs developed at Molecular Design Limited,» *Journal of Chemical Information and Modeling,* vol. 32, pp. 244-255, 5 1992.

[93] J. J. Dannenberg, «An Introduction to Hydrogen Bonding By George A. Jeffrey (University of Pittsburgh). Oxford University Press: New York and Oxford. 1997. ix 303 pp. 60.00. ISBN 0-19-509549-9.,» *Journal of the American Chemical Society,* vol. 120, pp. 5604-5604, 6 1998.

[94] S. Rajagopal et S. Vishveshwara, «Short hydrogen bonds in proteins,» *FEBS Journal,* vol. 272, pp. 1819-1832, 3 2005.

[95] R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger et F. M. Boeckler, «Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology,» *Journal of Medicinal Chemistry,* vol. 56, pp. 1363-1388, 1 2013.

[96] M. M. Harding, «Small revisions to predicted distances around metal sites in proteins,» *Acta Crystallographica Section D Biological Crystallography,* vol. 62, pp. 678-682, 5 2006.

[97] H. Zheng, D. R. Cooper, P. J. Porebski, I. G. Shabalin, K. B. Handing et W. Minor, «CheckMyMetal: a macromolecular metal-binding validation tool,» *Acta Crystallographica Section D Structural Biology,* vol. 73, pp. 223-233, 2 2017.

[98] M. Boehringer, H. Fischer, M. Hennig, D. Hunziker, J. Huwyler, B. Kuhn, B. M. Loeffler, T. Luebbers, P. Mattei, R. Narquizian, E. Sebokova, U. Sprecher et H. P. Wessel, «Aryl- and heteroaryl-substituted aminobenzo[a]quinolizines as dipeptidyl peptidase IV inhibitors,» *Bioorganic & Medicinal Chemistry Letters,* vol. 20, pp. 1106-1108, 2 2010.

[99] K. Kasahara et K. Kinoshita, «GIANT: pattern analysis of molecular interactions in 3D structures of protein–small ligand complexes,» *BMC Bioinformatics,* vol. 15, p. 12, 2014.

[100] K. Kasahara, M. Shirota et K. Kinoshita, «Comprehensive Classification and Diversity Assessment of Atomic Contacts in Protein–Small Ligand Interactions,» *Journal of Chemical Information and Modeling,* vol. 53, pp. 241-248, 12 2012.

[101] S. C. Althoen et R. McLaughlin, «Gauss-Jordan Reduction: A Brief History,» *The American Mathematical Monthly,* vol. 94, p. 130, 2 1987.

[102] L. Wright, X. Barril, B. Dymock, L. Sheridan, A. Surgenor, M. Beswick, M. Drysdale, A. Collier, A. Massey, N. Davies, A. Fink, C. Fromont, W. Aherne, K. Boxall, S. Sharp, P. Workman et R. E. Hubbard, «Structure-Activity Relationships in Purine-Based Inhibitor Binding to HSP90 Isoforms,» *Chemistry & Biology,* vol. 11, pp. 775-785, 6 2004.

[103] E. Casale, N. Amboldi, M. G. Brasca, D. Caronni, N. Colombo, C. Dalvit, E. R. Felder, G. Fogliatto, A. Galvani, A. Isacchi, P. Polucci, L. Riceputi, F. Sola, C. Visco, F. Zuccotto et F. Casuscelli, «Fragment-based hit discovery and structure-based optimization of aminotriazoloquinazolines as novel Hsp90 inhibitors,» *Bioorganic & Medicinal Chemistry,* vol. 22, pp. 4135-4150, 8 2014.

[104] W. Humphrey, A. Dalke et K. Schulten, «VMD: visual molecular dynamics.,» *Journal of molecular graphics,* vol. 14, n° 11, pp. 33--8, 27-8, 2 1996.

[105] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng et T. E. Ferrin, «UCSF Chimera?A visualization system for exploratory research and analysis,» *Journal of Computational Chemistry,* vol. 25, pp. 1605-1612, 2004.

[106] «The PyMOL Molecular Graphics System, Version 1.2r3pre».

[107] N. Huang, B. K. Shoichet et J. J. Irwin, «Benchmarking Sets for Molecular Docking,» *Journal of Medicinal Chemistry,* vol. 49, pp. 6789-6801, 11 2006.

[108] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis et P. S. Shenkin, «Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy,» *Journal of Medicinal Chemistry,* vol. 47, pp. 1739-1749, 3 2004.

[109] P. J. Ballester et W. G. Richards, «Ultrafast shape recognition for similarity search in molecular databases,» *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 463, pp. 1307-1321, 2 2007.

[110] S. Bu, L. Wang, P. Han, Z. Liu et K. Li, «3D shape recognition and retrieval based on multi-modality deep learning,» *Neurocomputing,* vol. 259, pp. 183-193, 10 2017.

[111] K. E. Hage, J.-P. Piquemal, Z. Hobaika, R. G. Maroun et N. Gresh, «Substituent-Modulated Affinities of Halobenzene Derivatives to the HIV-1 Integrase Recognition Site. Analyses of the Interaction Energies by Parallel Quantum Chemical and Polarizable Molecular Mechanics,» *The Journal of Physical Chemistry A,* vol. 118, pp. 9772-9782, 10 2014.

[112] R. S. Mulliken, «Structures of Complexes Formed by Halogen Molecules with Aromatic and with Oxygenated Solvents1,» *Journal of the American Chemical Society,* vol. 72, pp. 600-608, 1 1950.

[113] P. Murray-Rust et W. D. S. Motherwell, «Computer retrieval and analysis of molecular geometry. 4. Intermolecular interactions,» *Journal of the American Chemical Society,* vol. 101, pp. 4374-4376, 7 1979.

[114] A. Kovacs et Z. Varga, «Halogen acceptors in hydrogen bonding,» *Coordination Chemistry Reviews,* vol. 250, pp. 710-727, 3 2006.

[115] J. A. K. Howard, V. J. Hoy, D. OHagan et G. T. Smith, «How good is fluorine as a hydrogen bond acceptor?,» *Tetrahedron,* vol. 52, pp. 12613-12622, 9 1996.

[116] A. R. Voth, P. Khuu, K. Oishi et P. S. Ho, «Halogen bonds as orthogonal molecular interactions to hydrogen bonds,» *Nature Chemistry,* vol. 1, pp. 74-79, 4 2009.

[117] P. Auffinger, F. A. Hays, E. Westhof et P. S. Ho, «Halogen bonds in biological molecules,» *Proceedings of the National Academy of Sciences,* vol. 101, pp. 16789-16794, 11 2004.

[118] C. Dalvit, C. Invernizzi et A. Vulpetti, «Fluorine as a Hydrogen-Bond Acceptor: Experimental Evidence and Computational Calculations,» *Chemistry - A European Journal,* vol. 20, pp. 11058-11068, 7 2014.

[119] C. Dalvit et A. Vulpetti, «Weak Intermolecular Hydrogen Bonds with Fluorine: Detection and Implications for Enzymatic/Chemical Reactions, Chemical Properties, and Ligand/Protein Fluorine NMR Screening,» *Chemistry - A European Journal,* vol. 22, pp. 7592-7601, 4 2016.

[120] E. Carosati, S. Sciabola et G. Cruciani, «Hydrogen Bonding Interactions of Covalently Bonded Fluorine Atoms: From Crystallographic Data to a New Angular Function in the GRID Force Field,» *Journal of Medicinal Chemistry,* vol. 47, pp. 5114-5125, 10 2004.

[121] S. W. Rowlinson, J. R. Kiefer, J. J. Prusakiewicz, J. L. Pawlitz, K. R. Kozak, A. S. Kalgutkar, W. C. Stallings, R. G. Kurumbail et L. J. Marnett, «A Novel Mechanism of Cyclooxygenase-2 Inhibition Involving Interactions with Ser-530 and Tyr-385,» *Journal of Biological Chemistry,* vol. 278, pp. 45763-45769, 8 2003.

[122] C. Koch, A. Heine et G. Klebe, «Tracing the Detail: How Mutations Affect Binding Modes and Thermodynamic Signatures of Closely Related Aldose Reductase Inhibitors,» *Journal of Molecular Biology,* vol. 406, pp. 700-712, 3 2011.

[123] H. Wang, W. Wang et W. J. Jin, «-Hole Bond vs -Hole Bond: A Comparison Based on Halogen Bond,» *Chemical Reviews,* vol. 116, pp. 5072-5104, 2 2016.

[124] G. Gerebtzoff, X. Li-Blatter, H. Fischer, A. Frentzel et A. Seelig, «Halogenation of Drugs Enhances Membrane Binding and Permeation,» *ChemBioChem,* vol. 5, pp. 676-684, 4 2004.

[125] M. Hernandes, S. M. Cavalcanti, D. R. Moreira, W. Azevedo Junior et A. C. Leite, «Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design,» *Current Drug Targets,* vol. 11, pp. 303-314, 3 2010.

[126] Y. Lu, Y. Wang et W. Zhu, «Nonbonding interactions of organic halogens in biological systems: implications for drug discovery and biomolecular design,» *Physical Chemistry Chemical Physics,* vol. 12, p. 4543, 2010.

[127] H. Matter, M. Nazaré, S. Güssregen, D. Will, H. Schreuder, A. Bauer, M. Urmann, K. Ritter, M. Wagner et V. Wehner, «Evidence for C-Cl/C-Br⋯ Interactions as an Important Contribution to Protein-Ligand Binding Affinity,» *Angewandte Chemie International Edition,* vol. 48, pp. 2911-2916, 3 2009.

[128] S. Maignan, J.-P. Guilloteau, Y. M. Choi-Sledeski, M. R. Becker, W. R. Ewing, H. W. Pauls, A. P. Spada et V. Mikol, «Molecular Structures of Human Factor Xa Complexed with Ketopiperazine Inhibitors: Preference for a Neutral Group in the S1 Pocket,» *Journal of Medicinal Chemistry,* vol. 46, pp. 685-690, 2 2003.

[129] S. Sirimulla, J. B. Bailey, R. Vegesna et M. Narayan, «Halogen Interactions in Protein–Ligand Complexes: Implications of Halogen Bonding for Rational Drug Design,» *Journal of Chemical Information and Modeling,* vol. 53, pp. 2781-2791, 11 2013.

[130] L. A. Hardegger, B. Kuhn, B. Spinnler, L. Anselm, R. Ecabert, M. Stihle, B. Gsell, R. Thoma, J. Diez, J. Benz, J.-M. Plancher, G. Hartmann, D. W. Banner, W. Haap et F. Diederich, «Systematic Investigation of Halogen Bonding in Protein-Ligand Interactions,» *Angewandte Chemie International Edition,* vol. 50, pp. 314-318, 12 2010.

[131] G. Cavallo, P. Metrangolo, R. Milani, T. Pilati, A. Priimagi, G. Resnati et G. Terraneo, «The Halogen Bond,» *Chemical Reviews,* vol. 116, pp. 2478-2601, 1 2016.

[132] S. Aravinda, N. Shamala, C. Das, A. Sriranjini, I. L. Karle et P. Balaram, «Aromatic-Aromatic Interactions in Crystal Structures of Helical Peptide Scaffolds Containing Projecting Phenylalanine Residues,» *Journal of the American Chemical Society,* vol. 125, pp. 5308-5315, 5 2003.

[133] E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly et N. A. Meanwell, «Applications of Fluorine in Medicinal Chemistry,» *Journal of Medicinal Chemistry,* vol. 58, pp. 8315-8359, 7 2015.

[134] P. Zhou, J. Zou, F. Tian et Z. Shang, «Fluorine Bonding — How Does It Work In Protein-Ligand Interactions?,» *Journal of Chemical Information and Modeling,* vol. 49, pp. 2344-2355, 10 2009.

[135] E. V. Bartashevich et V. G. Tsirelson, «Interplay between non-covalent interactions in complexes and crystals with halogen bonds,» *Russian Chemical Reviews,* vol. 83, pp. 1181-1203, 12 2014.

[136] C. A. Hunter, J. Singh et J. M. Thornton, «π-π interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins,» *Journal of Molecular Biology,* vol. 218, pp. 837-846, 4 1991.

[137] C. A. Hunter, K. R. Lawson, J. Perkins et C. J. Urch, «Aromatic interactions,» *Journal of the Chemical Society, Perkin Transactions 2,* pp. 651-669, 2001.

[138] T. Clark, M. Hennemann, J. S. Murray et P. Politzer, «Halogen bonding: the σ-hole,» *Journal of Molecular Modeling,* vol. 13, pp. 291-296, 8 2006.

[139] L. A. Hardegger, B. Kuhn, B. Spinnler, L. Anselm, R. Ecabert, M. Stihle, B. Gsell, R. Thoma, J. Diez, J. Benz, J.-M. Plancher, G. Hartmann, Y. Isshiki, K. Morikami, N. Shimma, W. Haap, D. W. Banner et F. Diederich, «Halogen Bonding at the Active Sites of Human Cathepsin0.25emL and MEK1 Kinase: Efficient Interactions in Different Environments,» *ChemMedChem,* vol. 6, pp. 2048-2054, 9 2011.

[140] I. S. Gutiérrez, F.-Y. Lin, K. Vanommeslaeghe, J. A. Lemkul, K. A. Armacost, C. L. Brooks et A. D. MacKerell, «Parametrization of halogen bonds in the CHARMM general force field: Improved treatment of ligand–protein interactions,» *Bioorganic & Medicinal Chemistry,* vol. 24, pp. 4812-4825, 10 2016.

[141] C. A. Sotriffer, H. Gohlke et G. Klebe, «Docking into Knowledge-Based Potential Fields: A Comparative Evaluation of DrugScore,» *Journal of Medicinal Chemistry,* vol. 45, pp. 1967-1970, 5 2002.

[142] T. Langer et R. D. Hoffmann, Éds., Pharmacophores and Pharmacophore Searches, Wiley-VCH Verlag GmbH & Co. KGaA, 2006.

[143] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, J. Moult, T. Schwede et A. Tramontano, «Evaluation of the template-based modeling in CASP12,» *Proteins: Structure, Function, and Bioinformatics,* vol. 86, pp. 321-334, 12 2017.

[144] S. Griep et U. Hobohm, «PDBselect 1992–2009 and PDBfilter-select,» *Nucleic Acids Research,* vol. 38, pp. D318--D319, 9 2009.

[145] G. Wang et R. L. Dunbrack, «PISCES: a protein sequence culling server,» *Bioinformatics,* vol. 19, pp. 1589-1591, 8 2003.

[146] Akiyama, Onizuka, Noguchi et Ando, «Parallel Protein Information Analysis (PAPIA) System Running on a 64-Node PC Cluster.,» *Genome informatics. Workshop on Genome Informatics,* vol. 9, pp. 131-140, 1998.

[147] W. Li et A. Godzik, «Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,» *Bioinformatics,* vol. 22, pp. 1658-1659, 5 2006.

[148] K. Sikic et O. Carugo, «CARON – Average RMSD of NMR structure ensembles,» *Bioinformation,* vol. 4, pp. 132-133, 9 2009.

[149] L. Calvanese, G. D'Auria, A. Vangone, L. Falcigno et R. Oliva, «Analysis of the interface variability in NMR structure ensembles of protein–protein complexes,» *Journal of Structural Biology,* vol. 194, pp. 317-324, 6 2016.

[150] B. Schneider, J. Černý, D. Svozil, P. Čech, J.-C. Gelly et A. G. Brevern, «Bioinformatic analysis of the protein/DNA interface,» *Nucleic Acids Research,* vol. 42, pp. 3381-3394, 12 2013.

[151] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin et H. A. Carlson, «Binding MOAD, a high-quality protein ligand database,» *Nucleic Acids Research,* vol. 36, pp. D674--D678, 12 2007.

[152] R. Wang, X. Fang, Y. Lu et S. Wang, «The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures,» *Journal of Medicinal Chemistry,* vol. 47, pp. 2977-2980, 6 2004.

[153] E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata et D. Rognan, «sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank,» *Journal of Chemical Information and Modeling,* vol. 46, pp. 717-727, 3 2006.

[154] I. Wallach et R. Lilien, «The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding,» *Bioinformatics,* vol. 25, pp. 615-620, 1 2009.

[155] M. N. Drwal, G. Bret, C. Perez, C. Jacquemard, J. Desaphy et E. Kellenberger, «Structural Insights on Fragment Binding Mode Conservation,» *Journal of Medicinal Chemistry,* vol. 61, pp. 5963-5973, 6 2018.

[156] L. Breuza, S. Poux, A. Estreicher, M. L. Famiglietti, M. Magrane, M. Tognolli, A. Bridge, D. Baratin et N. Redaschi, «The UniProtKB guide to the human proteome,» *Database,* vol. 2016, p. bav120, 2016.

[157] «CHEMBL database release 23,» EMBL-EBI, 2017.

[158] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate et A. Bateman, «The Pfam protein families database: towards a more sustainable future,» *Nucleic Acids Research,* vol. 44, pp. D279--D285, 12 2015.

[159] C. J. A. Sigrist, E. Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret et I. Xenarios, «New and continuing developments at PROSITE,» *Nucleic Acids Research,* vol. 41, pp. D344--D347, 11 2012.

[160] W. Kabsch, «A solution for the best rotation to relate two sets of vectors,» *Acta Crystallographica Section A,* vol. 32, pp. 922-923, 9 1976.

[161] R. D. C. Team, «R: A Language and Environment for Statistical Computing,» Vienna, 2017.

[162] A. Bornot, C. Etchebest et A. G. Brevern, «Predicting protein flexibility through the prediction of local structures,» *Proteins: Structure, Function, and Bioinformatics,* vol. 79, pp. 839-852, 12 2010.

[163] T. Narwani, P. Craveur, N. Shinada, H. Santuz, J. Rebehmed, C. Etchebest et B. , «Dynamics and deformability of -, 310- and -helices,» *Archives of Biological Sciences,* vol. 70, pp. 21-31, 2018.

[164] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. Spoel, B. Hess et E. Lindahl, «GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit,» *Bioinformatics,* vol. 29, pp. 845-854, 2 2013.

[165] W. F. V. Gunsteren, Biomolecular Simulation: GROMOS 96 Manual and User Guide, Verlag der Fachvereine Hochschulverlag AG an der ETH Zurich, 1996.

[166] W. Kabsch et C. Sander, «Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,» *Biopolymers,* vol. 22, pp. 2577-2637, 12 1983.

[167] A. G. Brevern, C. Etchebest et S. Hazout, «Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks,» *Proteins: Structure, Function, and Genetics,* vol. 41, pp. 271-287, 2000.

[168] W. G. Touw, C. Baakman, J. Black, T. A. H. Beek, E. Krieger, R. P. Joosten et G. Vriend, «A series of PDB-related databanks for everyday needs,» *Nucleic Acids Research,* vol. 43, pp. D364--D368, 10 2014.

[169] T. J. Narwani, H. Santuz, N. Shinada, A. M. Vattekatte, Y. Ghouzam, N. Srinivasan, J.-C. Gelly et A. G. Brevern, «Recent advances on polyproline II,» *Amino Acids,* vol. 49, pp. 705-713, 2 2017.

[170] Y. Mansiaux, A. P. Joseph, J.-C. Gelly et A. G. Brevern, «Assignment of PolyProline II Conformation and Analysis of Sequence – Structure Relationship,» *PLoS ONE,* vol. 6, p. e18401, 3 2011.

[171] J. Franz, M. Lelle, K. Peneva, M. Bonn et T. Weidner, «SAP(E) – A cell-penetrating polyproline helix at lipid interfaces,» *Biochimica et Biophysica Acta (BBA) - Biomembranes,* vol. 1858, pp. 2028-2034, 9 2016.

[172] P. Craveur, A. P. Joseph, J. Esque, T. J. Narwani, F. NoÃl, N. Shinada, M. Goguet, S. Leonard, P. Poulain, O. Bertrand, G. Faure, J. Rebehmed, A. Ghozlane, L. S. Swapna, R. M. Bhaskara, J. Barnoud, S. TÃletchÃa, V. Jallu, J. Cerny, B. Schneider, C. Etchebest, N. Srinivasan, J.-C. Gelly et A. G. Brevern, «Protein flexibility in the light of structural alphabets,» *Frontiers in Molecular Biosciences,* vol. 2, 5 2015.

[173] G. E. Crooks, «WebLogo: A Sequence Logo Generator,» *Genome Research,* vol. 14, pp. 1188-1190, 5 2004.

[174] R. B. Best et J. Mittal, «Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences,» *Proteins: Structure, Function, and Bioinformatics,* vol. 79, pp. 1318-1328, 2 2011.

[175] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York et M. Karplus, «CHARMM: The biomolecular simulation program,» *Journal of Computational Chemistry,* vol. 30, pp. 1545-1614, 7 2009.

[176] H. W. Kuhn, «The Hungarian method for the assignment problem,» *Naval Research Logistics Quarterly,* vol. 2, pp. 83-97, 3 1955.

[177] J. M. Hancock et M. J. Zvelebil, Éds., Dictionary of Bioinformatics and Computational Biology, John Wiley & Sons, Inc., 2004.

[178] M. R. Hestenes, Conjugate Direction Methods in Optimization, Springer New York, 2012.

[179] C. J. Cramer et D. G. Truhlar, «Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics,» *Chemical Reviews,* vol. 99, pp. 2161-2200, 8 1999.

[180] R. M. Levy et E. Gallicchio, «COMPUTER SIMULATIONS WITH EXPLICIT SOLVENT: Recent Progress in the Thermodynamic Decomposition of Free Energies and in Modeling Electrostatic Effects,» *Annual Review of Physical Chemistry,* vol. 49, pp. 531-567, 10 1998.

[181] C. Etchebest, C. Benros, S. Hazout et A. G. Brevern, «A structural alphabet for local protein structures: Improved prediction methods,» *Proteins: Structure, Function, and Bioinformatics,* vol. 59, pp. 810-827, 4 2005.

Annexe 1

Column Name	Type	Information
CONTACT_ID	INTEGER	Contact Identifier
Ligand informations		
SMALL_MOL_CONF_ID	INTEGER	Ligand identifier
LIG_ELEMENT	VARCHAR	Ligand atomic element
LIG_ATOM_NAME	VARCHAR	Ligand PDB atom label
LIG_ATOM_NUMBER	INTEGER	Ligand PDB atom index
LIG_ATT_ATOM	VARCHAR	Ligand covalently bonded elements
LIG_HYBRIDISATION	VARCHAR	Ligand atom hybridization
LIG_HBA	INTEGER	Ligand is hydrogen bond acceptor
LIG_HBD	INTEGER	Ligand is hydrogen bond donor
LIG_NUM_LONE_PAIR	INTEGER	Ligand number of lone pair available
LIG_NUM_H	INTEGER	Ligand number of implicit hydrogen
LIG_IS_AROMATIC	INTEGER	Ligand is aromatic
Receptor informations		
STRUCTURE_STATE_ID	INTEGER	Structure of complex identifier
STR_RES_NAME	VARCHAR	Receptor PDB residue name
STR_RES_NUMBER	VARCHAR	Receptor PDB residue number
STR_CHAIN_CODE	VARCHAR	Receptor PDB residue chain
REC_ELEMENT	VARCHAR	Receptor atomic element
REC_ATOM_NAME	VARCHAR	Receptor PDB atom label
REC_ATOM_NUMBER	INTEGER	Receptor PDB atom index
RECEPTOR_ATT_ATOM	VARCHAR	Receptor covalently bonded elements
REC_HYBRIDISATION	VARCHAR	Receptor atom hybridization
REC_HBA	INTEGER	Receptor is hydrogen bond acceptor
REC_HBD	INTEGER	Receptor is hydrogen bond donor
REC_NUM_LONE_PAIR	INTEGER	Receptor number of lone pair available
REC_NUM_H	INTEGER	Receptor number of implicit hydrogen
REC_IS_AROMATIC	INTEGER	Receptor is aromatic
NOT_PROTEIN	INTEGER	Receptor is protein
REC_ATOM_XYZ	VARCHAR	Receptor atom 3D Coordinates in structure
Fragment information		
ATOM_INDEX_FRAGMENT	INTEGER	Ligand atom number in fragment
FRAGMENT_IDX	INTEGER	Fragment identifier
CTAB_3D	CLOB	Chemical Table of ligand fragment
LIG_FRAG_COM	VARCHAR	3D coordinates of ligand fragment center of mass
Geometric descriptors		
LIGAND_ANGLE	FLOAT	Angle of receptor atom compared to ligand vector
RECEPTOR_ANGLE	FLOAT	Angle of ligand atom compared to receptor vector
CROSS_ANGLE	FLOAT	Angle between ligand vector and receptor vector
DISTANCE	FLOAT	Distance between two atoms
DISTANCE_FRAG	FLOAT	Distance between center of mass of two fragments

Annexe 2

La dynamique moléculaire (DM) est une méthode de simulation informatique permettant l'étude du mouvement des atomes et des molécules. Lors d'une dynamique, des étapes successives d'un système sont générées à partir des lois du mouvement de Newton. La molécule analysée a ainsi une *trajectoire* décrivant la position et mouvement des particules du système en fonction du temps. Le déplacement des coordonnées spatiales est obtenu par la résolution des équations différentielles de la seconde loi de Newton (F = ma) :

$$\frac{\delta^2 x_i}{\delta t^2} = \frac{F_x}{m_i}$$

Cette équation décrit les mouvements d'une particule de masse *m* à travers sa coordonnée *x* avec F_x la force d'une particule dans cette direction. Les simulations de dynamique sont fondées sur l'assomption qu'un système est régi par l'hypothèse d'ergodicité, à savoir que tous les états intermédiaires sont largement équiprobables dans le temps. Une dynamique moléculaire débute par la définition des coordonnées initiales du système auxquelles vont être appliqués des vitesses initiales générées aléatoirement. L'évolution du système sera obtenue itérativement par un pas de temps prédéfini le plus souvent de l'ordre de la femtoseconde.

La position des atomes, leur vitesse respective ainsi que les forces qui leur sont appliqués dans le système sont calculées et mis à jour à chaque étape. Le procédé est répété de manière itérative jusqu'à ce que la trajectoire atteigne le temps de simulation désiré. Certains concepts doivent être appréhendés pour comprendre le fonctionnement de la dynamique moléculaire.

1. Champs de force

Afin de calculer l'énergie potentielle du système, des équations mathématiques ainsi que des paramètres ont été définies pour décrire un champ de force. L'énergie potentielle permet de quantifier la force exercée sur un atome. Un champ de force est donc un ensemble d'équations et de constantes censées reproduire les forces appliquées sur chaque particule d'un système.

Les constantes et équations des champs de forces sont issues d'expériences réalisées dans le domaine de la mécanique physique et de la chimique quantique. L'équation du potentiel

énergétique comprend la considération des énergies d'interaction covalentes ($E_{liée}$) et intermoléculaires ($E_{non-liée}$). Ces valeurs et fonctions varient en fonction du champ de force utilisé, néanmoins la formule de l'énergie potentielle d'interaction est généralement résumée à :

$$E_{total} = E_{liée} + E_{non-liée}$$

$$E_{liée} = E_{liaison} + E_{angle} + E_{dièdre}$$

$$E_{non-liée} = E_{electrostatique} + E_{Van\ der\ Waals}$$

Les valeurs énergétiques d'angles et de liaisons covalentes sont dérivées de fonctions énergétiques quadratiques tandis que les énergies d'interactions moléculaires sont dérivées des potentiels de Lennard-Jones et de Coulomb. La fonction énergétique *U(R)* est calculée selon la formule suivante :

$$U(R) = \sum_{liaison} k_i^{liaison}(r_i - r_0)^2 + \sum_{angle} k_i^{angle}(\theta_i - \theta_0)^2 + \sum_{dièdre} k_i^{dièdre}[1 + \cos(n_i\varphi_i + \delta_i)] + \sum_i \sum_{j \neq i} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\varepsilon r_{ij}}$$

Il existe différents champs de force ayant tous un champ d'application plus spécifique. Ainsi le champ de force AMBER [165] est souvent utilisé pour l'ADN et les protéines, tandis que CHARMM [175] est utilisé pour les petites molécules type ligands et macromolécules. Le champ de force GROMOS [165] a une application plus générale.

2. <u>Minimisation énergétique</u>

Une dynamique moléculaire permet l'exploration de l'ensemble des conformations d'une biomolécule en assumant que la conformation initiale se trouve dans un minimum énergétique. Or, les structures cristallographiques ne correspondent pas toujours à l'état d'équilibre optimal de la biomolécule du fait du processus même de la cristallisation. Il est donc essentiel de minimiser la fonction exprimant l'énergie potentielle du système avant de

débuter une dynamique molécule. Deux méthodes sont couramment utilisées pour minimiser l'énergie initiale d'un système : *steepest descent* [177] et *gradient conjugué* [178].

Steepest descent, en français algorithme du gradient, est la méthode la plus simple pour réaliser une minimisation énergétique. L'algorithme itératif cherche à chaque étape la pente la plus forte, opposée au gradient directionnel. Cet algorithme est très efficace lors des premières étapes de minimisation mais devient relativement lent dans la convergence finale. La méthode du gradient conjugué quant à elle, conserve les informations des étapes précédentes et optimise la recherche de direction à partir de cet historique.

3. La solvatation du système

Tous systèmes biologiques se situent dans un milieu aqueux, nécessitant ainsi la solvatation du système. La molécule est généralement entourée d'une boite de solvant avant l'étape de minimisation. Deux types de solvatation sont généralement distingués : la solvatation dite *implicite* [179] et *explicite* [180] (voir Figure 104). Dans le modèle implicite, le solvant est défini comme un continuum homogène polarisé dont les propriétés est proche de celles du solvant. Le modèle explicite, plus précis et réaliste mais plus coûteux informatiquement, modélise chaque molécule du solvant individuellement, résultant dans le calcul des forces, vitesses et positions de chaque molécule.

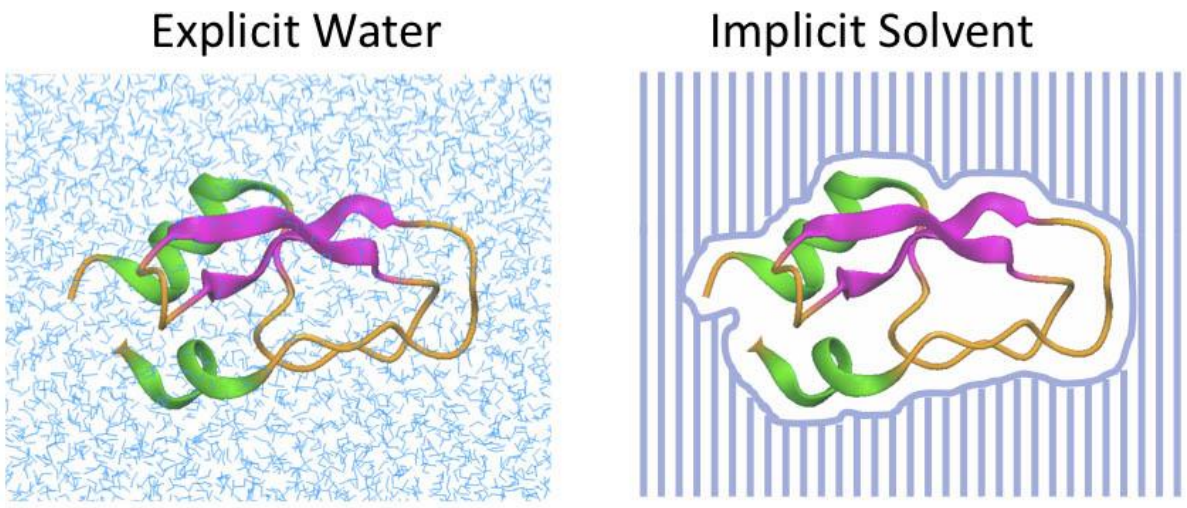


Figure 104 Description des deux modèles de solvant utilisés en dynamique moléculaire.

4. Conditions périodiques aux limites

Dans la majorité des dynamiques moléculaires, la boite de simulation est assez large pour limiter tout effet de bord (la molécule pouvant se voir à courte distance, ce qui biaise la

dynamique). Des conditions périodiques limites sont être appliquées afin de minimiser au mieux ces effets, simulant un environnement infini. La Figure 105 représente schématiquement ces conditions périodiques aux limites où le contact d'une molécule sur un bord la fait apparaitre sur le bord opposé. Pour obtenir ce résultat, la boite de simulation est dupliquée dans les trois directions cartésiennes. La conséquence directe de cette multiplication est l'augmentation du nombre d'interactions à calculer puisqu'il est nécessaire de prendre en compte les éléments présents dans les boites adjacentes. Une distance limite est défini pour le calcul du potentiel énergétique dans ces boites voisines.

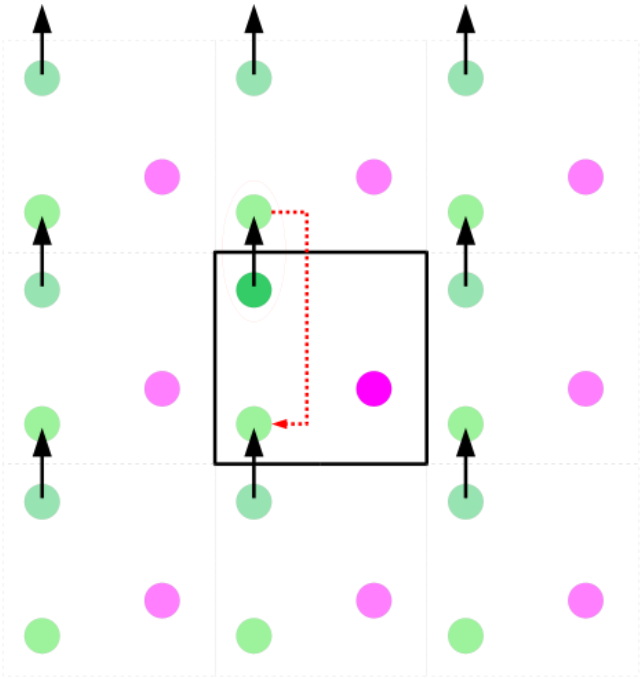


Figure 105 Représentation schématique de l'application des conditions périodiques aux limites.

5. Ensemble statistique

Les ensembles statistiques appliqués à la mécanique moléculaire constituent des contraintes spécifiques imposées au système. Différents ensembles sont couramment utilisés lors de dynamique tels que l'ensemble canonique NVT maximisant la conservation du nombre de molécule (N), du volume (V) et de la température du système (T). L'ensemble isotherme-isobarique NPT conserve le nombre de molécules (N), la pression (P) ainsi que la température (T) du système. Enfin, l'ensemble microcanonique NVE, en plus du nombre de molécules et du volume, l'énergie du système (E) est conservée.

Annexe 3

L'alphabet structural des Blocs Protéiques (BPs) a été développé au sein du laboratoire par de Brevern et collaborateurs en 2000 [167]. Composé de 16 lettres, nommées de *a* à *p*, chacune des lettres correspond à un pentapeptide de la chaine principale caractérisé par 8 angles dièdres ψ et ϕ (voir Figure 106 et Tableau 6). Leur construction en 2000 à partir de 342 protéines non-redondantes [167] et leur réévaluation en 2005 sur 717 protéines ont montré son extrême stabilité [181]. Développés pour caractériser une approximation fine de la structure locale des protéines, leur emploi s'est étendu à la prédiction de structure locale à partir de sa séquence.

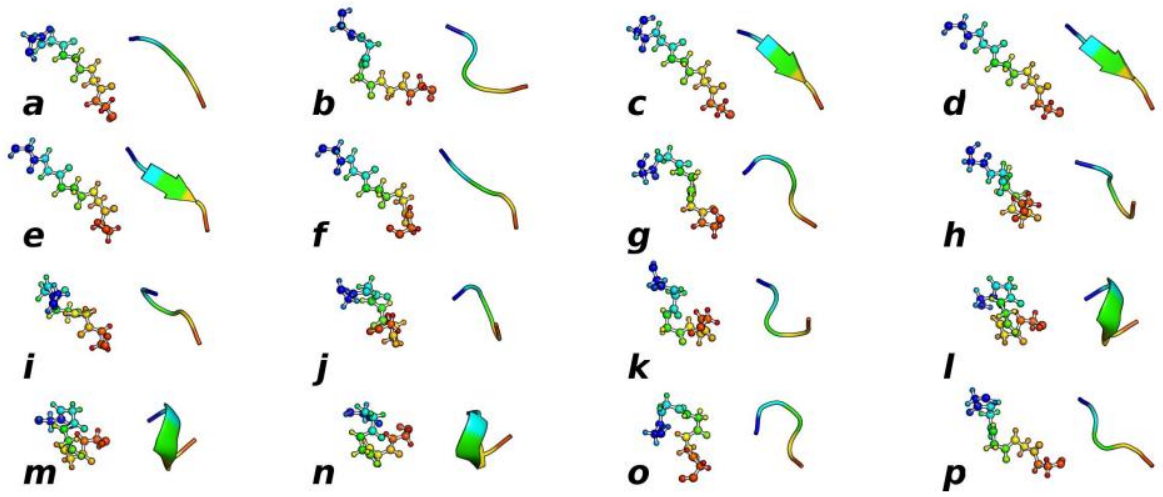


Figure 106 Représentation 3D des conformations représentatives de chaque Bloc Protéique.

BP	$\psi_{(n-2)}$	$\varphi_{(n-1)}$	$\psi_{(n-1)}$	$\varphi_{(n)}$	$\psi_{(n)}$	$\varphi_{(n+1)}$	$\psi_{(n+1)}$	$\varphi_{(n+2)}$
a	41.14	75.53	13.92	-99.8	131.88	-96.27	122.08	-99.68
b	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.9
c	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
d	141.98	-112.79	132.20	-114.79	140.11	-111.05	139.54	-103.16
e	133.25	-112.37	137.64	-108.13	133.00	-87.30	120.54	77.40
f	116.40	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
g	0.40	-81.83	4.91	-100.59	85.50	-71.65	130.78	84.98
h	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-90.88
i	130.68	-56.92	119.26	77.85	10.42	-99.43	141.40	-98.01
j	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
k	117.16	-95.41	140.40	-59.35	-29.23	-72.39	-25.08	-76.16
l	139.20	-55.96	-32.70	-68.51	-26.09	-74.44	-22.60	-71.74
m	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
n	-35.34	-65.03	-38.12	-66.34	-29.51	-89.10	-2.91	77.90
o	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
p	-27.09	-86.14	0.30	59.85	21.51	-96.30	132.67	-92.91

Tableau 6 Définition des angles dièdraux représentatifs de chaque bloc protéique, n représentant l'indice du résidu central.

Chaque BP représentant une conformation locale 3D spécifique, il est donc possible de représenter une structure 3D par la succession 1D de BPs assignés, illustré sur la Figure 107. Les BPs ainsi que les structures secondaires sont définis par des valeurs d'angles dièdres spécifiques, indiquant une relation forte entre les deux éléments. Cette relation peut être exprimée par le biais d'observations récurrentes mais ne correspond en aucun cas dans une relation directe. Le BP *d* par exemple reflète une conformation fréquemment retrouvées dans le cœur des brins β tandis que le groupe de BPs *a*, *b*, *c* ainsi que le groupe *e* et *f* sont plutôt situés aux extrémités N-terminales et C-terminales de ces mêmes brins respectivement. De

même, les BPs *m* décrivent le cœur des hélices α tandis que les groupes de BPs *k*, *l* et *n*, *o*, *p* vont représenter les régions adjacentes des hélices. Les boucles sont essentiellement représentées par les blocs *g*, *h*, *i* et *j*.

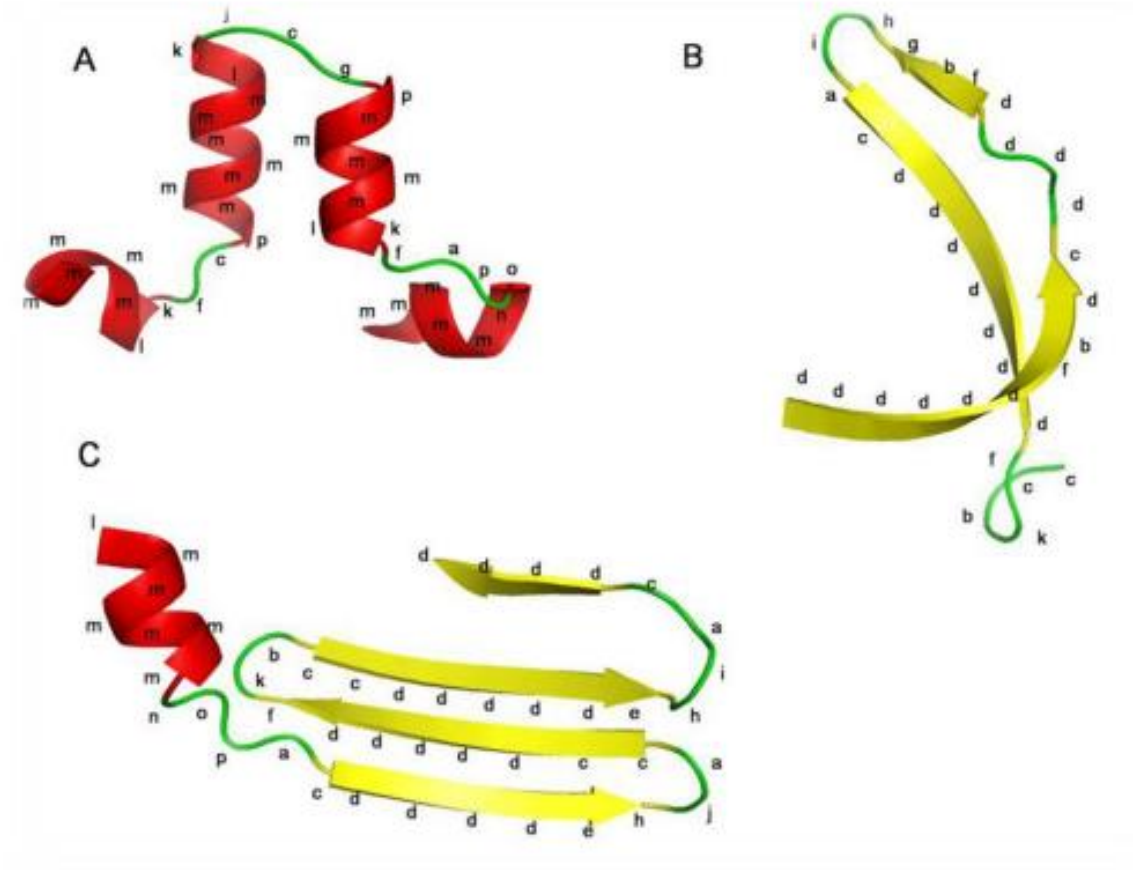


Figure 107 Correspondance entre structures 3D, structures secondaires et description 1D des Blocs Protéiques.

L'alphabet des BPs ainsi qu'un logiciel d'assignation des structures 3D en séquence BP sont accessibles au grand public https://github.com/pierrepo/PBxplore.