Recherche d'inhibiteurs de l'interaction Lutheran-Laminine par des techniques de modélisation et de simulation moléculaires

Noelly Madeleine

DSIMB, Dynamique des Structures et Interactions des Macromolécules Biologiques

(INSERM UMR-S 1134, Université Paris 7, Université de la Réunion, INTS)

Thèse de l'Université de La Réunion

Spécialité: Biologie Informatique

Présentée par

Mme MADELEINE Noëlly

Dirigée par

M. Fabrice GARDEBIEN

Pour obtenir le titre de Docteur de l'Université de la Réunion

---

Recherche d'inhibiteurs de l'interaction Lutheran-Laminine par des techniques de modélisation et de simulation moléculaires

---

Soutenue publiquement le 28 septembre 2017, devant le jury composé de:

Mme MITEVA Maria, Directrice de Recherche MTi, Inserm U973 - Université Paris Diderot	Rapporteur
M. MAIGRET Bernard, Directeur de Recherche Capsid team, CNRS–INRIA–LORIA - Université de Lorraine	Rapporteur
Mme MOUAWAD Liliane, Chargé de Recherche Institut Curie - CNRS UMR 9187–Inserm U1196	Examinatrice
M. STIGLIANI Jean-Luc, Maître de conférences Laboratoire de Chimie de Coordination du CNRS	Examinateur

Cette thèse à reçu le soutien financier de la Région Réunion.

À mes grands-parents que je ne cesserai jamais d'aimer,

À mes mamans,

À Edouard, à nous deux.

À ma famille,

À mes rayons de soleil Coco et Paumie.

"Pour inventer, il faut penser à côté."

Paul Souriau

Table des Matières

Introduction **8**

1 Méthodologie **23**

1.1 Recherche de protéines homologues ou analogues à Lu 23

1.2 Préparation des protéines 26

1.3 Élaboration d'un protocole de scoring à l'aide de la protéine CD80 27

1.4 Criblage virtuel de la chimiothèque ZINC appliqué à Lu 31

1.5 Les champs de force 35

1.6 Docking moléculaire 39

1.7 Fonctions de scoring secondaire 50

1.7.1 Fonction de scoring empirique 51

1.7.2 Fonction de scoring knowledge-based 57

1.7.3 Fonction de scoring basée sur le champ de force 58

1.8 Simulations de dynamique moléculaire 59

1.8.1 Simulations de dynamique moléculaire Langevin, LD 60

1.8.2 Simulations de dynamique moléculaire aux limites stochastiques ou *Stochastic Boundary Molecular Dynamics*, SBMD 61

2 Recherche de systèmes similaires Lu-protéine ou Lu-molécule **65**

3 Élaboration et validation d'un protocole de scoring **75**

3.1 Description de la pose sélectionnée pour les ligands 1 à 17 sur CD80 75
3.2 Recherche du protocole de scoring 80
3.2.1 Étape de relaxation en présence d'eau explicite 82
3.2.2 Évaluation des énergies d'interaction 91

4 Calcul des énergies d'interaction avec des méthodes de chimie quantique 103
4.1 Calcul des affinités des ligands avec la méthode PM6-DH2X 106
4.2 Calcul des affinités des ligands avec la méthode FMO 106
4.2.1 Description des différents protocoles testés 108
4.2.2 Choix de la base utilisée dans les calculs FMO-MP2 115
4.2.3 Comparaison des affinités calculées à partir des différentes variantes FMO avec les affinités expérimentales pour les ligands 1 - 8 et 14 - 16 118
4.3 Analyse des interactions par résidu 124

5 Recherche d'inhibiteurs d'interaction de Lu par criblage virtuel 133
5.1 Quel est l'intérêt d'utiliser une fonction de scoring secondaire et de réaliser une étape de relaxation dans un criblage virtuel ? 137
5.1.1 Quel est l'intérêt d'utiliser une fonction de scoring secondaire dans un criblage virtuel ? 137
5.1.2 Quel est l'intérêt de prendre en compte la flexibilité des complexes dans les calculs de score ? 139
5.2 Résultats obtenus sur les différentes étapes du criblage 140
5.3 Analyse des résultats obtenus à la fin du criblage 150
5.4 Analyse des 20 ligands de meilleurs scores 155
5.4.1 Réseau de liaisons hydrogène 158
5.4.2 Analyse des complexes formés entre Lu et chacun des 12 ligands restants 167

5.4.3 Différences entre les résultats DOCK6 et les résultats rDOCK 176

Conclusion et perspectives **180**

Bibliographie **183**

Annexes **201**

1. *Publication de la référence [1] qui présente les travaux réalisés pour le criblage de 395 601 molécules sur Lu.*

subsection**2.** *Résultats expérimentaux pour les trois meilleurs ligands issus du criblage de 395 601 molécules sur Lu.*

Introduction

La drépanocytose est une maladie génétique qui a été reconnue priorité de santé publique par l'OMS et l'UNESCO (après le cancer, l'infection par le VIH et le paludisme). C'est la maladie génétique la plus répandue dans le monde. En France, elle touche environ 15 000 personnes et, à La Réunion, elle concernerait un couple sur 65, ce qui correspond à environ 185 grossesses par an. Cette maladie, aussi appelée anémie falciforme, touche les globules rouges : on parlera de globules rouges drépanocytaires ou falciformes. Le globule rouge est formé d'hémoglobines composées chacune de deux chaînes alpha (α-globine) et de deux chaînes bêta (β-globine). Dans le cas du globule rouge drépanocytaire, le gène qui code pour la bêta-globine localisé sur le chromosome 11 a subit une mutation : l'hémoglobine anormale (hémoglobine S) résultante a tendance à polymériser et à s'agglomérer lorsque la concentration d'oxygène dans le sang est faible (hypoxie). Ceci conduit à la déformation des hématies c'est-à-dire que les globules rouges adoptent une forme de faucille au lieu de prendre la forme biconcave observée chez les personnes non malades. Ces globules rouges anormaux sont alors moins flexibles et se déplacent plus difficilement dans les vaisseaux sanguins que les globules rouges normaux. En plus du ralentissement provoqué par la forme en faucille des globules drépanocytaires, les malades sont plus vulnérables aux infections, ils sont sujets à une anémie causée par une destruction précoce des globules rouges anormaux (anémie hémolytique) et sont victimes de crises vaso-occlusives épisodiques très douloureuses.

Les causes des vaso-occlusions observées chez les personnes atteintes de drépanocytose sont multiples et les mécanismes les mieux connus sont présentés dans la Figure 1 (p. 10) [2, 3]. Le processus de vaso-occlusion chez les drépanocytaires est constitué de deux étapes : (i) la production et la liaison de réticulocytes de stress (globules rouges immatures) à l'endothélium vasculaire suivies de (ii) la forte adhésion des globules rouges vieillis et déformés à la matrice extracellulaire sous-endothéliale (lame basale). Dans

l'étape (i) ci-dessus, les réticulocytes de stress sont produits en grande quantité (érythropoïèse accrue) afin de contre-balancer l'anémie hémolytique observée chez les drépanocytaires. Ces réticulocytes de stress expriment les protéines membranaires CD36 et VLA-4 qui se lient respectivement aux protéines CD36 (grâce à un pont de thrombospondine) et VCAM-1, toutes les deux exprimées par l'endothélium vasculaire. L'adhésion de ces globules rouges immatures à l'endothélium vasculaire favorise le ralentissement du flux sanguin, ce qui a pour conséquence une diminution du taux d'oxygène dans le sang (hypoxie). Dans cet environnement faible en oxygène, l'hémoglobine S polymérise et s'agglomère entraînant ainsi une déformation en faucille (falciformation) des réticulocytes de stress au cours du temps : ils deviennent ainsi des globules rouges drépanocytaires falciformes. Les globules rouges drépanocytaires (vieillis et déformés) surexpriment la protéine membranaire Lutheran/BCAM qui se lie fortement et spécifiquement à la protéine Laminine 511/521 exposée au niveau de la lame basale de l'endothélium vasculaire enflammée (lors d'une inflammation causée par une infection par exemple) [4]. Ces interactions protéine-protéine citées ci-dessus entraînent le ralentissement du flux sanguin puis l'immobilisation des globules rouges drépanocytaires causant ainsi l'occlusion complète du vaisseau (vaso-occlusion) [3]. Dans cette thèse, nous nous sommes intéressés à la protéine Lutheran/BCAM pour laquelle des études *in vitro* montrent une forte implication dans la liaison des globules rouge drepanocytaire à la Laminine 511/521 [4, 5].

Les protéines Luthéran (Lu) et Basal Cell Adhesion Molecule (BCAM) font partie de la superfamille des immunoglobulines. Elles sont codées par le même gène et sont obtenues par épissage alternatif de l'ARNm correspondant. Lu et BCAM sont donc deux isoformes glycosylées respectivement de 85 kDa et 78 kDa qui diffèrent par la longueur de leur domaine cytoplasmique C-terminal. Par conséquent, ces glycoprotéines possèdent des domaines extracellulaires identiques composés d'un premier domaine N-terminal Ig de

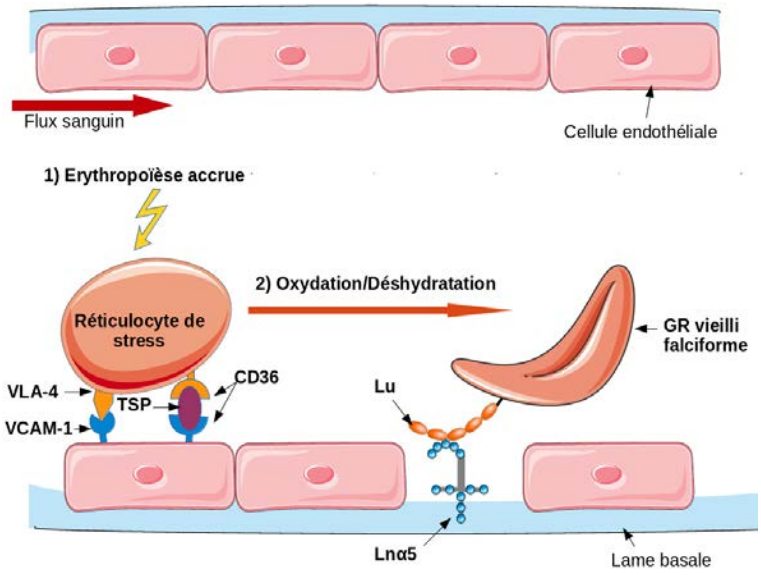


Figure 1: Mécanisme de la vaso-occlusion chez les drépanocytaires. Des réticulocytes de stress sont produits lors d'une érythropoïèse accrue. Ces réticulocytes de stress (globules rouges immatures) expriment les intégrines VLA-4 et CD36 qui se lient respectivement aux protéines VCAM-1 et CD36 exprimées par l'endothélium enflammé (par une infection par exemple). Les deux protéines CD36 se lient entre elles grâce à un pont de thrombospondine (TSP) libérée par activation de plaquettes. Cette première étape initie le ralentissement du flux sanguin, ce qui a pour conséquence une hypoxie puisque l'oxygène est moins bien transporté dans l'organisme. Se produit alors une désoxygénation de l'hémoglobine S : un processus d'oxydation et de déshydratation entraîne la falciformation du Globule rouge (GR) vieilli. Ce GR vieilli surexprime la protéine Luthéran/BCAM (Lu) qui se lie fortement à la protéine Laminine 511/521 (Lnα5) exposée au niveau de lame basale de l'endothélium enflammé. L'ensemble de ces phénomènes provoque des crises vaso-occlusives très douloureuses chez les drépanocytaires [2] (inspiré de l'ouvrage "La drépanocytose" [3]).

type V (domaine D1), d'un second domaine N-terminal Ig de type C1-set (domaine D2)[1] et de trois domaines N-terminal Ig de type C2-set (domaines D3 à D5) [7]. Dans cette étude, nous nous intéressons uniquement à la partie extracellulaire de Lu et de BCAM par laquelle elles se lient spécifiquement à la Laminine 511/521 [4, 5]. Puisque les protéines Lu et BCAM possèdent exactement les mêmes domaines extracellulaires, nous ne différencierons plus ces deux protéines dans la suite du manuscrit : elles seront toutes les deux désignées par Lu.

Les Laminines (Ln) sont des composants protéiques majeurs de la lame basale des cellules endothéliales et épithéliales. Ces protéines sont des hétérotrimères formés de trois sous-unités : α, β et γ qui forment 16 isoformes de Ln. Selon des études *in vitro* et *in vivo*, la protéine Lu ne va se lier qu'aux isoformes de la Laminine qui contiennent la chaîne α5 telles que Ln-511 et Ln-521 [4, 5, 8–10]. Ces isoformes seront appelées Lnα5 dans la suite du manuscrit. Lnα5 est composée de cinq domaines globulaires LG1-LG5 à son extrémité C-terminale. En plus d'interagir avec Lu, cette protéine peut aussi interagir spécifiquement avec quelques intégrines, α-dystroglycan et syndecan-4 [11]. Bien que les structures cristallographiques de plusieurs isoformes de Ln soient disponibles dans la PDB, celles contenant la chaîne α5 n'ont pas encore été résolues à ce jour.

Une caractéristique des globules rouges drépanocytaires est la surexpression de la protéine Lu à la surface de leur membrane. Cependant il a été montré que la présence de Lu en surface ne suffisait pas pour causer l'adhérence de ces cellules à l'endothélium vasculaire. En effet, afin de se lier à Lnα5, la protéine Lu doit être activée par phosphorylation du résidu Ser621 (grâce à une voie PKA) qui se trouve dans la partie cytoplasmique de Lu [12, 13]. Actuellement, il existe un seul traitement médicamenteux qui permet de réduire le nombre de crises vaso-occlusives chez les drépanocytaires : il s'agit du traitement à l'hydroxyurée [14]. Approuvée par la FDA (Food and Drug Administra-

[1] Ce domaine D2 a été classifié comme étant un domaine de type V dans Uniprot mais l'analyse structurale de ce domaine montre qu'il ressemble plûtot à un domaine Ig de type C1-set [6].

tion) des États-Unis, l'hydroxyurée empêche l'activation de la protéine Lu en inhibant la phosphorylation de la partie cytoplasmique de celle-ci. Bien que ce traitement réduit considérablement les crises vaso-occlusives chez les drépanocytaires, il peut avoir plusieurs effets secondaires tels qu'une hyperpigmentation de la peau et des nausées ainsi que des risques de myélosuppression et de neutropénie chez les enfants [14]. Compte tenu de ces effets secondaires, nous avons décidé de rechercher des petites molécules capables d'inhiber l'interaction Lu-Ln$\alpha 5$ par encombrement stérique, c'est à dire des molécules qui, en se liant à l'une des deux protéines, vont empêcher la liaison de l'autre protéine. Ce type de molécule est plus communément appelé inhibiteur d'interaction protéine-protéine (PPII). Dans les paragraphes qui suivent, nous discuterons des difficultés liées à la recherche de PPII, puis nous présenterons l'interaction Lu-Ln$\alpha 5$ et les étapes que nous avons suivies pour rechercher des inhibiteurs de cette interaction.

En raison de leur implication dans de nombreuses maladies, les interactions protéine-protéine (PPI) sont des cibles importantes dans la recherche de molécules à visée thérapeutique. La recherche de PPII est un travail difficile pour plusieurs raisons [15]. Premièrement, les protéines interagissent entre elles sur des surfaces larges (aux alentours de 1 500 à 3 000 Å^2) et relativement planes. Ainsi, contrairement aux enzymes qui possèdent un site actif profond et bien défini, la surface d'interaction protéine-protéine ne présente généralement pas de poches profondes dans lesquelles de petites molécules peuvent se lier. Deuxièmement, la surface d'interaction de deux protéines étant très large, il est difficile de définir un site de liaison précis sur lequel la recherche de PPII potentiels devrait se concentrer. Enfin, ces molécules inhibitrices doivent non seulement se lier sur une surface plane, mais doivent aussi résister à la liaison du partenaire biologique.

La protéine cible, c'est-à-dire la protéine sur laquelle les molécules inhibitrices se lient, fait une multitude d'interactions hydrophobes, de contacts de van der Waals ainsi que des interactions électrostatiques et des liaisons hydrogène avec son partenaire biologique [16]. Bien que les interactions entre deux protéines soient nombreuses et

s'étalent sur une large surface, seuls quelques résidus ou clusters de résidus (« hot spots ») concentrent la majorité de l'énergie d'interaction et sont déterminants pour la reconnaissance et l'affinité de liaison [17, 18]. Ces « hot-spots », détectées par des expériences de mutagenèse, constituent des régions de l'interface protéine-protéine qui peuvent être ciblées par les potentiels PPII [18].

Comme expliqué précédemment, cette étude consiste à trouver des PPII capables d'inhiber l'interaction Lu-Lnα5. En ciblant cette interaction, nous voulons supprimer ou réduire l'adhésion cellulaire responsable des crises vaso-occlusives observées chez les drépanocytaires. La recherche de PPII ciblant l'interaction Lu-Lnα5 est aussi d'un intérêt majeur dans la recherche de médicaments contre le cancer depuis qu'il a été découvert que l'interaction entre ces deux protéines favorisait la migration des cellules tumorales, notamment dans les cancers de la peau, des ovaires et du pancréas [19–21]. Cette étude est notamment originale puisque le seul résultat concluant indiquant l'inhibition par gêne stérique de l'interaction entre Lu et la Laminine a été obtenu à l'aide d'un anticorps se liant au deuxième domaine N-terminal de Lu [22]. Il est cependant très difficile et coûteux de développer une thérapie à partir d'anticorps ; la solution la mieux adaptée doit provenir d'une molécule de faible masse.

Afin de rechercher des inhibiteurs de l'interaction Lu-Lnα5, il est nécessaire d'analyser l'interface de liaison de ces deux protéines. L'interaction de ces protéines se fait au niveau des domaines LG1, LG2 et LG3 de Lnα5 [23] et des domaines D2 et D3 de Lu [6] (Figure 2). La structure de Lnα5 n'est pas connue. En revanche, celle des domaines D1 et D2 de Lu a été résolue par cristallographie aux rayons X et le domaine D3 a été obtenu avec la méthode SAXS (Small Angle X-ray Scattering) par Mankelow et collaborateurs en 2007 [6]. Selon des expériences de mutagenèse, le site de liaison de Lnα5 est notamment constitué de résidus chargés négativement sur les domaines D2 et D3 de Lu (Figure 2) [6]. Les résidus qui sont impliqués dans la liaison de Lnα5 sont montrés en jaune, orange et rouge dans la Figure 2. Alors que ces résidus se concentrent dans l'extrémité

haute du domaine D3 de Lu avec notamment les résidus Asp312 (D312, en rouge), Glu309 (E309, en orange) et Asp310 (D310, en orange) qui sont fortement impliqués dans la liaison de Lnα5, les résidus représentés en jaune (moyennement impliqués dans la liaison) sur la Figure 2 et qui se trouvent sur le domaine D2 s'étalent sur les deux faces de ce domaine avec un cluster de résidus sur l'extrémité basse (E132, D133, D198, D199) et un résidu sur l'extrémité haute (E180).

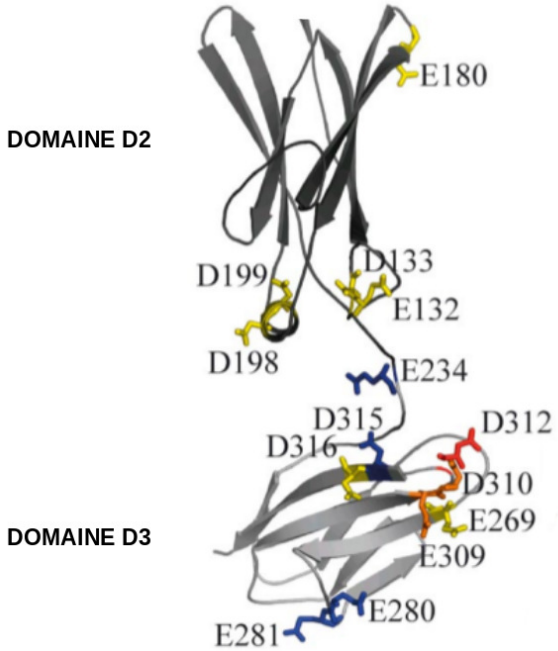


Figure 2: Structure des domaines D2 et D3 de Lu respectivement résolue par cristallographie aux rayons X et par la méthode SAXS. Les résidus qui apparaissent sur cette figure sont ceux qui ont été mutés en Ala afin de voir leur implication dans la liaison de Lnα5. Les résidus en bleu sont ceux pour lesquels les mutations n'ont pas changé l'affinité de Lu pour Lnα5 ; ceux en jaune, orange et rouge sont ceux qui ont contribuer à diminuer l'affinité de Lu pour Lnα5 de façon légère, marquée et importante [6] (Code PDB 2PET).

La jonction flexible qui lie les domaines D2 et D3 est longue d'environ huit acides

aminés et joue aussi un rôle important dans la liaison de Lnα5. Bien que les expériences de mutagenèses dirigées sur cette jonction n'aient pas affecté l'affinité de Lu pour Lnα5, il a été montré que la liaison entre ces deux protéines était abolie lorsque cette jonction était écourtée de trois résidus. Ce résultat implique que la liaison de Lnα5 sur Lu nécessite une réorientation du domaine D3 par rapport au domaine D2 qui est permise par la jonction qui lie ces deux domaines. C'est la flexibilité importante de cette jonction qui n'a pas permis de résoudre la structure du domaine D3 de façon précise avec la méthode de cristallographie aux rayons X [6].

Puisque seul le domaine D2 de Lu est à la fois impliqué dans la liaison de Lnα5 et résolue avec une bonne résolution de 1.7 Å (Code PDB 2PET), nous avons décidé de rechercher des PPII dirigés contre ce domaine uniquement. Pour cela, nous avons utilisé les techniques de modélisation et de simulation moléculaires afin de réaliser un criblage virtuel à partir d'une large bibliothèque de composés, chimiothèque, et d'en extraire ceux qui auront les propriétés désirées d'inhibition. Les techniques de modélisation et de simulation moléculaires ont été largement éprouvées dans le domaine de la recherche biomédicale et sont aujourd'hui devenues, pour la recherche de molécules à visée thérapeutique, des techniques incontournables parce qu'elles permettent à la fois d'accélérer les processus de découverte de ces molécules et, également, de minimiser leur coût de développement. Afin de réaliser un tel criblage virtuel, il est nécessaire de connaître le site de liaison de Lnα5 sur le domaine D2 de Lu.

Pour la recherche de PPII ciblant le second domaine de Lu, nous avons été confrontés à deux difficultés majeures. Premièrement, la surface d'interaction définie par les expériences de mutagenèse sur le domaine D2 de Lu est très étendue puisqu'elle couvre les deux faces du domaine, c'est-à-dire la face sur laquelle se trouvent les résidus Glu132, Asp133 et Glu180 et celle sur laquelle se trouvent les résidus Asp198 et Asp199 montrés sur la Figure 2 (p. 14). Cette surface étant trop étendue, il a été nécessaire de définir au préalable une zone plus restreinte du domaine D2 sur laquelle réaliser le criblage.

Deuxièmement, nous n'avions pas connaissance des autres partenaires de Lu malgré une recherche approfondie de ces partenaires dans la littérature. En raison de ces difficultés, il nous était impossible d'établir un protocole de criblage directement sur Lu.

Afin de contourner ces difficultés, nous avons établi une procédure de criblage qui s'articule en plusieurs étapes et qui sont présentées dans la Figure 3 : recherche d'un site de liaison sur le domaine D2 de Lu d'une part, et recherche d'un système similaire à Lu sur lequel établir un protocole de scoring d'autre part (a) ; recherche et validation d'un protocole de scoring sur le système similaire à Lu (étapes (b), (c) et (d)) ; application du protocole validé sur le domaine D2 de Lu (e). La recherche d'un système similaire a deux objectifs. Le premier est de permettre la déduction d'un site de liaison probable sur Lu à partir de la comparaison entre ce système similaire et le domaine D2 de Lu. Dans ce premier objectif, le système similaire doit être complexé à une autre molécule (protéine ou petite molécule), sans nécessairement de constante d'affinité associée. Le second objectif est l'élaboration d'un protocole de criblage à partir de ce système similaire comprenant une protéine (homologue ou présentant seulement une ressemblance locale, analogue) et des constantes d'affinité connues.

L'étape (a) consiste à identifier sur le domaine D2 de Lu un site de liaison plus restreint et adapté à la taille des molécules organiques trouvées dans la chimiothèque ZINC [24]. Elle consiste également à rechercher une protéine homologue ou analogue (dénomée système similaire dans la Figure 3) possédant des ligands d'affinités connues dans le but d'élaborer un protocole d'évaluation des énergies d'interaction entre la protéine et son ligand potentiel, également appelée scoring. L'élaboration ainsi que la validation de ce protocole de scoring sur un système similaire à Lu se fait en trois étapes : (b) le docking, c'est-à-dire le calcul des géométries des complexes entre une protéine et des composés ; (c) la relaxation des géométries des complexes afin de permettre un positionnement optimal du composé sur la protéine ; (d) l'évaluation des énergies d'interaction (scoring) entre la protéine similaire à Lu et les composés afin d'en déduire

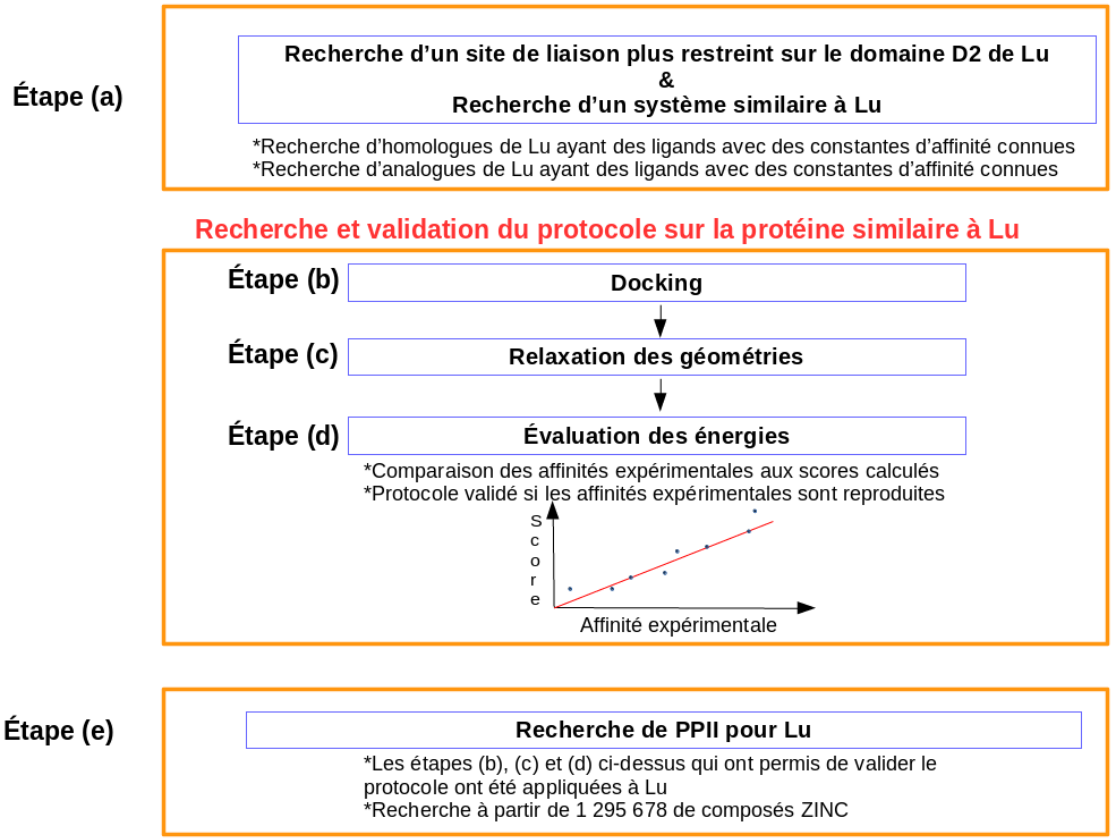


Figure 3: Procédure de criblage pour l'identification de composés ayant une forte probabilité de liaison à Lu. Les étapes (b), (c) et (d) constituent l'étape de recherche et de validation du protocole scoring sur la protéine similaire à Lu identifiée à l'étape (a). Lorsque le protocole est validé, celui-ci est appliqué au domaine D2 de Lu afin de trouver des inhibiteurs potentiels de l'interaction Lu-Lnα5 à partir de 1 295 678 composés issus de la chimiothèque ZINC.

ceux de meilleures affinités. Le docking dans l'étape (b) est nécessaire si, pour le système similaire, la géométrie expérimentale de chacun des complexes impliquant la protéine homologue/analogue et son ligand n'est pas connue. Dans l'étape (e), les étapes (b), (c) et (d) qui ont permis de valider le protocole sur la protéine similaire à Lu sont appliquées à Lu.

Pour l'étape (a) (Figure 3), nous nous sommes proposés de trouver une protéine homologue ou analogue à Lu formant un complexe avec une autre protéine ou un composé de faible masse et dont la structure est disponible dans la Protein Data Bank (PDB) [25]. Contrairement à une protéine homologue qui serait caractérisée par un pourcentage d'identité de séquences d'au moins 30 % sur les cinq domaines ou, à défaut, uniquement sur le domaine D2 de Lu, une protéine analogue ferait apparaître une similarité avec Lu plus localisée structuralement sur une région restreinte et qui définirait un site de liaison pour cet analogue. La connaissance du site d'interaction pour cet homologue/analogue nous renseignera sur le site d'interaction probable de Lu. Idéalement, si le partenaire de liaison de type protéique est lui-même un homologue à la Laminine, alors le site d'interaction de la Laminine sur Lu pourra également être mieux défini. D'autre part, pour l'étape (b), un protocole de scoring fiable pourra également être établi à partir d'une protéine homologue ou analogue à condition que cette dernière ait des ligands possédant des constantes d'affinité connues. Cette affirmation repose sur l'observation selon laquelle les performances d'un protocole de scoring dépendent de la protéine étudiée : les performances d'un même protocole appliqué à deux sites d'interaction (de deux protéines) qui diffèrent par leur topologie et/ou par leurs types de résidus ne seront pas nécessairement les mêmes, ce protocole étant performant par exemple dans un seul des deux cas. En revanche, les performances d'un protocole seront *a priori* les mêmes sur des protéines homologues ou analogues. Bien sûr, les performances attendues seront d'autant meilleures que le degré de similarité entre le site d'interaction de cet homologue/analogue et celui de Lu est élevé. Nous avons donc recherché dans cette étape

(a) une protéine homologue ou analogue possédant des ligands de constantes d'affinité connues et qui s'étendent du nanomolaire au micromolaire (ceci permettant d'assurer une hétérogénéité dans les valeurs de ces constantes).

Afin de rechercher un système similaire à Lu, nous avons réalisé des alignements de séquences, des alignements globaux de structures et des alignements locaux de structures. Puisque ces recherches n'ont pas été concluantes (pas d'homologues ou d'analogues de Lu ayant des ligands avec des constantes d'affinité connues ou permettant de définir un site de liaison sur le domaine D2 de Lu), nous avons décidé de construire le protocole de scoring sur la protéine CD80 qui présente, sur son domaine V, un site de liaison similaire à celui de Lu ainsi que plusieurs dizaines de ligands avec des constantes d'affinité connues. Cette protéine a été trouvée dans la banque de molécules TIMBAL [26]. La protéine CD80 satisfaisait les trois conditions suivantes: (i) une structure cristallographique de ce domaine existe dans la PDB ; (ii) une série de ligands avec des constantes de liaison connues sont disponibles ; (iii) son site de liaison partage certaines similitudes structurales avec celle prédite de Lu. Nous avons défini u on site de liaison le domaine D2 de Lu en utilisant une prédiction basée sur un consensus entre les résultats de plusieurs serveurs Web de prédiction.

La Figure 4 montre le canevas à partir duquel la méthodologie pour CD80 et Lu a été construite. Pour CD80, l'étape de docking consiste à trouver la pose des ligands de cette protéine, ligands dont les affinités sont par ailleurs connues. Pour Lu, l'étape de docking consiste en un criblage virtuel à partir de la chimiothèque ZINC. À l'étape de scoring, plusieurs méthodes de scoring ont été utilisées et celle qui a donné les meilleurs résultats dans le cas de CD80 a aussi été utilisée dans la recherche de ligands de Lu.

Le protocole de criblage associé aux étapes (b) à (d) (soit l'étape d'élaboration et de validation d'un protocole de scoring sur le système similaire à Lu) et défini dans la Figure 3 (p. 17) a donc été élaboré grâce à la protéine CD80 trouvée à l'étape (a). Ce protocole a été élaboré et perfectionné jusqu'à ce qu'il permette de reproduire les constantes

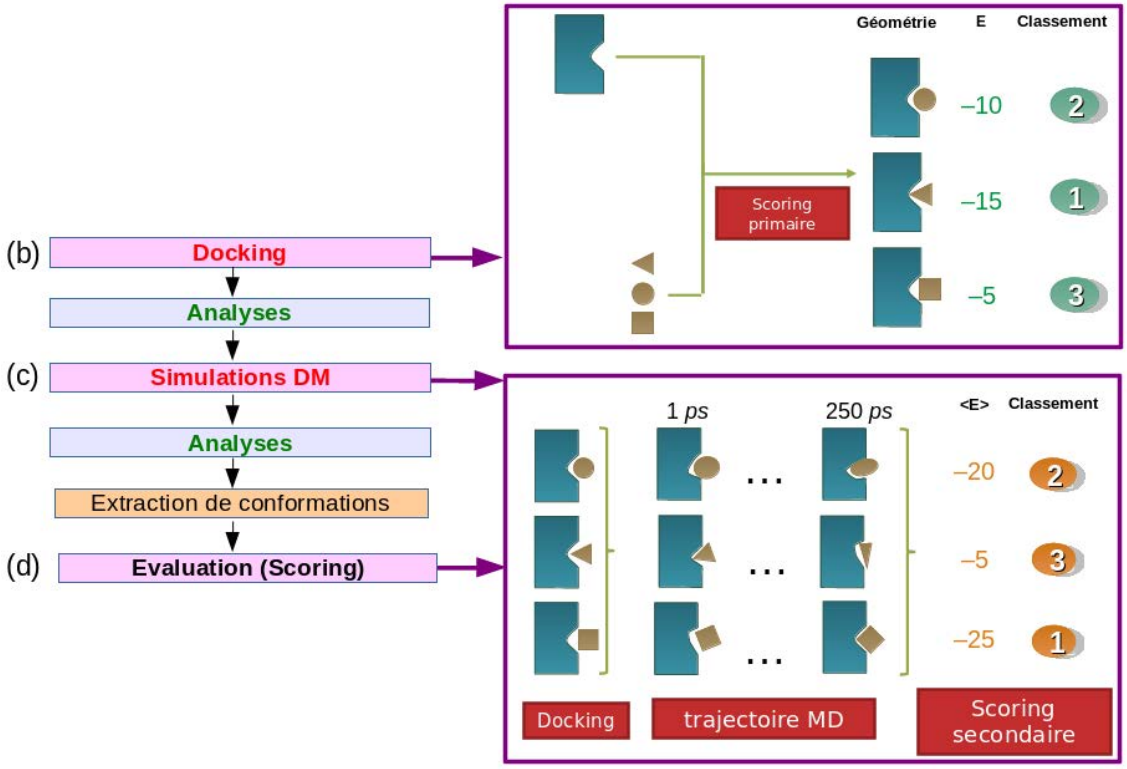


Figure 4: Canevas à partir duquel la méthodologie pour CD80 et Lu a été établie. Ce canevas contient les étapes (b) de docking, (c) de simulation DM et (d) de scoring (évaluation d'énergie) entre lesquelles nous avons réalisé des analyses.

d'affinité expérimentales pour les ligands de CD80 (les détails de ce protocole sont donnés dans la section 3.2). L'étape (b) de docking a été réalisée avec DOCK6, puis les complexes obtenus ont été relaxés à l'étape (c) avec des simulations de dynamique moléculaire (DM). À cette étape, nous avons testé des simulations DM sans eau, en présence d'eau explicite grâce à la méthode de dynamique moléculaire aux limites stochastiques (SBMD) et en présence d'eau implicite grâce à la méthode de dynamique Langevin (LD). Enfin, à l'étape (d), nous avons calculé les énergies des complexation (scores) issues des trajectoires obtenues à l'étape (c) grâce à différentes méthodes de calculs : des fonctions de scoring empiriques (XSCORE [27], Cyscore [28] et ID-Score [29]), knowledged-based (ITScore [30] et DrugScore [31]) et basées sur un champ de force (MM-PBSA [32]). Nous avons aussi réalisé des calculs de chimie quantique (méthodes semi-empiriques et Fragment Molecular Orbital, FMO [33]) afin d'évaluer les énergies des complexes.

Une fois le protocole de scoring validé, c'est à dire lorsqu'il est capable de reproduire les affinités expérimentales des quelques ligands choisis pour CD80, nous avons appliqué les étapes (b), (c) et (d) à Lu (étape (e) de la Figure 3, p. 17). Pour Lu, nous avons utilisé 1 295 678 composés de la chimiothèque ZINC [24] contrairement à CD80 pour lequel nous n'avons utilisé qu'un petit nombre de ligands. Dans l'étape (e), nous avons réalisé le docking avec les programmes DOCK6 [34] et rDOCK [35], suivie d'une simulation DM en présence d'eau explicite (méthode SMBD) et d'un calcul des énergies des complexes avec la méthode XSCORE. Les méthodes utilisées à l'étape (e) sont les méthodes qui ont été validées dans l'étape de validation du protocole de scoring sur CD80.

Suite à ce criblage virtuel sur le second domaine de Lu, trois molécules ont été testées par deux membres de l'équipe 1 expérimentale de notre UMR, Wassim El Nemer et Sylvie Cochet. Spécialiste de Lu, Wassim El Nemer a testé l'efficacité de ces molécules en utilisant la plateforme Venaflux qui permet de mimer le flux sanguin et ainsi de mesurer l'adhérence des globules rouge drépanocytaires à Lnα5 préalablement fixée. De cette étude expérimentale, deux molécules se sont révélées actives et font l'objet de deux

dépôts de brevet à l'heure actuelle auprès du Bureau Européen des Brevets.

Ce travail de thèse sera présenté comme suit dans le manuscrit : la méthodologie utilisée (chapitre 1) ; la procédure de recherche de système similaire à Lu (chapitre 2) et l'élaboration d'un protocole de scoring sur ce système ainsi que sa validation (chapitres 3 et 4) ; la recherche de PPII potentiel sur Lu à partir de composés issus de la chimiothèque ZINC (chapitre 5).

Chapitre 1

Méthodologie

Dans ce chapitre, nous présenterons d'abord la méthodologie utilisée dans cette thèse : la recherche de protéines homologues ou analogues à Lu (section 1.1) ; la préparation des protéines CD80 et Lu (section 1.2) ; l'élaboration d'un protocole de scoring à l'aide de la protéine CD80 (section 1.3) ; le criblage virtuel de la chimiothèque ZINC appliqué à Lu (section 1.4). Puis nous présenterons le principe du champ de force (section 1.5) avant de présenter les méthodes de docking moléculaire, de calculs d'énergies avec des fonctions de scoring secondaire (section 1.7) et la méthode de simulation de dynamique moléculaire (section 1.8).

1.1 Recherche de protéines homologues ou analogues à Lu

.

Dans la première étape de notre procédure de criblage (étape (a), Figure 3 à la p. 17), nous avons, dans un premier temps, recherché dans la littérature des ligands avec lesquels Lu était co-cristallisé. Puisque cette recherche a été sans succès, nous avons util-

isé quatre approches différentes pour identifier un site de liaison sur le domaine D2 de Lu et/ou identifier un homologue ou un analogue de Lu ayant des ligands avec des structures et des constantes d'affinité connues à partir duquel nous pouvons construire le protocole de scoring. Tout d'abord, une recherche par BLAST [36], à partir de la séquence entière de Lu ou de son domaine D2, a été réalisée pour trouver une protéine homologue en complexe avec un partenaire (protéine ou composé de faible masse) et pour laquelle une structure est disponible dans la PDB [25]. En second lieu, pour les protéines trouvées dans cette recherche qui avaient des pourcentages d'identité de séquence faibles avec Lu (moins du seuil de 30% par rapport à la séquence de D2), leur structure a été alignée sur celle du domaine D2 de Lu afin de vérifier si une relation d'homologie distante était possible. En troisième lieu, une autre stratégie pour trouver des homologues distants a consisté à utiliser des outils tels que PDBeFOLD [37], VAST [38], VAST+ [39], DALI [40] et l'outil de recherche de similarité 3D de la PDB [25]. En dernier lieu, une stratégie a été utilisée pour déterminer des protéines analogues uniquement : des alignements structuraux, non plus globaux mais locaux, avec Lu ont été utilisés dans ce but grâce à l'utilisation de POSSUM [41], GIRAF [42] et Phyre2 [43]. Bien que des analogues furent trouvés présentant une similarité localisée sur une certaine portion de la surface de D2, aucun n'a pu être trouvé co-cristallisé avec une autre protéine ou avec une petite molécule.

Puisque les stratégies citées ci-dessus n'ont pas permis d'identifier un site de liaison sur Lu, nous avons utilisé les résultats obtenus avec les expériences de mutagenèse [6] et une prédiction basée sur un consensus entre les résultats de plusieurs serveurs Web afin de prédire un site de liaison sur le domaine D2 de Lu. Parmi les méthodes de prédiction de site de liaison, on peut distinguer la méthode géométrique et la méthode énergétique. La méthode géométrique consiste à rechercher des cavités à la surface des protéines : ce sont les cavités les plus profondes qui sont généralement prédites comme sites de liaison. La méthode énergétique consiste à calculer les interactions entre une sonde (van

der Waals, électrostatique, hydrophobe, etc.) et la protéine afin de détecter les zones énergétiquement favorables à la liaison d'un ligand [44]. Le domaine D2 de Lu est un domaine de type C1-set, qui se caractérise par deux feuillets β antiparallèles définissant chacun une surface pseudo-plane (Figure 2, p. 14). Ces feuillets β organisés en sandwich définissent donc deux surfaces planes avec peu de cavités. La méthode la plus adaptée à la recherche de site de liaison sur ce domaine D2 est alors la méthode énergétique. Nous avons testé trois serveurs de prédiction de site de liaison qui utilisent la méthode énergétique : Q-siteFinder [45], FTSite [46] et SiteHound [47]. Ces serveurs Web utilisent différents types de sondes (carbone aromatique, phosphate oxygène, etc.) permettant de détecter les zones énergétiquement favorables à la liaison d'un ligand. Bien que le domaine D2 présente peu de cavités, celui-ci présente tout de même plusieurs micro-cavités qui sont définies par des chaînes latérales de résidus adjacents sur les surfaces. Afin de voir s'il était possible de prédire un site de liaison à partir des micro-cavités présentes à la surface du domaine D2 de Lu, nous avons testé les serveurs de prédiction eFindsite [48] et Surfnet [49] qui utilisent la méthode géométrique.

La recherche de protéines homologues ou analogues de Lu n'avait pas pour seul but de définir un site de liaison sur Lu. En effet, l'idée était d'établir un protocole de scoring sur une protéine semblable à Lu et qui possède plusieurs ligands avec des constantes d'affinité connues. Nous avons sélectionné la protéine CD80 de la base de données TIMBAL [26] qui liste des inhibiteurs d'interactions protéine-protéine. Ces inhibiteurs se lient au domaine V de CD80 dans le but d'inhiber les interactions CD80-CD28 et CD80-CTLA-4. Le site de liaison prédit (section 1.3) sur le domaine V de CD80 est similaire à celui prédit sur le domaine D2 de Lu. De plus, plusieurs dizaines de ligands de constantes d'affinité connues sont répertoriés pour ce domaine V de CD80 [50]. Le site de liaison de CD80 ainsi que sa comparaison avec celui prédit pour Lu sont présentés dans le chapitre 2. Dans le but de reproduire les affinités expérimentales d'un ensemble de ligands de CD80, nous avons testé plusieurs combinaisons de protocoles de docking, de simulations

de dynamique moléculaire (DM) ainsi que différentes méthodes de calcul d'énergie de liaison des ligands (respectivement les étapes (b) à (d) de la Figure 3, p. 17). Ces calculs d'énergie de liaison ont été réalisés à partir de calculs empiriques avec différentes fonctions de scoring et des calculs de chimie quantique avec la méthode FMO (Fragment Molecular Orbital) et la méthode PM6. Les détails relatifs aux étapes (b) à (d) (Figure 3, p. 17) utilisées pour établir le protocole de scoring sur CD80 d'une part et pour réaliser le criblage sur Lu d'autre part sont donnés dans les sections 1.3 et 1.4.

1.2 Préparation des protéines

Les structures cristallographiques de Lu et de CD80 ont été obtenues à partir de la PDB [25] (code PDB 2PET pour Lu, et 1I8L et 1DR9 pour CD80). Seuls les deux domaines N-terminaux de Lu et de CD80 sont disponibles dans ces structures respectives. PROCHECK [51] et WHATIF [52] sont des outils qui permettent un examen détaillé de la stéréochimie de la structure d'une protéine et ont été appliqués aux deux structures disponibles de CD80 qui ont toutes les deux une faible résolution de 3 Å. Étant donné que les problèmes stéréochimiques et de packing étaient similaires pour les deux structures, nous avons décidé d'optimiser la géométrie de la structure 1DR9. Ce choix a été motivé par la comparaison des facteurs de structure (B-factors[1]) des deux structures cristallographiques : les valeurs les plus faibles sont observées pour 1DR9 (41 ± 11 Å^2 pour 1DR9 et 52 ± 25 Å^2 pour 1I8L).

Dans le but d'améliorer la stéréochimie globale du domaine cible V de CD80 avant l'étape de docking (étape (b), Figure 3, p. 17), une procédure de minimisation à l'aide du programme CHARMM [53] a été réalisée après l'immersion de l'ensemble du domaine V dans une couche de molécules d'eau : seules les molécules d'eau ont d'abord

[1]Les B-factors renseignent la précision avec laquelle les positions atomiques sont connues et/ou sur la mobilité intrinsèque des atomes.

été minimisées par 1 000 pas de minimisation suivie de 5 000 autres pas pour la minimisation de l'ensemble du système. Après la minimisation, les problèmes stéréochimiques (en lien avec les liaisons, les angles de valence et de torsion du squelette et des chaînes latérales) ainsi que le principal problème de packing qui concernait le résidu Met43 du site d'interaction ont été résolus ou amoindris. La structure minimisée finale de CD80 a été utilisée dans l'étape suivante de docking.

Aucune minimisation de la structure cristallographique de Lu n'a été considérée étant donné (i) sa haute résolution (1.7 Å) et (ii) l'absence de problèmes stéréochimiques ou de packing pour le domaine ciblé (résidus 116 à 231) d'après les outils de vérification PROCHECK et WHATIF.

1.3 Élaboration d'un protocole de scoring à l'aide de la protéine CD80

Pour valider le protocole de criblage, nous avons d'abord appliqué les étapes (b) à (d) au domaine V de CD80 telles qu'elles ont été définies dans l'introduction (Figure 3, p. 17). Cependant, aucun des composés qui se lient à ce domaine V n'a été co-cristallisé avec CD80 et les seules informations que nous avons sur leur région de liaison sur le domaine V de cette protéine sont des résultats issus des expériences de mutagenèses qui montrent que la mutation des résidus Asn48 (N48K) ou Trp50 (W84A) diminue fortement l'affinité de ces composés pour CD80 [54]. L'effet de la mutation du résidu Asn48 a été testé en même temps que la mutation des résidus Met47 et Ile49. Bien que l'affinité de CD80 pour les composés soit diminuée lorsque cette protéine est mutée au niveau des résidus Asn48, Met47 et Ile49, nous avons constaté que les chaînes latérales de ces deux derniers résidus sont enfouies à l'intérieur de la protéine. Les chaînes latérales des résidus Met47 et Ile49 ne sont donc pas susceptibles d'interagir directement avec les composés.

Afin, de délimiter précisément le site de liaison des inhibiteurs sur CD80, nous avons utilisé les serveurs Web de prédiction de sites de liaison SiteHound [47], eFindSite [48] et Surfnet [49], conjointement avec les deux résidus importants, Asn48 et Trp50, pour la liaison des inhibiteurs selon les expériences de mutagenèses [50, 54, 55]. Dix-sept composés ont été choisis avec une large gamme de valeurs IC50, allant de 4 nM à 1 300 nM. Dans l'étape (b) (Figure 3, p. 17), les calculs de docking ont été réalisés avec DOCK6 [34]. Dans cette étape de docking, les ligands sont traités comme flexibles alors que la protéine est traitée comme rigide. Au cours du docking, un ou plusieurs fragments d'ancrage, tel que défini et exposé en détail dans la sous-section 1.6 (p. 42), sont positionés dans 100 orientations différentes sur le site de liaison préalablement identifié. Le reste du ligand est ensuite construit liaison après liaison, avec une rotation autour de chaque liaison ajoutée par incréments de 10° pour identifier l'orientation la plus favorable basée sur l'énergie d'interaction (van der Waals et les contributions électrostatiques seulement) entre la protéine et le composé en cours de construction [34]. L'exposant pour la contribution répulsive de l'énergie de Lennard-Jones a été pris à neuf. Le facteur diélectrique a été choisi à quatre sans le calcul des interactions électrostatiques afin de mimer un effet d'écran dû à la solvatation.

Les énergies d'interaction protéine-ligand calculées uniquement avec les contributions de van der Waals et électrostatiques conduisent à un classement initial des 17 composés étudiés pour CD80 (scoring primaire après l'étape de docking (b), Figure 3, p. 17 et Figure 4, p. 20). Cependant, ce classement manque généralement de précision puisque l'effet du solvant est négligé ou imprécis et que la protéine est traitée comme rigide. C'est pourquoi les énergies d'interaction protéine-ligand ont été réévaluées avec les fonctions de scoring secondaire MM-PBSA [32], XSCORE [27], DrugScore [31], Cyscore [28], ITScore [30] et ID-Score [29] afin de les comparer aux 17 valeurs IC50 expérimentales. Avant cette réévaluation de l'énergie réalisée à l'étape (d), ces composés ont été soumis à des simulations DM à l'étape (c) pour tenir compte de la flexibilité des pro-

téines et de l'influence du solvant sur la liaison des ligands (Figures 3, p. 17 et 4, p. 20). Dans cette étape de relaxation, la protéine et le composé peuvent changer leur structure de manière à renforcer ou à améliorer leur interaction et leur complémentarité. Par exemple, un donneur de liaison hydrogène de la protéine peut se réorienter de façon à faire face à un accepteur de liaison hydrogène du composé. Cette étape permet donc un raffinement général de la structure pour la procédure de scoring secondaire suivante, l'étape (d). Dans cette étape de relaxation des complexes protéine-ligand, le champ de force CHARMM36 a été utilisé pour la protéine et le champ de force CGENFF a été utilisé pour les ligands (combinaison de champs de force ci-après dénommée CHARMM36-CGENFF) [56]. Afin d'améliorer les contacts entre les ligands et CD80, certains paramètres d'angles de torsion et de valence ainsi que les charges des ligands ont été modifiés de manière à reproduire les données issues des calculs de chimie quantique réalisés avec Gaussian09 [57] et MOPAC2012 (version 15.052L) [58] (ces modifications seront détaillées dans le chapitre 4).

Trois représentations alternatives de l'environnement aqueux dans lequel les complexes sont simulés ont été testées pour les simulations DM : absence d'eau, représentation implicite et représentation explicite pour l'eau. Pour la représentation implicite du solvant, on a réalisé des simulations de dynamique moléculaire Langevin (LD) [59] qui permettent de modéliser les effets du solvant en termes de friction et de chocs intermoléculaires avec les solutés (protéine et ligand). Ces frictions et notamment les chocs entre les solutés et les molécules d'eau résultent de l'agitation thermique des molécules d'eau (l'eau est absente dans ce modèle mais les effets sont mimés). Pour la représentation explicite du solvant, des simulations DM aux limites stochastiques (SBMD) [60] ont été réalisées pour chaque complexe protéine-ligand préalablement immergé dans une sphère de molécules d'eau d'un rayon de 32 Å. Cette taille de sphère est suffisante pour former une large couche d'eau autour du site de liaison. Un potentiel a été ajouté sur la surface la plus externe de la sphère d'eau (soit à la limite entre les molécules d'eau et le

vide) afin d'éviter l'évaporation.

Toutes les simulations LD et SBMD ont été effectuées avec le programme CHARMM [53]. Les énergies des complexes protéine-ligand issus de l'étape de docking ont été minimisées avant de chauffer les complexes de 0 à 300 K suivant un incrément de 10 K. Les étapes de chauffage (60 ps) et d'équilibration (80 ps) ont été suivies d'une étape de production de 200 ps. Ce court temps de simulation est suffisant pour une relaxation locale des géométries en présence d'eau. Les temps de simulation ne pouvaient être plus grands puisque notre but était d'appliquer ce protocole sur le criblage d'une grande quantité de molécules issues de la banque ZINC afin de trouver des ligands pour Lu. Il est à noter que lors de simulations longues avec des paramètres de champ de force pour le ligand qui sont imparfaits (angle(s) de torsion ou charge(s) par exemple), le système peut dériver vers des géométries improbables. Dans les simulations, les interactions électrostatiques et de van der Waals ont été tronquées en utilisant un « switching » des forces entre 11 et 14 Å. De plus, nous avons utilisé un pas d'intégration de temps de 1 fs ainsi que l'algorithme SHAKE [61].

Les coordonnées des géométries ont été sauvegardées toutes les 250 fs, ce qui fait un total de 800 géométries parmi lesquels 100 géométries ont été sélectionnées de façon équidistante afin de permettre, à l'étape (d), les calculs d'énergie avec les fonctions de scoring secondaire (XSCORE, ITScore, ID-Score, DrugScore, Cyscore, MM-PBSA) et les méthodes de chimie quantique (Figure 3, p. 17).

Calculs de chimie quantique

Des calculs de chimie quantique ont été réalisés avec les programmes Gaussian09 [57] et MOPAC2012 [58] afin d'optimiser les structures des ligands uniquement ou de réaliser des scans optimisés sur des angles dièdres. Ces résultats nous ont permis d'améliorer les paramètres des ligands dans le champ de force CGENFF. Les scans optimisés ont permis de calculer l'énergie potentielle en fonction de la variation de certains angles dièdres. Les

détails des angles étudiés pour chacun des 17 composés sélectionnés se trouvent dans la sous-section 3.2.1. Les charges des ligands ont été réajustées dans le champ de force CGENFF suivant les charges calculées par chimie quantique en appliquant soit une règle de trois, soit en réalisant un ajustement relatif. Ces modications seront présentées dans la section 3.2.1 (p. 83). Les modifications des paramètres dans le champ de force CGENFF ont finalement permis d'améliorer les interactions entre les ligands et CD80 (résultats présentés dans le chapitre 4).

Afin de décrire plus précisément les interactions entre CD80 et les ligands, des calculs de chimie quantique ont été réalisés en utilisant la méthode *ab initio* FMO (Fragment Molecular Orbital) [33] et semi-empirique PM6-DH2X de MOPAC [58]. Le principe général de la méthode FMO est expliqué dans le chapitre 4 (p. 103 ; les détails peuvent être trouvés dans de nombreuses références [62–64]). Dans ces calculs réalisés avec le programme GAMESS-US [65], le modèle de solvatation PCM ainsi que le niveau de calcul MP2 (FMO-PCM-MP2) ont été utilisés dans le but de reproduire les affinités expérimentales. Les détails des différentes méthodes de calcul réalisées avec la méthode FMO-PCM-MP2 sont donnés dans la sous-section 4.2.1 (p. 108). Nous avons aussi réalisé des calculs FMO dans le vide avec la méthode RI-MP2 (plus rapide que MP2) à l'aide du programme PAICS [66, 67] afin d'évaluer uniquement la contribution énergétique de chaque résidu en contact avec les ligands dans la pose sélectionnée.

1.4 Criblage virtuel de la chimiothèque ZINC appliqué à Lu

Dans l'étape (e) de la Figure 3 (p. 17), un premier criblage a d'abord été réalisé sur 395 601 composés issus de la chimiothèque ZINC suivant le protocole que nous avons publié en 2016 (Figure 6 de la référence [1] ; publication disponible en Annexe). De

ce criblage nous avons obtenu deux PPII qui se sont révélés actifs selon les tests *in vitro* réalisés par Wassim El Nemer et Sylvie Cochet (les détails expérimentaux sont fournis en Annexe).

Dans le présent travail, un nouveau criblage à partir de 1 295 678 composés issus de la même chimiothèque a été réalisé et dans lequel la méthodologie a été modifiée afin de permettre, nous l'espérons, son perfectionnement. Dans cette section, et dans le reste du manuscrit, seuls les détails du protocole utilisé pour le deuxième criblage (criblage à partir de 1 295 678 composés) seront présentés.

Tous les composés (1 295 678) issus de ZINC sont issus de la catégorie « All clean » (qui sont disponibles dans le commerce). Dans cette catégorie, aucune molécule ne possède de groupement pouvant initier une toxicité [24]. À l'étape de docking (étape (b), Figure 4, p. 20), deux programmes de docking ont été utilisés : DOCK6 [34] et rDOCK [35]. Chacun de ces programmes a permis de complexer 1 295 678 composés sur Lu traité de façon rigide. Seuls les groupements terminaux OH et $NH_3{}^+$ des résidus qui se trouvent dans le site de liaison de Lu ont été rendus flexibles pendant l'étape de docking avec rDOCK uniquement.

Les différentes étapes du protocole appliqué à Lu pour le criblage de 1 295 678 composés sont montrées dans la Figure 1.1. Une seule conformation (pose) par ligand a été générée par DOCK6 (car DOCK6 ne permet de garder que la meilleure pose par ligand) alors que trois poses par ligand (pose 1, pose 2 et pose 3) ont été générées par rDOCK (étape (b), de la Figure 1.1). Pour un ligand donné dans rDOCK, les poses 1, 2 et 3 correspondent aux trois meilleures énergies de complexation calculées. Puis, seuls les 5 000 meilleurs composés classés par la fonction de scoring primaire de DOCK6 ont été retenus ; de même, les 5 000 meilleurs composés classés par la fonction de scoring primaire de rDOCK associés à chacune des poses 1, 2 et 3 ont été retenus (soit une liste de 5 000 composés pour chacune des poses). Nous avons utilisé le programme MATCH [68] afin de générer les paramètres CGENFF pour les ligands.

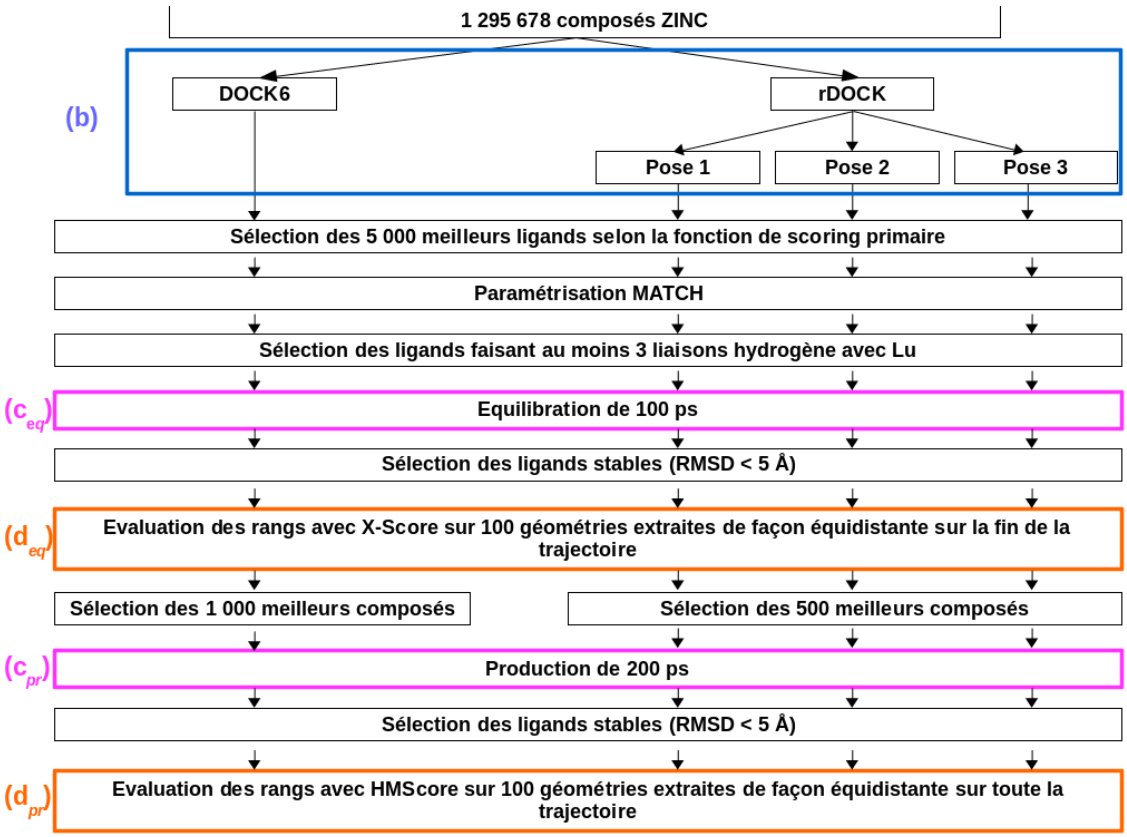


Figure 1.1: Méthodologie utilisée pour le criblage de 1 295 678 sur le domaine D2 de Lu. L'étape (b) encadrée en bleu correspond à l'étape de docking. Des étapes d'analyse ont permis de sélectionner des composés (ligands) après les étapes d'équilibration (c_{eq}) et de production (c_{pr}) ; (c_{eq}) et (c_{pr}) constituant des étapes de relaxation (étape (c) de la Figure 3, p. 17). Après chacune de ces étapes de relaxation, les énergies des géométries ont été évaluées avec la fonction de scoring X-Score (étape (d_{eq})) et HMScore (étape (d_{pr})) (étape (d) de la Figure 3, p. 17). Ces deux fonctions de scoring font partie de la méthode de calcul XSCORE (présentée dans la sous-section 1.7.1, p. 51). Les étapes (d_{eq}) et (d_{pr}) ont respectivement été réalisées sur 100 géométries extraites à la fin de la trajectoire d'équilibration et sur l'ensemble de la trajectoire de production. Pour sélectionner les ligands stables après l'étape d'équilibration (c_{eq}) et après l'étape de production (c_{pr}), les calculs de RMSD ont été réalisés respectivement sur la fin et sur l'ensemble des trajectoires correspondantes.

Avant l'étape d'équilibration (étape (c_{eq})), le nombre de liaisons hydrogène établies entre les ligands retenus et Lu a été comptabilisé avec CHARMM. Puis, des étapes d'analyse ont permis de sélectionner des ligands après l'étape de simulation d'équilibration (c_{eq}) et après l'étape de simulation de production (c_{pr}). Puisque les deux ligands actifs issus du premier criblage de 395 601 composés font chacun trois liaisons hydrogène avec Lu, nous avons décidé de ne sélectionner que les ligands faisant au moins trois liaisons hydrogène avec la protéine dans le second criblage de 1 295 678 composés. Les étapes DM (étapes (c_{eq}) et (c_{pr})) des simulations SBMD (dynamique moléculaire aux limites stochastiques) ont été réalisées sur ces ligands avec 60 ps de chauffage, 100 ps d'équilibration (étapes (c_{eq}) et (c_{pr})) et 200 ps de production (étape (c_{pr}) uniquement). Tout comme pour les simulations SBMD réalisées sur CD80 dans l'étape de validation du protocole de scoring, nous avons utilisé une sphère d'eau d'un rayon de 32 Å et les énergies des complexes protéine-ligand ont été minimisées avant de chauffer les complexes de 0 à 300 K suivant un incrément de 10 K. L'analyse des trajectoires de dynamique moléculaire a permis d'éliminer les ligands instables après chacune des étapes (c_{eq}) et (c_{pr}). Un ligand a été considéré comme instable lorsque la moyenne de RMSD calculée entre la pose initiale (pose de docking) et la pose correspondante à chacune des géométries extraites de la trajectoire était supérieure à 5 Å. La valeur de 5 Å a été choisie afin de ne pas exclure des ligands pour lesquels un ou plusieurs groupements périphériques subissent une réorientation au cours des simulations. Les calculs de RMSD ont été réalisés deux fois : (i) sur la fin de la trajectoire d'équilibration obtenue à l'étape (c_{eq}) et (ii) sur l'ensemble de la trajectoire de production obtenue à l'étape (c_{pr}).

L'étape (d) d'évaluation des énergies de complexation de la Figure 4 (p. 20) a été réalisée deux fois pendant le criblage des 1 295 678 composés. D'abord, les affinités moyennes des ligands ont été réévaluées avec la fonction X-Score à partir de 100 géométries issues de la trajectoire d'équilibration (étape (d_{eq}) de la Figure 1.1). Cette étape a permis de sélectionner les 1 000 meilleurs ligands dont les géométries initiales ont été générées

par DOCK6. Pour chacune des poses 1, 2 et 3 générées par rDOCK, les 500 meilleurs ligands ont été sélectionnés à cette même étape. Puis, les affinités ont été réévaluées une seconde fois avec la fonction HMScore sur 100 géométries issues de la trajectoire de production (étape (d_{pr})). Les fonctions X-Score et HMScore sont les fonctions de scoring implémentées dans la méthode de calcul XSCORE présentée à la page 51. Enfin, le top-20 des ligands provenant de l'ensemble des résultats rDOCK et DOCK6 a été analysé et commenté.

1.5 Les champs de force

Le champ de force est un ensemble de fonctions mathématiques et de paramètres qui permettent de décrire les structures des molécules ainsi que les interactions intramoléculaires et/ou intermoléculaires. Il permet de calculer de manière simplifiée les contributions énergétiques associées aux atomes d'un système moléculaire.

Plusieurs techniques de modélisation moléculaire utilisent un champ de force. C'est le cas du docking moléculaire, de certaines des techniques permettant l'évaluation des énergies d'interaction (fonctions de scoring) et de la dynamique moléculaire, techniques qui seront respectivement présentés dans les sections 1.6 (p. 39), 1.7 (p. 50) et 1.8 (p. 59). Dans le cas du docking moléculaire, qui consiste à générer différentes conformations d'un ligand sur une protéine, un champ de force est utilisé afin d'évaluer les énergies des complexes formés et ainsi de ne garder que les complexes de meilleure énergie. Les fonctions de scoring secondaire utilisent un champ de force en général plus élaboré que celui utilisé dans la technique de docking moléculaire afin d'évaluer plus précisément les énergies d'interaction. Enfin, la technique de dynamique moléculaire utilise un champ de force afin de simuler les mouvements des molécules au cours du temps.

Dans les trois techniques citées ci-dessus, les champs de force utilisent un découpage de l'énergie potentielle en deux types principaux: l'un correspondant aux contributions

covalentes ($E_{interactions\ liées}$), l'autre aux contributions non covalentes ($E_{interactions\ non\text{-}liées}$). Des expressions mathématiques associées à différentes sous-contributions énergétiques permettent de décrire ces contributions covalentes et non covalentes. Dans ces expressions, les paramètres ainsi que les fonctions mathématiques peuvent varier mais restent apparentés. Nous prendrons dans la suite l'exemple du champ de force CHARMM très populaire pour illustrer le concept de champ de force.

L'énergie potentielle d'une molécule ou d'un système plus complexe se décompose de la façon suivante :

$$E = E_{interactions\ liées} + E_{interactions\ non\text{-}liées} \quad (1.1)$$

avec:

$$E_{interactions\ liées} = E_{liaisons\ covalentes} + E_{angles\ de\ valence} + E_{angles\ dièdres} + E_{angles\ dièdres\ impropres} \quad (1.2)$$

$$E_{interactions\ non\text{-}liées} = E_{van\ der\ Waals} + E_{électrostatiques} \quad (1.3)$$

Contribution des liaisons covalentes et des angles de valence De manière simplifiée, les fluctuations des distances de liaison covalente ainsi que celles des angles de valence autour de leur valeur de référence s'apparentent aux mouvements décrits par des oscillateurs harmoniques. Ces fluctuations sont donc communément modélisées par des systèmes à ressort [69]. L'énergie associée à ces systèmes peut être calculée d'après les équations 2.4. Ces équations comprennent un paramètre représentant la valeur d'équilibre obtenue lorsqu'aucune force n'est appliquée (distance r_0 et angle θ_0) et une constante de force K_r qui modélise l'amplitude des fluctuations autour de la valeur d'équilibre.

$$E_{liaisons\ covalentes} = \sum^{liaisons\ covalentes} \frac{1}{2} K_r (r - r_0)^2 \quad (1.4)$$

$$E_{angles\ de\ valence} = \sum^{angles\ de\ valence} \frac{1}{2} K_\Theta (\Theta - \Theta_0)^2 \quad (1.5)$$

Contribution des angles dièdres classiques Les angles dièdres classiques correspondent à la rotation autour d'une liaison covalente, également appelée torsion. Le terme $E_{\textit{angles dièdres}}$ ne représente pas, à lui seul, la variation d'énergie due à la torsion. Il est utilisé pour compenser les contributions des autres termes au profil énergétique de torsion comme, entre autres, l'énergie de van der Waals entre deux atomes séparés par trois liaisons covalentes. L'énergie de torsion liée à la rotation autour d'une liaison covalente est modélisée par une fonction périodique dont la plus simple expression est présentée dans l'équation 1.6.

$$E_{\textit{angles dièdres}} = \sum^{\textit{angles dièdres}} K_\Phi(1 + cos(n\Phi - \delta)) \tag{1.6}$$

L'utilisation de la fonction cosinus permet d'obtenir des profils énergétiques périodiques, *i.e.* présentant plusieurs valeurs d'angle dièdre classique qui correspondent à des valeurs minimales locales de l'énergie[2]. Le paramètre n, appelé multiplicité ou périodicité de la rotation, est lié au nombre de minima énergétique par périodicité. Le paramètre δ, appelé phase, permet de définir les valeurs d'angle dièdre pour lesquelles le terme $E_{\textit{angles dièdres}}$ est minimal. Une constante de force K_Φ est utilisée afin de représenter les hauteurs de barrière énergétique séparant les minima. Ce paramètre permet donc de limiter la possibilité de rotation autour d'une liaison covalente. Afin d'améliorer la modélisation de l'énergie qui résulte de la rotation autour d'une liaison covalente, il est souvent utile d'étendre le terme $E_{\textit{angles dièdres}}$ en utilisant une série de Fourier. Sans l'aide de ces séries, il est souvent impossible de représenter les profils énergétiques les plus complexes. Par exemple, pour la rotation autour de la liaison covalente formée entre les deux atomes centraux du butane, deux expressions doivent être combinées pour définir le terme $E_{\textit{angles dièdres}}$. Une des expressions possède une périodicité de 2 et une phase de 180°, l'autre une périodicité de 3 et une phase de 0°. La combinaison de ces deux

[2]La fonction cosinus étant périodique sur 360° (cos(x)$\in [-1;1]$), pour obtenir une énergie positive ou nulle, il est nécessaire d'ajouter la valeur 1 aux résultats obtenus par la fonction cosinus.

expressions permet d'augmenter les énergies correspondant aux conformations décalées gauche et droite afin d'obtenir le profil énergétique attendu pour cette rotation dans le butane.

Contribution des angles dièdres impropres L'angle dièdre impropre représente la possibilité pour la position d'un quatrième atome D lié à l'atome central B d'un angle de valence $\widehat{ABC}$ de dévier par rapport au plan (ABC). L'utilité principale du terme $E_{\textit{angles dièdres impropres}}$ est de contrôler la planéité de certaines parties telles que les liaisons peptidiques ou les systèmes π. La forme de ce terme (équation 1.7) est inspirée de celle décrite précédemment pour $E_{\textit{liaisons covalentes}}$ et $E_{\textit{angles de valence}}$ avec le paramètre ϕ_0 permettant de représenter la valeur de l'angle correspondant à la géométrie d'équilibre et la constante de force K_{imp} qui limite la déviation par rapport à cette valeur d'angle.

$$E_{\textit{angles dièdres impropres}} = \sum^{\textit{angles dièdres impropres}} \frac{1}{2} K_{imp}(\phi - \phi_0)^2 \tag{1.7}$$

L'énergie électrostatique Le terme $E_{\textit{électrostatique}}$ (voir équation 1.8) est l'énergie associée à la force de Coulomb [70].

$$E_{\textit{électrostatique}} = \sum_{\textit{paire d'atomes ij}}^{\textit{couples d'atomes non-liés}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.8}$$

Afin de calculer l'énergie d'interaction entre deux atomes i et j, $E_{\textit{électrostatique}}$ tient compte des charges partielles q_i et q_j de ces deux atomes ainsi que de la distance r_{ij} séparant ces deux atomes. La constante ϵ_0 représente la valeur de la permittivité du vide.

L'énergie de van der Waals Le terme de van der Waals $E_{\textit{van der Waals}}$ permet de représenter l'énergie associée aux forces de dispersion de London [69] (partie attractive en $-\frac{1}{r^6}$ de l'énergie) et celle associée à la répulsion interélectronique due au principe d'exclusion de Pauli [71] (partie répulsive en $\frac{1}{r^{12}}$). De part sa double composition (parties attractive/répulsive), le terme $E_{\textit{van der Waals}}$ est typiquement représenté par un potentiel de

Lennard-Jones 6-12 [72] :

$$E_{\textit{van der Waals}} = \sum_{\substack{\textit{couples d'atomes non-liés} \\ \textit{paire d'atomes ij}}} \epsilon_{ij}[(\frac{R_{min,ij}}{r_{ij}})^{12} - 2(\frac{R_{min,ij}}{r_{ij}})^{6}] \tag{1.9}$$

Le paramètre ϵ_{ij}, appelé paramètre de Lennard-Jones, est dépendant des types atomiques de i et de j et définit la valeur absolue du minimum énergétique pour l'interaction de van der Waals entre les atomes i et j. Le paramètre $R_{\textit{min, ij}}$, appelé rayon du cœur répulsif, définit la distance de séparation entre les atomes i et j pour laquelle l'énergie de van der Waals est minimale. Lorsque r_{ij} prend la valeur $R_{\textit{min, ij}}$, les forces répulsives et attractives s'annulent, ce qui entraîne $E_{\textit{van der Waals}}$ = -ϵ_{ij}. En général, les paramètres du terme $E_{\textit{van der Waals}}$ pour les différentes paires d'atomes sont obtenus à partir des règles de combinaison de Lorentz-Berthelot [73, 74].

1.6 Docking moléculaire

La méthode de docking moléculaire existe depuis plus de 30 ans [75,76] et est largement utilisée aujourd'hui dans la recherche de molécules médicament [77]. Cette méthode consiste à prédire la conformation d'un ligand (molécule ou macromolécule) ainsi que son orientation dans une région donnée d'une protéine cible. La recherche de la conformation d'un ligand se fait en deux étapes : (i) la génération de différentes conformations possibles du ligand (poses) grâce à un algorithme de recherche, différents complexation sont obtenus ; et (ii) le calcul des énergies des complexes grâce à une fonction de scoring (fonction de scoring primaire) qui permet d'évaluer la pertinence des différentes conformations adoptées par ce ligand. Cette deuxième étape permet de faire un classement des scores (énergies calculées par la fonction de scoring) afin d'identifier les poses les plus probables du ligand. Le processus est le même lorsqu'on cherche à discriminer différents ligands selon leurs affinités pour la protéine cible (Figure 1.2) : l'algorithme de recherche va générer une ou plusieurs poses pour chaque ligand puis la fonction de scor-

ing primaire va permettre d'identifier le ligand qui se lie le plus fortement à la protéine cible.

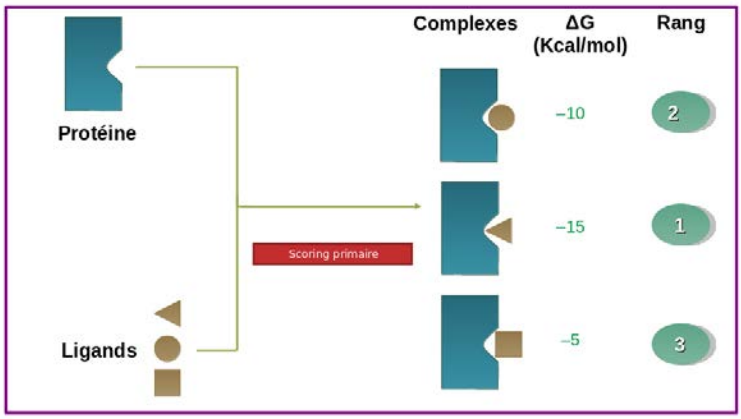


Figure 1.2: Principe du docking moléculaire illustré. Après la complexation des ligands sur la protéine, l'énergie de chacun des complexes est évaluée grâce à une fonction de scoring primaire qui calcule les énergies d'interaction entre chaque ligand et la protéine. Puis les ligands sont classés selon leur énergie de liaison à la protéine (score). Dans cet exemple, plus l'énergie est négative, plus le complexe protéine-ligand correspondant est stable et le rang associé bas.

Il existe deux types de docking : le docking rigide et le docking flexible. Ces deux méthodes diffèrent par la façon dont le ligand et la protéine sont traités pendant la recherche de la meilleure conformation. En effet, dans le docking rigide, le ligand et la protéine sont gardés dans leur géométrie initiale sans possibilité de changement de leur structure au cours de la recherche de la meilleure pose (ni angle, ni mouvement de rotation n'est autorisé). En revanche, dans le docking flexible, le programme prend en compte les modifications de la structure initiale du ligand, par exemple en changeant ses angles de torsion de manière à permettre une interaction plus forte avec la protéine. Pour la protéine, la flexibilité de quelques résidus peut aussi être prise en compte. Cette

méthode de docking flexible permet d'obtenir des poses *a priori* plus réalistes puisqu'elle prend en compte l'adaptation du ligand au site de liaison. Seule la méthode de docking flexible sera présentée plus en détail.

Docking flexible

Comme indiqué ci-dessus, la méthode de docking flexible permet de prendre en compte l'adaptation du ligand lors de sa liaison à la protéine. Ainsi, en permettant au ligand d'être flexible par des mouvements de rotation, des changements d'angles, mais aussi des mouvements de translation, on favorise de meilleures interactions entre le ligand et la protéine. Il existe trois grandes catégories d'algorithme de recherche de conformation qui prennent en compte la flexibilité du ligand : les algorithmes de recherche systématique qui regroupent les méthodes de reconstruction incrémentales et les méthodes de recherche exhaustives ; les algorithmes de recherche aléatoire qui regroupent, entre autres, les méthodes de Monte-Carlo et les algorithmes génétiques ; enfin, les algorithmes de recherche déterministe qui regroupent les méthodes de dynamique moléculaire et de minimisation d'énergie [76].

L'algorithme de recherche systématique consiste à explorer tous les mouvements de rotation possible du ligand : tous les degrés de liberté sont testés (0° à 360°) avec un pas incrémental choisi. Comme son nom l'indique, l'algorithme de recherche aléatoire permet de générer plusieurs poses d'un ligand en faisant des mouvements de translation, de torsion, mais aussi de rotation de façon aléatoire. Puis, une fonction de probabilité permet de retenir les poses les plus probables uniquement. L'algorithme de recherche déterministe permet aussi de simuler l'espace conformationel de la protéine par des méthodes de dynamique moléculaire ou de minimisation d'énergie. Il est difficile de prendre en compte la flexibilité d'une protéine qui possède beaucoup plus de degrés de liberté qu'une molécule d'une vingtaine d'atomes. Afin de contourner ce problème, cer-

tains programmes de docking tels que GLIDE [78] et FLEXX [79] permettent de prendre en compte la flexibilité de certains résidus choisis par l'utilisateur.

Alors que DOCK6 [34] utilise la méthode de reconstruction incrémentale (algorithme de recherche systématique), rDOCK [35] utilise une combinaison des méthodes Monte-Carlo et d'algorithme génétique (algorithme de recherche aléatoire) associé à des étapes de minimisation permettant de prendre en compte la flexibilité de certains groupements pour quelques résidus (algorithme de recherche déterministe). Chacune de ces méthodes utilisées dans les programmes DOCK6 et rDOCK est présentée dans les sous-sections qui suivent.

DOCK6

Algorithme de recherche de conformations Le programme DOCK6 [34] utilise la méthode de reconstruction incrémentale qui est une des méthodes de l'algorithme de recherche systématique. Cette méthode consiste à reconstruire le ligand fragment par fragment dans le site de liaison de la protéine cible suivant plusieurs étapes (Figure 1.3). Le ligand est d'abord décomposé en plusieurs fragments avant de définir un fragment de base qui est souvent un cycle : il s'agit de l'ancre (anchor). Dans un premier temps, l'ensemble des cycles est traité comme rigide. Si la molécule ne contient pas de cycle, c'est le plus gros fragment qui sera utilisé comme ancre.

Une fois que l'ancre a été définie, celui-ci est orienté de façon rigide dans le site actif préalablement défini par l'utilisateur. Plusieurs orientations de l'ancre dans le site de liaison sont testées et leurs énergies d'interaction (van der Waals et électrostatique) avec la protéine sont calculées avant d'être minimisées. Puis, les différentes orientations de l'ancre sont classées selon leur score (énergie d'interaction) afin de n'en retenir que les meilleures. L'étape suivante consiste à rajouter progressivement les autres fragments (parties flexibles) du ligand sur l'ancre. Chaque fragment rajouté à l'ancre est traité comme suit et dans cet ordre : (i) test de plusieurs orientations possibles du fragment ce

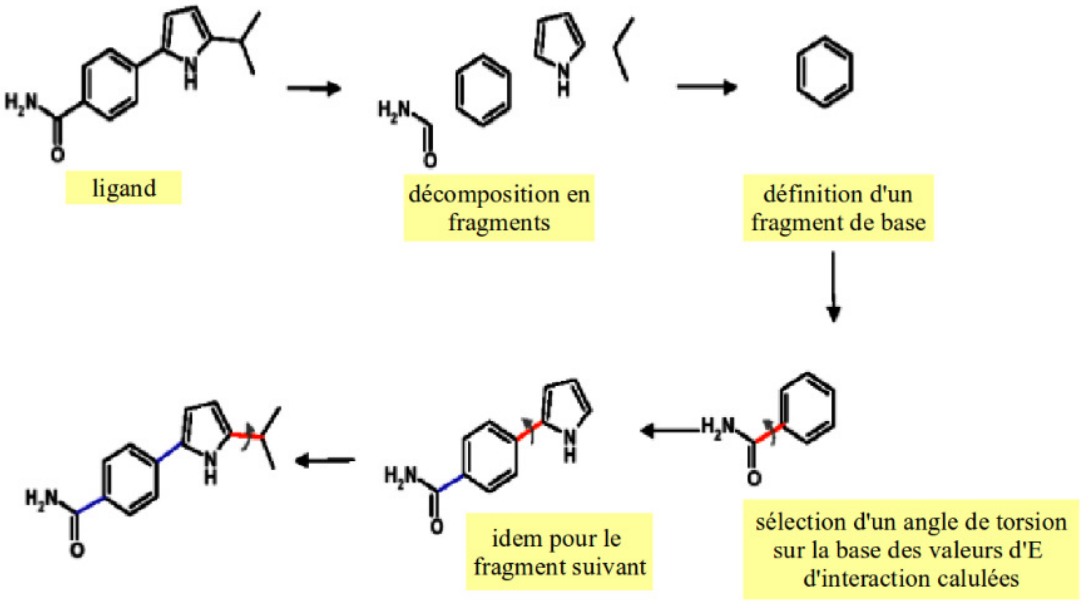


Figure 1.3: Méthode de reconstruction incrémentale du ligand dans le site actif qui permet de choisir les meilleurs angles de torsion de façon à optimiser l'interaction avec la protéine (la protéine n'apparaît pas ici pour raison de clarté).

qui permet d'obtenir une structure partielle du ligand dans différentes orientations ; (ii) calcul de l'énergie d'interaction (score) entre la structure partielle du ligand et la protéine ; (iii) optimation du complexe ligand partiel — protéine. Puis, un nouveau fragment est ajouté à la structure partielle du ligand dont seules les meilleures orientations sont conservées comme précédemment. Cette procédure se poursuit jusqu'à ce que le ligand soit entièrement reconstitué. Plusieurs conformations pour le ligand sont donc générées puisque les meilleures orientations sont gardées au cours de la construction du ligand.

Fonction de scoring primaire Chaque complexe protéine-ligand généré avec DOCK6 est évalué avec une fonction de scoring primaire qui permet d'estimer l'affinité de liaison entre une protéine et chacune des poses du ligand générées pendant le docking. Dans le cas de ce programme de docking, la fonction de scoring primaire se définit par la somme des énergies dues aux contributions de van der Waals et électrostatiques respectivement calculées selon un terme de Lennard-Jones et un potentiel de Coulomb.

rDOCK

Algorithme de recherche de conformations Le programme rDOCK utilise une combinaison de méthodes faisant partie des catégories d'algorithmes de recherche systématique et déterministe afin de générer les poses les plus probables pour chaque ligand. Pour générer la pose d'un ligand donné, rDOCK utilise trois variantes d'un algorithme génétique qui sont chacune suivie d'une étape de la méthode Monte-Carlo réalisée à faible température et d'une étape de minimisation. Le principe général de l'algorithme génétique qui est présenté dans la Figure 1.4 ainsi que celui de la méthode Monte-Carlo seront présentés dans cette sous-section.

Le principe des algorithmes génétiques, aussi appelé algorithme évolutionniste, est de mimer l'évolution biologique décrite par Darwin en 1859 [80]. Pour cela, les al-

gorithmes génétiques s'inspirent de phénomènes observés en biologie : la sélection, la recombinaison et la mutation. En biologie, l'information génétique est portée par des chromosomes constitués de gènes. Dans le cas, par exemple de la fécondation, les chromosomes du gamète mâle se recombinent avec ceux du gamète femelle (chromosomes parents) pour former une nouvelle combinaison de gènes : c'est la formation des chromosomes filles. Ces chromosomes filles sont similaires aux chromosomes parents mais diffèrent au niveau de quelques gènes en raison de ce brassage génétique mais aussi en raison d'une ou plusieurs mutations occasionnelles. Dans le cas de l'algorithme génétique appliqué au contexte du docking moléculaire, le chromosome représente une pose du ligand et les gènes qui constituent ce chromosome représentent les différents paramètres liés à la pose tels que les angles de torsion ainsi que les mouvements de rotation et de translation du ligand dans l'espace. Après avoir généré différentes poses de ligand, une fonction de survie (fitness function) permet de sélectionner les poses associées aux meilleurs scores (c'est à dire les poses dans lesquelles les ligands interagissent le mieux avec la protéine) : on obtient les chromosomes parents. Puis, toujours en mimant, ce qui est observé en biologie, certains paramètres des poses initiales du ligand (gènes des chromosomes parents) vont s'échanger (recombinaison) et d'autres vont être modifiés (mutations) pour obtenir de nouvelles poses qui diffèrent des poses initiales par quelques paramètres (chromosomes filles). Enfin, la fonction de survie permet d'évaluer chacune de ces nouvelles poses et seules les poses qui sont associées aux meilleurs scores sont gardées et deviennent à leur tour des «chromosomes parents» [81] (Figure 1.4).

Dans le cas du programme rDOCK, les paramètres (gènes) qui constituent les chromosomes sont le centre de masse du ligand, les mouvements de translation du ligand (en coordonnées cartésiennes), les angles dièdres qui permettent les mouvements de rotation du ligand et ceux qui permettent les mouvements de rotation des groupements OH et $NH_3{}^+$ terminaux de certains résidus. Les chromosomes parents sont formés de façon aléatoire à partir du site de liaison préalablement définie par l'utilisateur. Puis ces

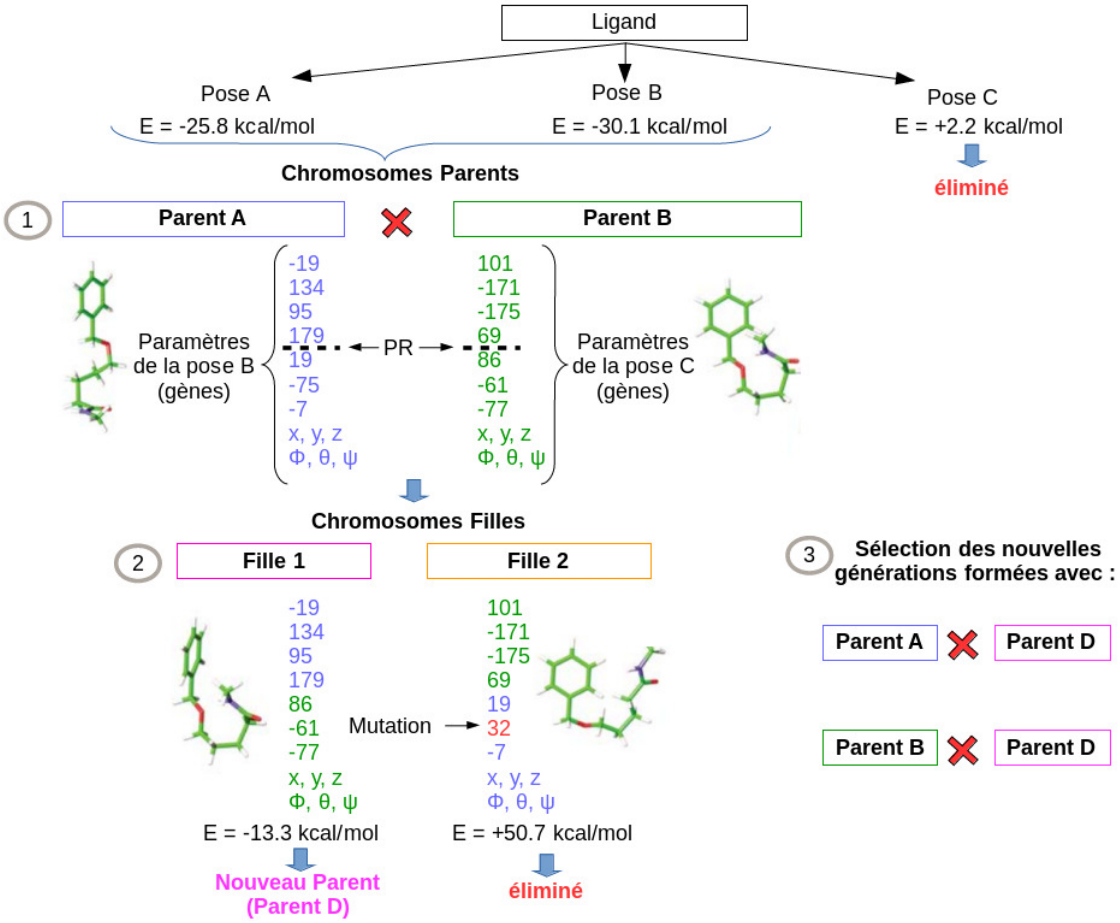


Figure 1.4: Principe général de l'algorithme génétique. Après avoir généré plusieurs poses d'un ligand, les énergies de liaison E ont été calculées grâce à une fonction de survie. Les poses de meilleures énergies deviennent les chromosomes parents. La croix rouge entre les couples de parents (parents A et B ; parents A et D ; parents B et D) symbolise la recombinaison des paramètres (gènes) qui constituent chaque parent. Chaque chromosome parent (pose) est formé de gènes (paramètres) qui peuvent être par exemple les angles de torsion, les angles de valence, les mouvements de translation (en coordonnées cartésiennes x, y, z) ou l'orientation du ligand dans le site de liaison (en coordonnées polaires...). Le point de recombinaison (PR) est le point au niveau duquel les informations (gènes) des chromosomes parents A et B (respectivement en bleu et en vert) vont s'échanger pour former les chromosomes filles (Fille 1 et Fille 2). Le chromosome Fille 2, qui a aussi subi une mutation ponctuelle (c'est-à-dire une modification d'un paramètre (gène)) est éliminé puisque son énergie de liaison est insatisfaisante alors que le chromosome Fille 1, qui constitue une pose de bonne énergie, est gardé et devient à son tour un chromosome parent (Parent D). Puis, d'autres poses du ligand sont générées grâce à la recombinaison des parents restants (Parents A, B et D) avant d'être sélectionnées.

chromosomes parents (poses initiales) vont subir différentes étapes de recombinaison et de mutations afin de donner des chromosomes filles qui seront sélectionnés grâce à un algorithme de survie. La méthode de sélection utilisée est la roulette qui consiste à sélectionner les poses (chromosomes filles) les plus probablement adaptées au site de liaison.

L'étape suivante est l'étape de génération de pose avec la méthode Monte-Carlo (MC). Dans la méthode MC, la pose du ligand est modifiée de façon aléatoire par des changements séquentiels d'un ou plusieurs paramètres liés aux mouvements de rotation, translation et/ou l'orientation du corps du ligand sur le site de liaison. L'énergie de la pose nouvellement formée (pose n+1) est comparée à celle de la pose qui la précède (pose n). Si la pose n+1 possède une énergie de complexation meilleure que celle de la pose n alors la pose n est éliminée et la pose n+1 est gardée. Dans le cas contraire, la pose n+1 est soumise au critère de Métropolis pour décider de son acceptation ou de son rejet. Le critère de Métropolis est dépendant de la température : plus la température est élevée et plus les poses de haute énergie seront gardées. Dans le cas de rDOCK, la méthode MC est utilisée à basse température, ce qui ne permet pas d'obtenir des conformations de hautes énergies. Puis, d'autres modifications sont réalisées sur la pose gardée (de nouveau appelée pose n) pour obtenir une nouvelle pose n+1 qui sera à son tour évaluée.

Chaque variante de l'algorithme génétique utilisé dans rDOCK (rDOCK utilise trois types d'algorithmes génétiques dénoncés variantes ici) est suivie d'une étape MC et d'une étape de minimisation. L'utilisation de ces différentes combinaisons d'algorithmes permet de générer et de sélectionner les poses les plus probables c'est-à-dire les poses les plus énergétiquement favorables avant une étape d'optimisation (minimisation) finale. L'étape de minimisation finale permet d'atteindre les minimums locaux d'énergie afin d'avoir des complexes protéine-ligand stables.

Fonction de scoring primaire La fonction de scoring primaire utilisée dans rDOCK est plus complète que celle utilisée dans DOCK6. En effet, alors que DOCK6 ne prend en compte que les interactions de van der Waals et électrostatiques établies entre le ligand et la protéine, rDOCK calcul les énergies des complexes de la façon suivante grâce à la fonction de scoring S^{tot} :

$$S^{\text{tot}} = S^{\text{inter}} + S^{\text{intra}} + S^{\text{site}} + S^{\text{restraint}} \tag{1.10}$$

Le terme S^{inter} est le terme qui prend en compte les interactions intermoléculaires (définies ci-dessous). Il s'agit d'un terme très important puisqu'il calcule les énergies d'interaction entre la protéine et le ligand. À ce terme sont ajoutés d'autres termes qui permettent d'affiner les calculs d'énergie en prenant en compte d'autres paramètres : les termes S^{intra}, S^{site} et $S^{restraint}$. Le terme S^{intra} prend en compte les interactions intramoléculaires du ligand et représente donc l'énergie relative de la conformation du ligand. Le terme S^{site} représente l'énergie relative à la flexibilité des groupements terminaux -OH et $-NH_3{}^+$ des résidus du site actif. Enfin, le terme $S^{restraint}$ prend en compte les contraintes externes qui vont influencer la liaison du ligand sur la protéine : il s'agit d'un ensemble de contraintes non physiques telles que, par exemple, l'utilisation de pharmacophores qui va permettre d'orienter le docking. Dans chacun des termes S^{inter}, S^{intra} et S^{site}, on calcule les contributions énergétiques apportées par les interactions de van der Waals (potentiel de Lennard-Jones), les interactions polaires attractives (telles que les liaisons hydrogène) et répulsives (telle que l'interaction entre deux donneurs de liaisons hydrogène) et l'effet de solvatation obtenu grâce au calcul de la surface accessible au solvant. Enfin, pour prendre en compte les effets du solvant lors du docking, nous avons utilisé l'algorithme dock_solv.prm. D'autres détails peuvent être trouvés dans le manuel de rDOCK [35].

Les programmes de docking DOCK6 et rDOCK, disponibles gratuitement, utilisent donc des techniques différentes afin de générer les poses et de discriminer les meilleures des moins bonnes conformations. Le tableau 1.1 résume les caractéristiques de chacun

Tableau 1.1: Tableau récapitulatif des principales caractéristiques des programmes de docking DOCK6 et rDOCK.

	DOCK6	rDOCK
Méthode de recherche de pose	· Reconstruction incrémentale · Minimisation	· Algorithme génétique · Monte Carlo · Minimisation
Fonction de scoring	· **Terme intermoléculaire :** vdW, électrostatique	· **Terme intermoléculaire (1) :** vdW, liaison hydrogène, interaction polaire de courte distance, énergie de torsion, énergie de solvatation. · **Terme intramoléculaire ligand :** idem que **(1)**. · **Terme intramoléculaire protéine (site de liaison) :** idem que **(1)**. Note : *énergie de torsion uniquement calculée pour les groupements terminaux -OH et $NH3^+$ des résidus
Solvatation	· Permittivité relative dans le terme électrostatique, $\varepsilon = 4\varepsilon_0$	· Possibilité de faire du docking avec des molécules d'eau cristallographique représentées de façon explicite · Solvatation implicite au travers du calcul de la surface accessible

de ces programmes. Un avantage du programme rDOCK est la possibilité de prendre en compte des molécules d'eau cristallographiques lors du docking. Ces molécules d'eau, représentées de façon explicite, peuvent être traitées de façon rigide ou flexible et peuvent avoir des mouvements de rotation et/ou de translation selon les options choisies par l'utilisateur. La prise en compte de molécules d'eau cristallographiques est importante dans l'étape du docking, notamment lorsque celles-ci participent à la liaison du ligand en formant des ponts [82]. Dans cette étude, nous avons considéré trois molécules d'eau cristallographiques de Lu dans les calculs de docking de 1 295 678 molécules sur cette protéine. Le choix de ces molécules d'eau est expliqué dans la section 5.2.

1.7 Fonctions de scoring secondaire

Les fonctions de scoring permettent d'estimer l'affinité des ligands dans les différentes poses générées par le programme de docking. Elles doivent pouvoir identifier les bonnes poses en décrivant au mieux les interactions entre la protéine et le ligand [83]. On peut distinguer les fonctions de scoring primaire et les fonctions de scoring secondaire. La fonction de scoring primaire est celle qui est intégrée au programme de docking. Celle-ci doit pouvoir décrire les interactions protéine-ligand de façon assez rapide afin d'évaluer chacune des poses générées par le programme de docking. En raison de cette contrainte de rapidité, la description des interactions protéine-ligand par ces fonctions de scoring primaire est peu précise. Il est donc toujours préférable de procéder à un nouveau classement des poses obtenues suite au docking grâce à une fonction de scoring secondaire plus élaborée. Les calculs étant plus longs par ligand, celle-ci ne peut s'appliquer que sur un nombre restreint de ligand. Cette fonction décrira toujours mieux les interactions protéine-ligand qu'une fonction de scoring primaire puisqu'elle prend en considération beaucoup plus de contributions énergétiques. Les fonctions de scoring (primaire et secondaire) diffèrent entre elles par les contributions énergétiques qui sont pris en compte dans les calculs des affinités. Leur efficacité est notamment dépendante du système sur lequel la fonction a été construite [84] : une fonction de scoring construite sur une métalloprotéine ne donnera pas de bons résultats si elle est appliquée à un système très différent tel qu'un anticorps par exemple.

Il existe trois catégories de fonction de scoring : (i) empirique (ii) basée sur le champ de force et (iii) basée sur les connaissances ou « *knowledge-based*». Dans cette étude, nous avons testé chacune de ces catégories : les trois fonctions de scoring empirique XSCORE [27], ID-Score [29] et Cyscore [28], les deux fonctions scoring *knowledge-based* ITScore [30] et DrugScore [31] et une fonction de scoring basée sur un champ de force MM-PBSA [32].

1.7.1 Fonction de scoring empirique

Le principe de la fonction de scoring empirique est de calculer les affinités grâce à une somme de termes individuels pondérés par des coefficients afin de reproduire les affinités expérimentales d'un set donné de ligands. Ces termes peuvent être par exemple le nombre de donneurs de liaison hydrogène, le nombre de liaisons flexibles, le nombre de contacts de van der Waals, etc. Cette catégorie de fonction de scoring est très rapide du fait de la simplicité de son équation. Le tableau 1.2 résume les caractéristiques des fonctions de scoring empirique XSCORE, ID-Score et Cyscore testées dans cette étude.

Les trois fonctions de scoring prennent en compte les interactions de van der Waals, les liaisons hydrogène, les effets de désolvatation du ligand et de la protéine lors de leur liaison ainsi que la flexibilité du ligand. Chacune de ces énergies est calculée différemment selon les fonctions de scoring. Ces différences seront commentées dans les paragraphes qui suivent. Dans cette section, nous développerons davantage la méthode de calcul XSCORE que les autres méthodes qui se sont montrées moins efficaces que XSCORE.

XSCORE

XSCORE est composée de trois fonctions pouvant être utilisées ensemble (équation [1]) ou individuellement : HPScore (équation [2]), HMSscore (equation [3]) et HSScore (équation [4]).

X-Score = (**HPScore** + **HMScore** + **HSScore**) / 3 [1]

$$\textbf{HPScore} = C_{0,1} + C_{VDW,1} \times [VDW] + C_{HB,1} \times [L.H] + C_{HP} \times [Paire\ hydrophobe] + C_{RT,1} \times [liaison flexible] \quad [2]$$

Tableau 1.2: Caractéristiques principales des fonctions de scoring secondaire XSCORE, ID-Score et Cyscore. Les valeurs (d_{ij}), f(θ_1), f(θ_2) dont les scores dépendront respectivement de la distance entre les atomes i et j, de l'angle établi entre les atomes DR, D et A, et de l'angle établi entre les atomes D, A et AR (Tableau 1.2, p. 52). Les atomes DR et AR correspondent respectivement à l'hétéroatome lié à l'atome D et à l'hétéroatome lié à l'atome A (Figure 1.5).

	X-Score	ID-Score	Cyscore
Interactions de van der Waals (potentiel de Lennard-Jones)	· Seuls les atomes lourds sont considérés dans les calculs d'énergie. · Valeur du score associé à l'interaction de van der Waals entre deux atomes ne dépasse pas 100. · Potentiel de Lennard-Jones 8-4	· Toutes les paires d'atomes sont considérées · 20 descripteurs qui diffèrent selon les types d'atomes qui sont en contact	· Atomes d'hyrogène ignorés · Potentiel de Lennard-Jones 10-6
Liaison hydrogène	· 3 descripteurs géométriques (d_{ij}), f(θ_1), f(θ_2) : $0 \leq f(d_{ij}) \leq 1$: selon distance calulée entre i et j $0 \leq f(\theta_1) \leq 1$: selon angle calculé entre DR, D et A $0 \leq f(\theta_2) \leq 1$: selon angle calculé entre D, A et AR } Figure 2.5 · Pas de différenciations entre les liaisons hydrogène chargés et neutres.	· Potentiel de liaison hydrogène · 10 descripteurs différents	· Potentiel de liaison hydrogène adapté selon l'étude de Goodford en 1985.
Entropie	· La flexibilité des groupements terminaux -CH3, -NH_2, -OH et -X (halogène) est ignorée · La flexibilité des portions cycliques est ignorée · Valeur de score entre 0 et 1 attribuée à un atome selon le nombre de liaisons flexibles auquel il est lié. Si l'atome est lié à plus de 2 liaisons flexibles, un faible score de 0.5 lui est attribué.	· Somme de liaisons flexibles et pourcentage d'atomes lourds non lipophiles (équation de Eldridge et collaborateurs)	· Somme des liaisons flexibles
Effet hydrophobe	· 3 fonctions empiriques : - Algorithme de surface hydrophobe (HSScore), calcul de la surface accessible du solvant (SAS) - Algorithme de contacts hydrophobes (HPScore) - Algorithme de complémentarité (« matching ») hydrophobe (HMScore)	· Effet de désolvatation : - LogP - PSA - Volume du ligand - Effet hydrophobe du site de liaison - SASA	· Facteur de courbure calculé sur la surface accessible au solvant (SAS)
Autre		· Interactions électrostatiques · Interactions π-π · Interactions métal – ligand · Effet de désolvatation · Complémentarité de forme · Complémentarité physico-chimique	

$$\textbf{HMScore} = C_{0,2} + C_{VDW,2} \times [VDW] + C_{HB,2} \times [L.H] + C_{HM} \times [Complementarite\ hydrophobe] + C_{RT,2} \times [liaisonflexible] \quad [3]$$

$$\textbf{HSScore} = C_{0,3} + C_{VDW,3} \times [VDW] + C_{HB,3} \times [L.H] + C_{HS} \times [Surface\ hydrophobe] + C_{RT,3} \times [liaisonflexible] \quad [4]$$

Dans les équations ci-dessus, les termes *VDW* et *L.H* correspondent aux contributions liées respectivement aux interactions de van der Waals et aux liaisons hydrogène. Les termes « Paire hydrophobe », « Complémentarité hydrophobe » et « Surface hydrophobe » correspondent aux contributions liées à l'effet hydrophobe occasionné par la désolvatation des complexes. Le terme « liaison flexible » correspond au terme d'entropie. Les termes $C_{0,1}$, $C_{0,2}$ et $C_{0,3}$ sont des constantes spécifiques aux équations respectives [2], [3] et [4]. Les termes $C_{VDW,x}$, $C_{HB,x}$, $C_{RT,x}$ (avec x = 1, 2 ou 3) correspondent respectivement aux coefficients spécifiques des termes *VDW*, *L.H* et « liaison flexible ». Enfin, les termes C_{HP}, C_{HM} et C_{HS} correspondent aux coefficients respectifs des termes « Paire hydrophobe », « Complémentarité hydrophobe » et « Surface hydrophobe ».

Les fonctions HPScore, HMScore et HSScore ne diffèrent que par les coefficients cités ci-dessus ainsi que les termes qui permettent de calculer l'effet hydrophobe (« Paire hydrophobe », « Complémentarité hydrophobe » et « Surface hydrophobe »). L'effet hydrophobe se définit comme le rapprochement des groupements non polaires afin d'exclure les molécules d'eau lors de la désolvatation de la protéine et du ligand.

Dans la fonction HPScore, l'effet hydrophobe est calculé grâce à un algorithme de contacts hydrophobes et correspond à la somme arithmétique des paires d'atomes hydrophobes en contact entre la protéine et le ligand. La force de chaque contact hydrophobe est évaluée par la distance entre les deux atomes : un score de 1 est attribué

lorsque la distance entre les deux atomes est optimale et un score de 0 est attribué lorsque la distance est supérieure à la distance de référence (soit une distance égale à la distance de référence additionnée à 2 Å). Dans la fonction HMScore, des scores supérieurs à 0 (c'est à dire favorables) sont décomptés si un atome hydrophobe du ligand se trouve dans une région hydrophobe du site de liaison : on parle de correspondance ou complémentarité hydrophobe. La valeur du score associée à cette fonction est aussi dépendante du LogP_i qui correspond à la contribution de l'atome *i* dans la lipophilicité du ligand. Enfin, la fonction HSScore prend en compte l'effet hydrophobe grâce au calcul de la surface accessible du ligand et de la protéine.

Les liaisons hydrogène contribuent à l'énergie de liaison et favorisent la stabilité du ligand sur la protéine. La formation d'une liaison hydrogène dépend de la distance entre un hétéroatome donneur de liaison hydrogène (D) et un hétéroatome accepteur de liaison hydrogène (A) ainsi que de l'angle formé entre l'atome D, l'atome d'hydrogène et l'atome A. XSCORE décrit la liaison hydrogène suivant trois descripteurs géométriques : d_{ij}, f(θ_1), f(θ_2) dont les valeurs dépendront respectivement de la distance entre les atomes i et j, de l'angle établi entre les atomes DR, D et A, et de l'angle établi entre les atomes D, A et AR (Tableau 1.2, p. 52 et Figure 1.5). Les atomes DR et AR correspondent respectivement aux atomes liés à D et à A.

Le terme d'entropie est pris en compte en considérant les liaisons flexibles du ligand. Comparée aux fonctions de scoring ID-Score et Cyscore, la fonction XSCORE fait beaucoup d'approximations en ignorant les mouvements de rotation de certains groupements terminaux et les groupements cycliques (Tableau 1.2). De plus, la fonction XSCORE tente de ne pas surestimer les énergies dues aux mouvements de rotation des liaisons flexibles liées par un même atome (liaisons flexibles croisées) en attribuant un faible score à ce dernier. On parle de liaisons flexibles croisées lorsque les mouvements de ces liaisons flexibles interfèrent entre elles, contrairement aux liaisons flexibles isolées qui sont libres de rotation (Figure 1.6, p. 56). Ainsi, le mouvement d'une liaison flexible

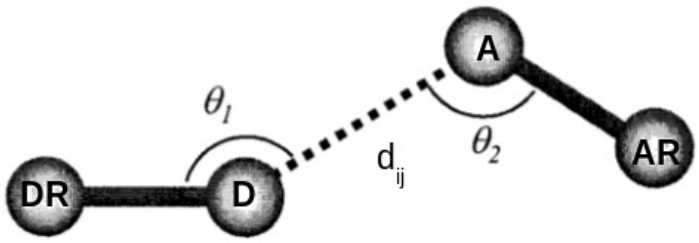


Figure 1.5: Paramètres permettant de calculer la contribution énergétique de la liaison hydrogène dans la fonction de scoring XSCORE. Les distances sont calculées entre l'atome accepteur A et l'atome donneur D. Les atomes DR et AR correspondent aux atomes liés à l'atome A et à l'atome D. L'angle θ_1 est l'angle formé entre les atomes DR, D et A. L'angle θ_2 est l'angle formé entre les atomes D, A et AR.

est atténué lorsqu'il a lieu en même temps que celui d'une autre liaison flexible liée au même atome. Ce cas de figure n'est pas pris en considération dans les fonctions de scoring classique qui traitent donc les liaisons flexibles croisées et les liaisons flexibles isolées de la même façon.

Enfin, la fonction XSCORE utilise une version inhabituelle du potentiel de Lennard-Jones en 8-4 pour calculer les énergies d'interaction de van der Waals. Par exemple, XSCORE ne considère que les atomes lourds (Tableau 1.2, p. 52).

Cyscore et ID-Score

Comme la fonction de scoring XSCORE, Cyscore utilise la fonction de Lennard-Jones sans considérer les atomes d'hydrogène pour calculer les énergies de van der Waals. Pour les contributions énergétiques apportées par les liaisons hydrogène et les liaisons flexibles, Cyscore utilise le potentiel utilisé dans le programme de Grid de Goodford [85] et, comme la plupart des fonctions de scoring, il comptabilise toutes les liaisons flexibles sans différencier les liaisons flexibles « isolées » des liaisons flexibles « croisées ».

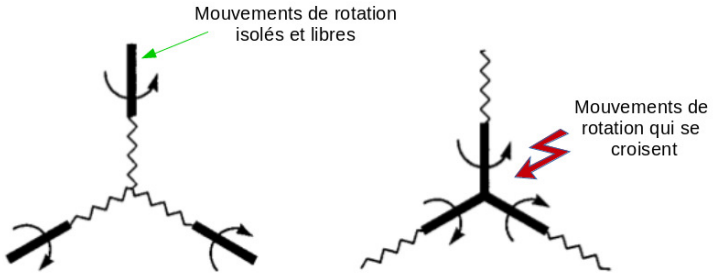


Figure 1.6: Liaisons flexibles isolées (à gauche) et croisées (à droite) [27]. Les mouvements de rotation des liaisons flexibles croisées sont atténués comparativement à des mouvements de rotation des liaisons flexibles isolées qui sont libres. Les liaisons flexibles croisées sont liées à un même atome. Une liaison flexible isolée est éloignée d'une autre liaison flexible isolée par au moins une liaison (représentée en zig-zag).

La particularité de Cyscore est la méthode utilisée pour décrire l'effet hydrophobe : le facteur de courbure de la surface accessible au solvant. Ce facteur permet de prendre en compte la forme du site de liaison en différenciant les surfaces planes des surfaces concaves et convexes. Les détails de cette méthode peuvent être trouvés dans l'étude de Yang Cao et Lei Li [28].

En plus des termes énergétiques énoncés au début de cette sous-section 1.7.1 (interactions de van der Waals, liaisons hydrogène, effets de désolvatation et flexibilité du ligand), ID-Score prend en compte les interactions électrostatiques, les interactions $\pi - \pi$, les interactions métal-ligand, la complémentarité de forme, les correspondances physico-chimiques entre les groupements du ligand et ceux de la protéine et l'effet de désolvatation (Tableau 1.2 (p. 52). La fonction ID-Score permet donc une description assez complète des interactions protéine-ligand. La particularité de cette fonction de scoring est le grand nombre de descripteurs (paramètres) qui sont considérés pour le calcul d'un seul terme : par exemple, les interactions de van der Waals sont décrites avec pas moins de 20 descripteurs. Un total de 50 descripteurs est utilisé dans la fonction ID-Score [29].

1.7.2 Fonction de scoring knowledge-based

La fonction de scoring *knowledge-based* se base sur une étude statistique des interactions protéine-ligand observées dans les structures résolues expérimentalement. Cette fonction de scoring est construite grâce à une analyse statistique des paires d'atomes protéine-ligand. Le principe est donc d'obtenir des informations énergétiques (potentiels de paires) à partir d'information structurale, cela en utilisant le logarithme de l'équation de Boltzmann. Bien que cette catégorie de fonction de scoring soit aussi rapide que les fonctions de scoring empirique, elle est limitée par l'utilisation d'un état de référence. L'état de référence est l'état dans lequel les interactions interatomiques sont nulles. Cet état est un état imaginaire, non disponible dans les banques de structures et qui est construit sur la base de l'approximation quasi-chimique [86]. L'approximation quasi-chimique considère qu'il existe un équilibre chimique entre les paires d'atomes non liés. Il est difficile d'obtenir un état de référence idéal puisque les méthodes utilisées pour calculer cet état de référence ne sont pas suffisamment précises [87]. La fonction de scoring ITScore [30] a été construite grâce à une méthode itérative qui permet de contourner ce problème lié à l'état de référence. Le principe de la méthode itérative est d'améliorer les potentiels de paires extraites des structures de complexes protéine-ligand jusqu'à ce que la fonction puisse discriminer les bonnes des mauvaises poses adoptées par les ligands.

La construction de la plupart des fonctions de scoring *knowledge-based* se base sur des structures de complexes protéine-ligand (le cas également de ITScore). En revanche, pour construire la fonction de scoring DrugScore (DSX) [31], plutôt que d'utiliser des structures de complexes protéine-ligand, des cristaux de petites molécules organiques ont été utilisés afin d'y extraire des potentiels de paires dépendant des distances interatomiques. La particularité de cette fonction est qu'elle prend aussi en compte l'effet de désolvatation grâce à des potentiels de paires basés sur la surface accessible au solvant. De plus, cette fonction de scoring possède aussi des potentiels de paires basés sur les an-

gles de torsion, ce qui permet de relaxer les poses des ligands et d'éviter des incohérences dans les angles de torsion.

1.7.3 Fonction de scoring basée sur le champ de force

La fonction de scoring basée sur le champ de force calcule les énergies des interactions physiques qui sont établies entre la protéine et le ligand telles que les interactions non-liées (van der Waals et électrostatiques) et les interactions liées (énergies de torsion, contributions de liaisons covalentes et des angles de valence). Les paramètres de champ de force utilisés dans ces calculs sont dérivés de données expérimentales et de calculs *ab initio* de chimie quantique. Cette méthode de calcul peut être combinée au modèle de solvatation implicite GB/SA ou PB/SA afin de prendre en compte l'effet du solvant. C'est le cas de la méthode MM-PBSA.

La méthode MM-PBSA consiste à calculer les énergies de liaison grâce à la méthode de mécanique moléculaire (MM) et à prendre en compte les effets de solvatation de façon implicite grâce à la méthode PB/SA (*Poisson-Boltzmann Surface Area*). Une variante de cette méthode est la méthode GB/SA (*Generalized Born Surface Area*). Les termes PB et GB permettent de calculer la composante électrostatique qui correspond à l'interaction entre l'eau et le soluté. Lors de la désolvatation, ces termes calculent le coût énergétique occasionné par la désolvatation des solutés (protéine et ligand). Quant au terme SA, il calcule la variation de la surface accessible au solvant lors de la complexation. Dans la partie MM de ces méthodes, les interactions liées et non-liées citées précédemment sont prises en compte dans les calculs d'énergie de liaison.

1.8 Simulations de dynamique moléculaire

Les simulations de dynamique moléculaire (DM) consistent à simuler le mouvement des molécules au cours du temps en calculant la position des atomes grâce à l'équation de Newton et à une équation dite d'intégration telle que celle de Verlet [69]. Pour cela, il est nécessaire d'utiliser un champ de force qui permet de calculer de manière simplifiée les contributions des forces issues de la mécanique classique [69,88] (section 1.5, p. 35). Le calcul de ces forces sur chaque atome permet de simuler la dynamique du système. Dans cette étude, des simulations DM ont été réalisées afin de relaxer les complexes protéine-ligand issus du docking (Figure 1.7). Grâce à la méthode DM, la protéine peut adapter sa conformation, en particulier celle de ses chaînes latérales par rapport au ligand et vice-versa.

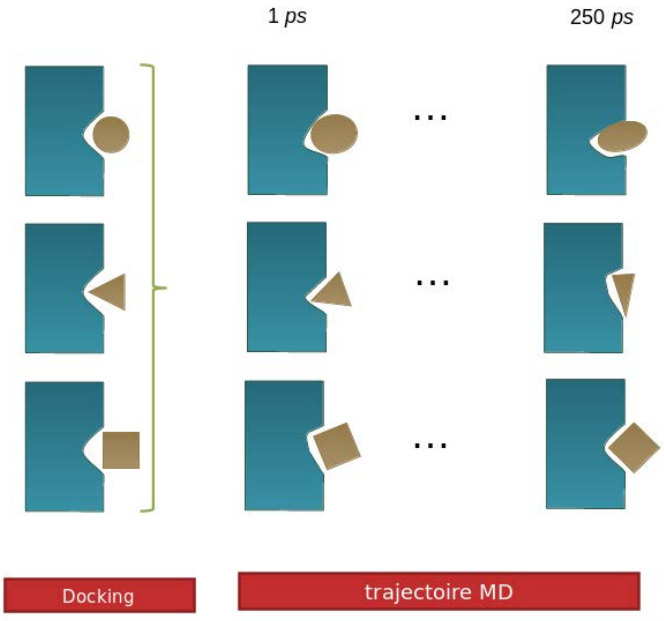


Figure 1.7: La simulation DM permet de relaxer les complexes protéine-ligand au cours du temps.

L'ensemble des simulations de dynamique moléculaire présentée dans cette étude ont été réalisées en utilisant le programme CHARMM [89] avec l'algorithme d'intégration des équations newtoniennes du mouvement *leapfrog* [90] et un pas de temps d'une femtoseconde (fs). Les distances des liaisons covalentes formées avec des atomes d'hydrogène ont été fixées à leurs valeurs de référence (r_0) en utilisant l'algorithme *SHAKE* [61]. Les contraintes appliquées par *SHAKE* permettent d'éliminer les fluctuations de ces liaisons covalentes qui interviennent à des fréquences élevées, *i.e.* dans des temps très courts ($\sim$0,5 fs). En utilisant *SHAKE*, l'énergie correspondant aux liaisons covalentes formées avec des atomes d'hydrogène n'intervient pas dans l'énergie finale.

Lors des simulations de dynamique moléculaire, deux représentations du solvant peuvent être utilisées: représentation implicite et représentation explicite. Contrairement à la représentation explicite, dans une représentation implicite du solvant, les molécules d'eau ne sont pas incluses dans le système. Cette représentation implicite permet de tenir compte de certains effets relatifs au solvant (effet d'écran des charges atomiques et/ou friction entre soluté et solvant) tout en diminuant les temps de calcul nécessaires. Pour la représentation explicite du solvant, il est possible de diminuer les temps de calcul en limitant le nombre de molécules d'eau incluses dans le système.

Dans les sous-sections suivantes, les méthodes de simulation utilisées dans cette étude employant une solvatation implicite ou explicite sont décrites.

1.8.1 Simulations de dynamique moléculaire Langevin, LD

Afin de diminuer la mobilité des atomes et de représenter plus précisément les effets du solvant, l'équation newtonienne du mouvement peut être remplacée par l'équation de Langevin (équation 2.11) [59].

$$m_i\vec{a_i} = m_i\frac{d\vec{v_i(t)}}{dt} = \vec{F_i} - m_i\gamma_i\vec{v_i}(t) + \vec{R_i}(t) \tag{1.11}$$

La force $\vec{F_i}$ est calculée à partir de l'équation générale du potentiel énergétique par la

relation: $\vec{F_i} = -\frac{\partial}{\partial \vec{r_i}} V(\vec{r_1}, \vec{r_2}, ..., \vec{r_N})$ avec $(\vec{r_1}, \vec{r_2}, ..., \vec{r_N})$ les coordonnées cartésiennes des atomes 1 à N. Le paramètre γ_i correspond au coefficient de friction de la particule i tandis que le vecteur $\vec{R_i}$(t) représente la fonction aléatoire de fluctuation du soluté modélisant la collision avec le solvant. En fait, la fonction aléatoire est utilisée afin de compenser la perte énergétique due au phénomène de friction. Tandis que la force de friction dissipe l'énergie du système, la fonction aléatoire fournit une énergie additionnelle au système[3]. Ce phénomène est appelé théorème de fluctuation-dissipation [92] et traduit la relation existante entre γ_i et $\vec{R_i}(t)$ permettant de maintenir la température du système constante. Du fait de ce phénomène, les dynamiques de Langevin possèdent donc un thermostat intégré. Dans toutes les simulations LD réalisées dans cette thèse, la valeur du paramètre γ_i a été fixée à 50 ps^{-1}.

1.8.2 Simulations de dynamique moléculaire aux limites stochastiques ou *Stochastic Boundary Molecular Dynamics*, SBMD

La méthode *Stochastic Boundary Molecular Dynamics* (SBMD) [60, 93] a été imaginée dans le but de limiter le nombre de molécules du solvant contenues dans le système pour des simulations d'un processus localisé dans l'espace, ne faisant intervenir qu'une partie du système [94]. Cette méthode est particulièrement intéressante pour l'étude précise d'une région spécifique d'un système comme par exemple le site d'interaction protéine-ligand. Elle permet, lors des calculs, de ne pas prendre en compte la plupart des atomes n'intervenant pas directement dans le phénomène étudié. Elle est également très utile

[3]Ainsi, le seul paramètre de l'équation de Langevin qui peut être ajusté correspond à γ_i. Plus la valeur du paramètre γ_i tend vers zéro, plus la simulation de dynamique de Langevin se rapproche d'une simulation de dynamique moléculaire classique. Il est également important de noter que la valeur optimale du coefficient de friction peut dépendre du modèle de solvatation utilisé et, pour certains modèles, une valeur élevée peut permettre d'obtenir des résultats plus en accord avec les données expérimentales [91].

pour vérifier l'importance de certaines molécules d'eau dans la structure d'une protéine ou d'un complexe, telles que des molécules d'eau cristallographique pouvant jouer un rôle important dans les interactions protéine-ligand par exemple. La méthode SBMD a été maintes fois utilisée pour l'étude de systèmes biologiques et il a été montré qu'elle permet une description correcte des mouvements et interactions intervenant au niveau du site actif [95].

La méthode SBMD repose sur un découpage du système en trois régions: réaction, tampon et réservoir (figure 1.8). Les atomes de la région de réaction sont soumis à une dynamique classique sans contrainte tandis que les mouvements de ceux de la région tampon sont régis par l'équation de Langevin (équation 2.11)[4]. La région réservoir entoure la zone d'interaction (régions tampon et de réaction) et fournit un champ de force statique où les atomes sont contraints, ce qui permet de maintenir les propriétés d'équilibre des atomes de cette zone. Des molécules d'eau[5] sont incluses seulement dans les régions de réaction et tampon et peuvent diffuser librement entre ces deux régions [60]. Afin d'éviter l'apparition de problèmes liés aux effets de bord, il est nécessaire de définir avec précaution les parties du système contenues dans les différentes régions.

[4]L'espace entre les régions de réaction et tampon doit être assez faible afin d'éviter d'avoir à utiliser un modèle de thermostat sophistiqué.

[5]Pour les simulations SBMD, la version de CHARMM du modèle de molécules d'eau TIP3P a été utilisée [89].

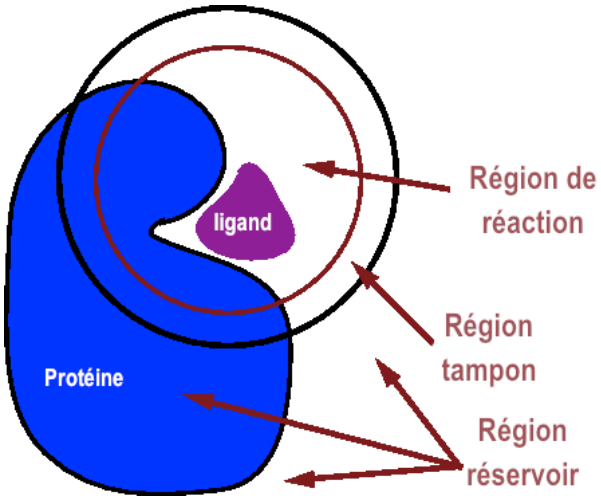


Figure 1.8: Représentation du découpage du système lors des simulations SBMD.

Chapitre 2

Recherche de systèmes similaires Lu-protéine ou Lu-molécule

Pour rechercher des inhibiteurs de l'interaction Lu-Lnα5 pouvant de lier au domaine D2 de Lu, il aurait fallu dans un contexte idéal, connaître le site de liaison de petites molécules ou de protéines sur Lu et avoir plusieurs ligands de Lu avec des structures et des constantes d'affinité connues avec lesquels un protocole de criblage pourrait être calibré. Puisque nous n'avons pas trouvé de tels partenaires pour Lu, nous avons recherché un système similaire à celui de Lu. La recherche d'un système similaire a deux objectifs. Le premier est de permettre la déduction d'un site de liaison probable sur Lu à partir de la comparaison entre ce système similaire et le domaine D2 de Lu. Dans ce premier objectif, le système similaire doit être complexé à une autre molécule (protéine ou petite molécule), sans nécessairement de constante d'affinité associée. Le second objectif est l'élaboration d'un protocole de criblage à partir de ce système similaire comprenant une protéine (homologue ou présentant seulement une ressemblance locale, analogue) et des constantes d'affinité connues.

Une recherche BLAST a permis d'identifier la protéine NCAM-1 qui possède 36%

d'identité de séquence avec Lu (pour les séquences entières de Lu et NCAM-1). Cependant, aucun ligand lié à NCAM-1 n'a été trouvé. Cette recherche BLAST a aussi donné une liste de structures de protéines qui possèdent un faible taux d'identité de séquences avec Lu (moins de 30%). Étant donné que l'homologie entre deux protéines n'est pas garantie lorsque le taux d'identité de séquence est aussi faible, les structures correspondantes ont été comparées à celle du domaine D2 de Lu. Lorsque les deux types d'alignement (séquences et structures) ne révélaient pas d'écart dans les structures secondaires, d'une part, et montraient une topologie similaire des domaines respectifs d'autre part, une recherche de structure protéine-protéine ou protéine-ligand a été réalisée dans la littérature et dans les banques de données ChemBL [96] et BindingDB [97] à partir de la structure trouvée similaire au domaine D2 de Lu. Parmi les résultats BLAST, de nombreuses protéines d'adhésion cellulaires ont été trouvées. Par exemple les protéines Nectine-1, Nectine-2 et Nectine-4 présentaient des taux d'identité de séquences entre 13 % et 16 % avec des RMSD entre 1.5 et 1.9 Å pour les alignements des structures respectives (Tableau 2.1). Cependant, aucun ligand lié à ces Nectines n'a été trouvé.

Tableau 2.1: Pourcentage d'identité et RMSD calculé pour les protéines identifiées dans la recherche de séquence ou de structure homologue ou analogue de Lu. Les pourcentages associés à chaque homologue ou analogue ont été calculés pour tous les domaines noté D*N* (à partir du N-terminal) aligné avec le domaine D2 de Lu avec *N* qui est le domaine N-terminal de la protéine. Les alignements de structures ont été calculés entre le domaine D2 de Lu et les domaines des homologues ou analogues pour lesquels un fort pourcentage d'alignement de séquence a été constaté.

Protéine	code PDB	Identité (%)	RMSD (Å)
RAGE	3CJJ	20 (D1), 14 (D2)	2.2
CD200R	4BFG	12 (D1)	2.1
VCAM-1	1VSC	16 (D1), 14 (D2), 17 (D3), 19 (D4), 13 (D5), 19 (D6), 13 (D7)	1.7
ICAM-1	1IAM	10 (D1), 12 (D2), 9 (D3), 12 (D4), 11 (D5)	1.7
NCAM-1	2NCM	16 (D1), 12 (D2), 16 (D3), 19 (D4), 15 (D5)	2.0
NCAM-2	2XY1	16 (D1), 10 (D2), 19 (D3), 14 (D4), 15 (D5)	1.8
Nectin-1	3ALP	15 (D1), 16 (D2)	1.9
Nectin-2	4FMK	15 (D1), 14 (D2)	1.6
Nectin-4	4FRW	16 (D1), 13 (D2)	1.5

Des recherches d'analogues de Lu ont ensuite été réalisées avec VAST [38], VAST+ [39], Delta-NCBI-BLAST [98], DALI [40], PDBeFOLD [37] et l'outil 3D-similarity dans la PDB [99]. Des résultats ont été obtenus tels que le premier domaine de la protéine CD200R qui possède un taux d'identité de séquence de 12 % et un RMSD de 2.1 Å selon l'alignement structural (Tableau 2.1). La recherche de protéines ou de molécules partenaires à CD200R a permis de trouver son partenaire biologique CD200 (code PDB 4BFI). Dans cette structure cristallographique CD200R et CD200 interagissent au niveau de leur domaine V. Il n'y a cependant pas de similarité observée avec le domaine D2 de Lu.

Les serveurs de recherche PDBeFOLD, VAST et VAST+ ont révélé les deux autres analogues RAGE et VCAM-1 (Tableau 2.1). De plus, VAST et l'outil de recherche 3D-similarity ont permis de trouver les analogues NCAM-2 et ICAM-1 (Tableau 2.1). Cependant, malgré une recherche étendue, aucun ligand qui cible les domaines analogues de NCAM-2 et VCAM-1 n'a été trouvé. En revanche, la structure d'un partenaire se liant à ICAM-1 est connue : il s'agit de LFA-1. Toutefois, la liaison de ICAM-1 à LFA-1 nécessite un ion Mg^{2+}. Ce mode de liaison est différent de celui de Lu avec Lnα5 puisque l'ion Mg^{2+} n'est pas nécessaire pour cette interaction. En effet, il a été montré que la liaison de Lu à Lnα5 n'était pas inhibée en présence d'EDTA, capable de chélater les ions Mg^{2+} [22]. D'autres part, la structure cristallographique de RAGE montre que cette protéine interagit avec son ligand biologique S100A13 (code PDB 2LE9). Cependant, l'analyse de l'interface de liaison révèle que les résidus impliqués dans l'interaction ne sont pas identiques ou même similaires à ceux trouvés sur les faces se trouvant près des résidus Glu132/Asp133 ou Asp198/Asp199 de Lu.

Enfin, des protéines présentant des similarités cette fois-ci locales avec le second domaine de Lu ont été recherchées. Les serveurs suivants ont été utilisés pour rechercher des poches similaires aux poches peu profondes détectées sur le second domaine de Lu : POSSUM, GIRAF, CMASA et SiteComp [41,42,100,101]. En plus de cette recherche, Site-

Comp identifie les différences entre les sites de liaison et détecte les résidus impliqués dans la liaison des ligands. Ces serveurs ont donné des résultats insatisfaisants : seule une liste de protéines n'ayant pas de région d'interaction similaire à la région proche des résidus Glu132/Asp133 ou Asp198/Asp199 de Lu et/ou n'ayant pas de ligands connus a été trouvée.

Finalement, la recherche du site de liaison sur le domaine D2 de Lu a été réalisée grâce aux serveurs de prédiction. Les deux serveurs Q-siteFinder [45] et SiteHound [47] ont prédit une large zone de liaison proche des résidus Glu132/Asp133. Cette zone de liaison prédite est plane en raison des brins bêta qui la constituent et est principalement définie par des résidus hydrophiles non chargés et par des résidus hydrophobes (Figure 2.1).

Puisque les résultats précédents n'ont pas permis de trouver un homologue ou analogue de Lu qui possède plusieurs ligands avec des structures et des constantes d'affinité connues et permettant de construire le protocole de criblage, nous avons recherché une protéine semblable à Lu dans les bases de données 2P2I [102] et TIMBAL [26] qui listent les protéines impliquées dans les interactions protéine-protéine et leurs inhibiteurs. Dans cette liste, la protéine CD80 a été retenue. En effet, CD80 possède deux domaines extracellulaires Ig-like : un domaine V et un domaine C2-set. La structure cristallographique de code PDB 1I8L révèle que CD80 interagit avec son ligand biologique CTLA-4 au niveau du domaine V. De plus, pour cette protéine, il existe une série de ligands qui se lient au domaine V de CD80 afin d'inhiber son interaction avec CTLA-4 et CD28. Cependant, le site de liaison précis ainsi que le mode de liaison de ces ligands sur CD80 sont inconnus. Afin de définir plus précisément un site de liaison sur le domaine V de CD80, on s'est appuyé sur : (i) la structure cristallographique du complexe CD80—CTLA-4 (code PDB 1I8L) ; (ii) les prédictions des sites de liaison grâce à des serveurs Web ; (iii) les études de mutagenèse réalisées sur le domaine V de CD80 qui montrent que les résidus Asn48 et Trp50 sont très importants pour la liaison des ligands inhibiteurs [54].

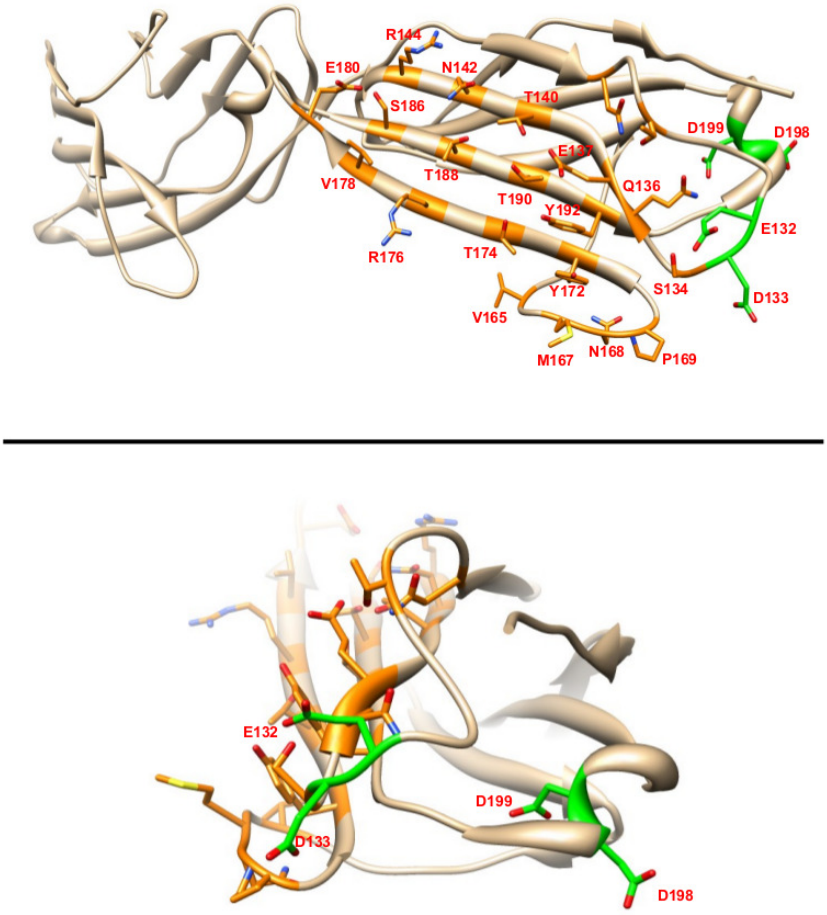


Figure 2.1: Structure expérimentale des deux premiers domaines N-terminaux de Lu. Le site de liaison prédit est délimité par les résidus en orange et par le résidu Glu132 sur le domaine D2. Les résidus Glu132, Asp133, Asp198 et Asp199 qui sont importants pour la liaison de la Laminine selon les expériences de mutagenèse sont représentés en vert [6].

Les serveurs Web suivants ont été utilisés : eFindSite [48], Surfnet [49], FT-Site [46] et Site-Hound [47]. Seuls eFindSite, SiteHound et Surfnet ont prédit des sites d'interaction en accord avec les études de mutagenèse [54] et les interactions CD80-CTLA-4 observées dans la structure cristallographique (code PDB 1I8L). Les serveurs eFindSite et SiteHound ont prédit les résidus Tyr31, Arg29, Met43, Ser44, Val83, Leu85 et Thr41 comme ayant un potentiel de liaison à un partenaire. En plus de ceux-là, les sondes hydroxyles et phosphate-oxygène de SiteHound ont prédit Asn48 et Trp50 comme résidus importants (et connus comme tels selon les expériences de mutagenèse) et l'ensemble des résidus en interaction avec CTLA-4 dans la structure cristallographique [103] à l'exception des résidus Val83, Leu85 et Leu97. Le site de liaison prédit par ces serveurs Web est montré dans la Figure 2.2a. Malgré le faible taux d'identité de séquences de 16 % entre le domaine C1-set de Lu et le domaine V de CD80, l'analyse de ces deux domaines révèle des similitudes en termes de topologie et de polarité (Figure 2.2). En effet, pour les deux protéines, les serveurs de prédiction définissent un site de liaison plat formé de brins bêta. De plus, ces régions ont des polarités similaires : des résidus pouvant former des liaisons hydrogène sont retrouvés en proportion importante (Ser, Thr, Asn, Gln), ainsi que des résidus hydrophobes (aromatique et aliphatique) et des résidus chargés (Arg, Glu). Dans les parties (c) à (e) de la Figure 2.2, les squelettes peptidiques des domaines V de CD80 (en vert) et C1-set de Lu (en kaki) ont été superposés. Tout d'abord, on peut observer que les brins bêta respectifs au centre des sites de liaison prédits sont presque parfaitement superposés (dans le cas de la Figure 2.2e, les deux structures ont été légèrement décalées pour une question de clarté). De plus, les résidus chargés se trouvent essentiellement à la périphérie du centre du site de liaison de chacune des protéines Lu et CD80 (Figure 2.2c, d) alors que les résidus hydrophobes ainsi que des résidus Thr se trouvent dans le centre du site de liaison des deux protéines (Figure 2.2e). Le centre du site de liaison prédit pour CD80 et Lu est encerclé en violet dans la Figure 2.2. Sachant que la pertinence d'un protocole de scoring est dépendant

du système protéine-ligand, les similarités observées en termes de topologies et de polarités entre les deux domaines respectifs confortent l'hypothèse qu'un protocole capable de reproduire les résultats expérimentaux de CD80 donnera aussi des résultats fiables pour Lu. Enfin, CD80 a été choisi puisqu'il possède plusieurs dizaine de ligands avec des constantes d'affinités connues (IC50) [50].

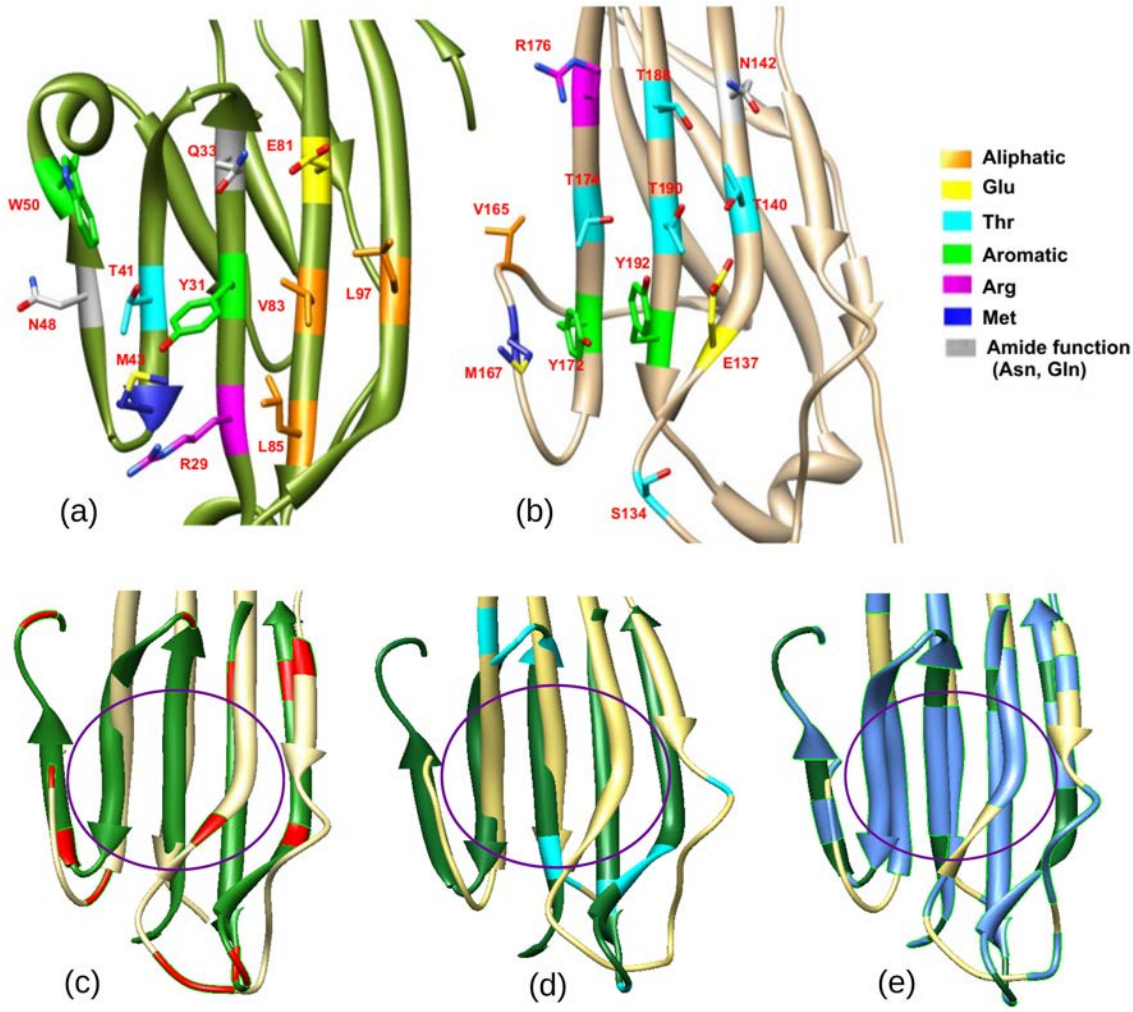


Figure 2.2: Comparaison entre les sites de liaisons prédits sur (a) le domaine V de CD80 et (b) le domaine D2 (C1-set) de Lu. La superposition de ces deux domaines (c à d) montre qu'ils partagent une même structure plane et que la distribution des résidus sur les sites de liaison de chacun des domaines est similaire avec (i) des résidus chargés négativement (en rouge) et positivement (en bleu) (respectivement (c) et (d)) à la périphérie du centre de leur site de liaison (entouré en violet) ; (ii) des résidus hydrophobes (en bleu) au centre de leur site de liaison (e).

Chapitre 3

Élaboration et validation d'un protocole de scoring

Dans ce chapitre, nous présenterons d'abord les 17 ligands de CD80 que nous avons choisi d'étudier ainsi que la pose que nous avons sélectionnée pour ces ligands sur la protéine. Puis nous présenterons les différentes étapes qui ont permis d'élaborer, de perfectionner et de valider le protocole de scoring sur le système CD80-ligand.

3.1 Description de la pose sélectionnée pour les ligands 1 à 17 sur CD80

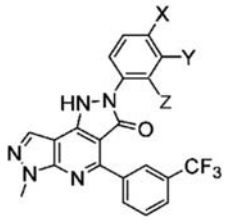
Parmi les 17 ligands de CD80 étudiés ici, 16 possèdent un même cœur 6-methyl-4-(3-[trifluoromethyl]phenyl) – dihydrodipyrazolopyridinone et diffèrent par des groupements plus ou moins gros (ligands **1** à **16** du Tableau 3.1). Le 17ème ligand (ligand **17**, Tableau 3.1), possède un cœur légèrement différent dans lequel un cycle pyrazole est manquant. Les ligands seront divisés en plusieurs parties dans la suite du manuscrit : le cœur (entouré en rose), la partie nord (entourée en bleu) et la partie sud (entourée en vert)

(Figure 3.1). La partie nord est constituée d'un cycle phényl lié à des halogènes (atomes de fluor ou de chlore) et/ou à des groupements plus imposants tels que COO^- et des groupements cycliques. Enfin la partie sud est constituée d'un groupement phényl lié à un groupement CF_3 ou à un groupement NO_2, NH_2 ou O-Me. Le choix de ces 17 inhibiteurs de CD80 s'est fait de manière à avoir un maximum de conditions à satisfaire, ce qui a conduit à sélectionner les inhibiteurs les plus complexes structuralement parmi ceux qui sont disponibles. Ces inhibiteurs ont également été choisis de manière à avoir une large gamme de valeurs de constantes de liaison. Puisque ces ligands possèdent des structures similaires, il est très probable qu'ils adoptent la même position (pose) sur CD80 [104].

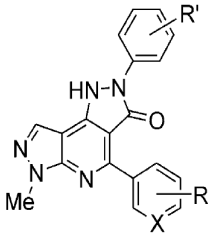
La pose de ces ligands sur CD80 n'est pas connue. De plus, dans le cas de cette protéine, le site prédit d'interaction définit une large région de la surface de cette protéine et correspond à une quantité considérable de poses possibles pour un ligand donné. Une analyse parallèle d'une cinquantaine de poses pour chacun des ligands choisis pour cette étape et de la relation entre la structure de plusieurs autres ligands et leur affinité déterminée expérimentalement a permis d'identifier les poses plausibles de ces ligands sur le domaine V de CD80. L'analyse s'est également basée, en plus de la relation structure-affinité mentionnée, sur des conditions de géométrie des deux partenaires d'interaction qui favorisent *a priori* une énergie d'interaction satisfaisante dénoncées comme «conditions optimales d'interaction» dans le reste du manuscrit c'est-à-dire des conditions qui permettent une correspondance entre groupements hydrophiles, d'une part, et entre groupements hydrophobes d'autre part, un nombre maximum de liaisons hydrogène entre la protéine et le ligand et un non-enfouissement des groupements donneurs et accepteurs de liaisons hydrogène de chacun des partenaires.

La pose que nous avons sélectionnée pour les ligands étudiés ici est montrée dans la Figure 3.2. Deux études nous ont permis de sélectionner cette pose (pose la plus probable des ligands sur CD80) : (i) l'étude de Erbe et al qui montre que les mutants W50A

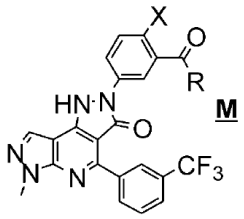
Tableau 3.1: Structures des composés utilisés pour la validation du protocole de scoring. Les structures des ligands **1** à **16** diffèrent par les groupements X, Y, Z, R ou R'. Les ligands **9** à **12** possèdent des groupements R en position méta ou para (respectivement M et P). Les ligands **10** et **11** ont le même groupement R en position méta M. Le ligand **17** possède un coeur différent de celui des autres ligands.



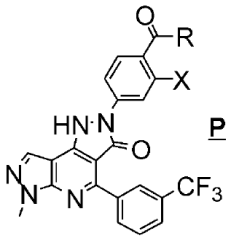
Ligand	Y	X	Z	IC50 (nM)
1	H	Cl	H	60 ± 17
2	COO^-	F	H	70 ± 14
3	Cl	COO^-	H	16 ± 2
4	F	COO^-	H	210 ± 20
5	F	H	H	20 ± 3
6	H	F	F	1000 ± 250
7	H	H	F	300 ± 80
8	COO^-	Cl	H	7 ± 3



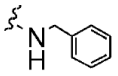
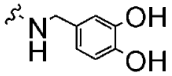
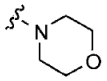
Ligand	R'	X	R	IC50 (nM)
13	3-F	CH	3-OMe	82 ± 9
14	3-F	CH	3-NO_2	4 ± 1
15	4-F	CH	3-NO_2	50 ± 13
16	4-F	CH	3-NH_2	400 ± 100

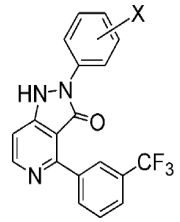


M



P

Ligand	R	Position de R	X	IC50 (nM)
9		M	H	120 ± 75
10		M	H	9 ± 1
11	Idem que 10	M	Cl	22 ± 4
12		P	H	13 ± 4



Ligand	X	IC50 (nM)
17	3-F	1300 ± 100

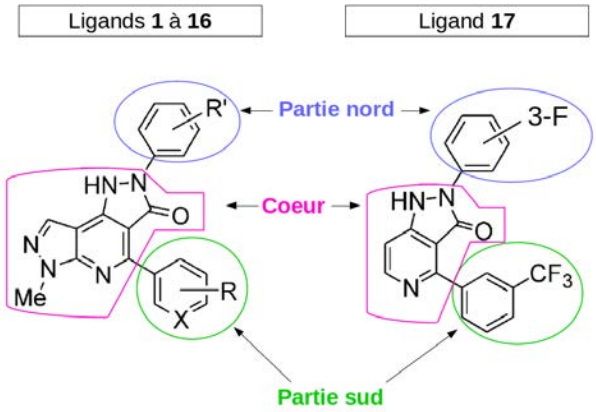


Figure 3.1: Les ligands sont composés de trois parties: la partie nord, la partie sud et le coeur qui sont respectivement entourés en bleu, vert et rose. Le ligand **17** possède un cœur différent des ligands **1** à **16**.

et N48K diminuent fortement l'affinité des inhibiteurs pour CD80 (Figure 3.2) [54] et (ii) l'étude de Green et al qui permet la relation entre structure et affinité de plus de 70 ligands [50]. Cette dernière étude a montré que la liaison NH du cœur des ligands conditionne une forte affinité, ce qui suggère que celle-ci doit être impliquée dans une liaison hydrogène avec CD80. En effet, lorsque cet azote est méthylé (NH transformé en N-CH_3), on constate une très forte diminution de l'affinité des ligands pour la protéine : l'azote ne pouvant probablement plus donner d'hydrogène à un atome électronégatif accepteur O ou N de la protéine pour former une liaison hydrogène. Par conséquent les ligands doivent adopter une pose dans laquelle (i) ils sont en interaction avec les résidus Asn48 et Trp50 et (ii) la liaison NH du cœur des ligands intervient dans une liaison hydrogène avec CD80. Ces conditions sont respectées dans la pose sélectionnée (Figure 3.2) : (i) les ligands font des contacts $\pi - \pi$ avec le résidu Trp50 (configuration perpendiculaire ou T-shape) et des liaisons hydrogène et/ou des contacts de van der Waals avec le résidu Asn48, (ii) la liaison NH du cœur des ligands forme une liaison hydrogène avec le résidu Gln33. Il est intéressant par ailleurs de noter que ce résidu Gln33

est impliqué dans une liaison hydrogène avec CTLA-4 dans la structure cristallographique correspondante (code PDB 1I8L) [103].

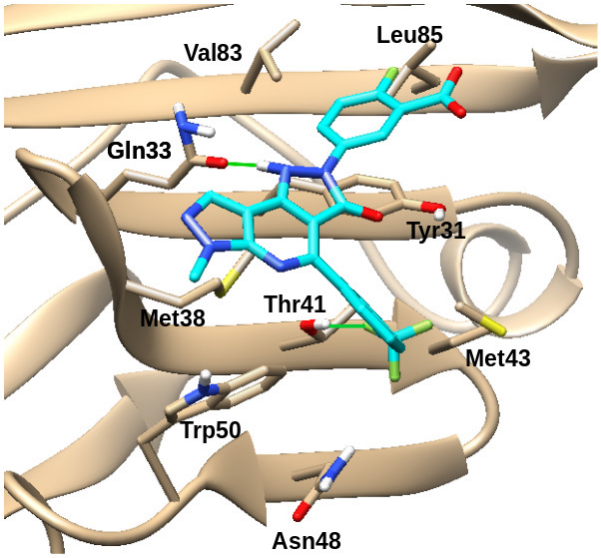


Figure 3.2: Pose commune aux ligands **1** à **17** sur CD80 (exemple du ligand **8**). Les liaisons hydrogène sont représentées en trait plein vert. La liaison hydrogène entre le cœur des ligands et le résidu Gln33 est présente dans tous les complexes étudiés.

De plus, dans cette pose, plusieurs des 17 ligands forment aussi une liaison hydrogène avec le résidu Thr41, notamment celle impliquant le groupement CF_3 de la partie sud des ligands. Enfin, on peut observer que les résidus Tyr31, Lys37, Met38 et Val83 entrent aussi en contact avec les ligands. Ces interactions seront plus longuement décrites avec des calculs de chimie quantique dans le chapitre 4.

La pose décrite ici (pose sélectionnée) respecte les conditions fixées précédemment d'interaction entre CD80 et les ligands, permettant des conditions optimales d'interaction p. 76). En effet, dans cette pose (i) il y a une bonne correspondance entre les groupements hydrophiles de CD80 et ceux des ligands, d'une part, et entre les groupements hydrophobes de CD80 et ceux des ligands d'autre part, (ii) au moins deux liaisons hy-

drogène sont formées entre CD80 et les ligands (avec les résidus Gln33 et Thr41) et (iii) aucun groupement donneur ou accepteur de liaison hydrogène de la protéine ou des ligands n'est enfoui.

3.2 Recherche du protocole de scoring

La recherche du protocole de scoring est une étape importante qui permet de valider le protocole qui sera appliqué à Lu. Nous considérerons que le protocole est validé lorsque (i) une corrélation raisonnable est obtenue entre les valeurs expérimentales de IC50 et les énergies d'interaction calculées (scores) et (ii) le ligand lié à la protéine est stable pendant la trajectoire de la dynamique moléculaire. Idéalement, une relation linéaire devrait relier les valeurs expérimentales d'affinité à celles calculées, mais cette condition n'est presque jamais remplie. En effet, il est souvent difficile d'obtenir un coefficient de corrélation entre les scores et les affinités expérimentales qui dépasse 0.71 [32], et cela, peu importe la méthode de calcul (fonction de scoring secondaire) utilisée.

Comme indiqué dans l'introduction de ce manuscrit, le protocole de criblage qui sera appliqué à Lu après sa validation sur CD80 contient les étapes (b) à (d) suivantes : (b) docking des ligands sur la protéine, (c) relaxation des complexes protéine-ligand issus du docking avec des simulations DM et (d) calcul de scores afin d'évaluer l'affinité des ligands pour CD80. Les étapes (c) et (d) constituent le protocole de scoring. La procédure de recherche des meilleures conditions et des meilleurs paramètres qui nous ont permis d'obtenir un protocole de scoring robuste sont expliquées dans le paragraphe qui suit et est illustrée sur la Figure 3.3.

L'étape (d) de calcul de scores est réalisée uniquement si les trajectoires obtenues à l'étape (c) de simulation DM (étape de relaxation) respectent les conditions suivantes (dénoncées "conditions optimales de simulations DM" dans la suite du manuscrit) : (i) le ligand est stable au cours de la trajectoire, (ii) les conditions d'interaction entre CD80 et

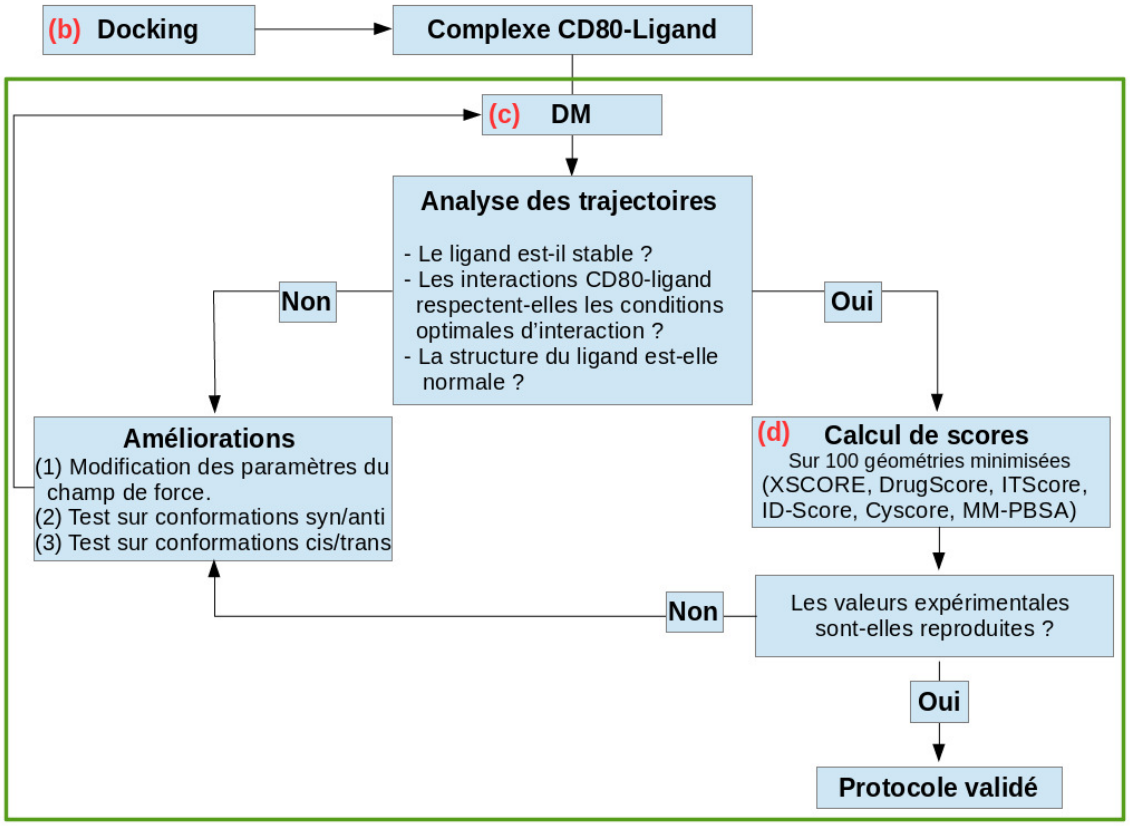


Figure 3.3: Procédure de recherche du protocole de scoring (étapes (c) et (d), encadré vert) sur CD80 avec 17 inhibiteurs d'affinités expérimentales connues. Les étapes '**Analyse des trajectoires**' et '**Améliorations**' sont des étapes intermédiaires qui ont été utilisées pour perfectionner le protocole. Les conditions optimales d'interaction CD80-ligand sont celles qui sont présentées à la page 76.

le ligand respectent les conditions optimales d'interaction (p. 76) au cours de la trajectoire et (iii) aucune anomalie structurale n'est observée dans la géométrie du ligand au cours de la trajectoire. Un exemple d'anomalie structurale dans la géométrie d'un ligand est la non-planéité des cycles aromatiques. Si une de ces conditions n'est pas respectée, plusieurs améliorations sont réalisées telles que des modifications des paramètres de champ de force du ligand, c'est-à-dire que nous avons modifié les paramètres de champ de force des ligands afin d'améliorer les contacts de ces ligands avec CD80. Pour certains ligands, nous avons aussi tenté d'améliorer les résultats en utilisant différentes conformations (syn/anti, cis/trans) qui seront présentées dans la sous-section 3.2.2 (p. 91). Ces améliorations ont été réalisées jusqu'à ce que les trajectoires soient validées. Le nombre de trajectoires calculé ainsi que les détails des simulations DM seront donnés dans la sous-section 3.2.1.

À l'étape (d) de calcul de scores, différentes méthodes de calcul sont testées et les améliorations citées ci-dessus sont réalisés jusqu'à ce que les scores calculés reproduisent les affinités expérimentales. Des calculs de scores ont aussi été réalisés sur différentes conformations possibles pour certains ligands (conformations syn/anti, cis/trans). Dans les sous-sections suivantes, nous présenterons les méthodes utilisées ainsi que les différentes améliorations que nous avons réalisées afin (i) d'obtenir des trajectoires qui respectent les conditions citées ci-dessus à l'étape de relaxation (étape (c) de la Figure 3, p. 17) et (ii) de reproduire les affinités expérimentales des ligands à l'étape d'évaluation des énergies d'interactions protéine-ligand (étape (d) de la Figure 3, p. 17).

3.2.1 Étape de relaxation en présence d'eau explicite

Pour chacun des complexes CD80-ligand, des simulations DM ont permis de relaxer le système (étape (c), Figure 3.3). La trajectoire d'une simulation DM dépend des conditions initiales telles que la conformation de départ, les différents paramètres utilisés

et le jeu de vitesses initiales. En changeant une de ces conditions citées ci-dessus, on explorera une autre zone de l'espace conformationel : une trajectoire différente sera donc obtenue. Afin d'explorer au mieux l'espace conformationnel de nos systèmes CD80-ligand, 10 simulations DM qui diffèrent par leurs vitesses initiales ont été réalisées pour chaque complexe. Les simulations DM ont été réalisées en présence d'eau explicite avec la méthode de dynamique moléculaire aux limites stochastiques (SBMD). Dans la suite de cette sous-section, nous justifierons le choix de la méthode de DM utilisée pour l'étape de relaxation puis nous présenterons les différentes modifications du champ de force qui ont permis d'obtenir des trajectoires qui respectent les conditions optimales de simulations DM (p. 80).

Choix de la méthode de dynamique moléculaire pour l'étape de relaxation

Comme présenté dans la Figure 3.3 (p. 81), l'étape de calcul de score est réalisée uniquement si l'étape de DM est validée. La qualité des résultats de scoring est donc fortement dépendante de la qualité des trajectoires obtenues à l'étape de relaxation puisque les calculs de scores sont réalisés sur les géométries issues de ces trajectoires. Il faut donc s'assurer que la méthode de dynamique moléculaire utilisée reproduise au mieux le comportement des molécules dont les mouvements dépendent aussi du milieu dans lequel ces molécules sont simulées. Afin de choisir la méthode de DM la plus adaptée à notre système et à la méthode de criblage (dans le cas de Lu), nous avons réalisé des simulations DM (i) sans eau, (ii) en présence d'eau implicite (simulations LD) et (iii) en présence d'eau explicite (simulations SBMD) sur les complexes formés entre CD80 et un sous-ensemble de ligands (ligands **1** à **8** du Tableau 3.1, p. 77) sur lequel nous nous sommes focalisés pour l'optimisation de la procédure.

Nous avons observé une instabilité des ligands dans les trajectoires issues des simulations sans eau, notamment pour les ligands chargés **2** et **8** pour lesquels les groupements COO^- interagissent fortement avec les groupements NH_3^+ de la protéine. Ce phénomène

n'est pas observé dans les simulations en présence d'eau (implicite ou explicite) car l'eau atténue ces interactions charge-charge en faisant écran.

L'analyse des trajectoires obtenues avec chacune des méthodes LD et SBMD a permis d'évaluer la stabilité de chacun des ligands au cours des simulations, leur géométrie ainsi que les interactions établies par ceux-ci. De cette analyse, nous avons constaté que les paramètres initiaux du champ de force CGENFF n'étaient pas optimaux sur le cœur des ligands, ce qui conduisait à une instabilité de ces ligands sur la protéine pendant les simulations DM. Une analyse du cœur des ligands **1** à **8** a permis de constater que l'instabilité de ces ligands au cours des trajectoires était due à : (i) des constantes de force de torsion incorrectes au niveau du cœur des ligands, (ii) une valeur d'angle de valence $\angle NNH$ adjacent au groupement CO du cœur des ligands qui n'était pas en accord avec celle des structures de ces ligands optimisées par chimie quantique au niveau HF/6-31G*, et (iii) une mauvaise qualité des charges des atomes qui constituent le cœur des ligands. Les points (i) à (iii) seront développés dans le paragraphe ci-après.

Dans le cas (i) des constantes de force de torsion incorrectes, on observait que le cœur des ligands n'était pas plan, ce qui paraissait inhabituel puisqu'il est constitué de cycles aromatiques. La planéité du cœur a été vérifiée par calcul de chimie quantique. La conformation inhabituelle observée pour cette partie des ligands avait pour conséquence une perte de contacts de cette même partie (le cœur des ligands) avec la protéine au cours des simulations. Afin d'avoir un cœur plan pour tous les ligands et ainsi améliorer leurs contacts avec la protéine, nous avons ajusté les constantes de forces de torsion des quatre angles dièdres $\angle ABCF$, $\angle DCBI$, $\angle CFGJ$ et $\angle QGFK$ (montrés sur la Figure 3.4) qui constituent le cœur des ligands.

Dans le cas (ii) de l'angle de valence $\angle NNH$ incorrect, la liaison NH du cycle pyrazole du cœur des ligands était dans le plan du cycle (à gauche sur la Figure 3.5) : une telle orientation ne permettait pas la formation d'une liaison hydrogène optimale entre cette liaison NH et l'atome d'oxygène (accepteur) de la chaîne latérale du résidu Gln33. Afin

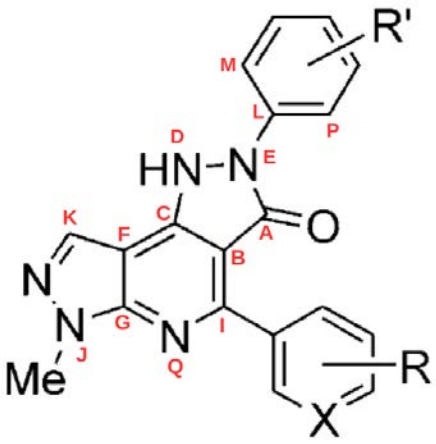


Figure 3.4: Les constantes de force de torsion ont été modifiées pour les angles dièdres $\angle ABCF$, $\angle DCBI$, $\angle CFGJ$ et $\angle QGFK$ qui constituent le cœur des ligands **1** à **16**. Pour le ligand **17**, seule la constante de force de torsion de l'angle $\angle DCBI$ a été ajustée. Les constantes de force de torsion des angles dièdres $\angle DELP$ et $\angle AELM$ qui font la jonction entre le cœur et la partie nord des ligands **1** à **17** a été modifiée.

de vérifier l'orientation de cette liaison NH, nous avons minimisé le ligand **8** (Figure 3.2, p. 79) avec le programme Gaussian09 en utilisant la méthode *ab initio* HF-6-31G*. Selon ces calculs, la liaison NH n'est pas plane, mais elle pointe vers l'oxygène de la chaîne latérale du résidu Gln33 (Figure 3.5). Nous avons donc ajusté l'angle de valence $\angle NNH$ du cycle pyrazole de façon à être en accord avec les calculs HF.

De plus, nous avons constaté que les charges des atomes d'azote qui constituent le cœur des ligands au niveau CGENFF n'étaient pas en accord avec celles calculées au niveau MP2/6-31G* (cas (iii) ci-dessus). Nous avons alors ajusté les charges des atomes d'azotes qui constituent le cœur des ligands de façon à respecter les tendances observées dans les résultats de chimie quantique. Par exemple, pour les atomes C4, N3 et N2 qui sont respectivement le carbone du groupement méthyle du coeur, l'azote lié à ce groupement méthyle et l'azote lié à N2, les charges respectives de ces atomes étaient de +0.20, -0.55 et +0.22 dans le champ de force CGENFF. Ces charges étaient très différentes de celles qui ont été obtenues par chimie quantique et qui sont de +0.73, -0.53 et -0.25 pour les atomes respectifs C4, N3 et N2. Nous avons alors ajusté les charges de ces atomes dans le champ de force CGENFF de la façon suivante : +0.75, -0.38 et -0.27. Ces charges modifiées suivent la même tendance que les charges calculées par chimie quantique avec, notamment la charge de l'atome N2 qui est devenue négative. L'ajustement de l'angle de valence $\angle NNH$ du cycle pyrazole et des charges sur les atomes d'azotes du cœur des ligands a favorisé la formation d'une liaison hydrogène forte qui est maintenue tout le long des trajectoires entre le cœur des ligands et Gln33.

Les modifications du champ de force sur le cœur des ligands ont permis d'améliorer les contacts des ligands avec la protéine CD80 au cours des simulations LD et SBMD.

Une évaluation des affinités des ligands avec la fonction de scoring secondaire XS-CORE a permis de comparer les résultats obtenus avec la méthode SBMD (en présence d'eau explicite) à ceux obtenus avec la méthode LD (en présence d'eau implicite) de la façon suivante : (i) 100 géométries sélectionnées et qui sont associées à chacun des lig-

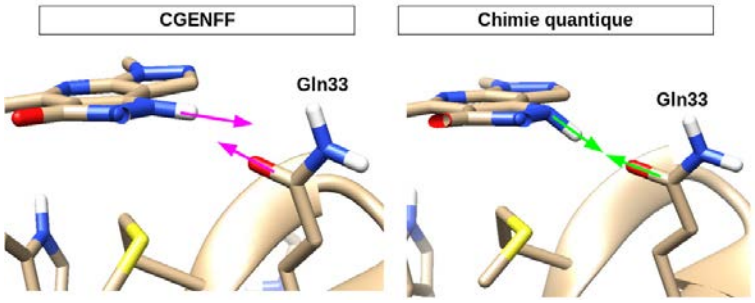


Figure 3.5: Orientation de la liaison NH du cœur des ligands après optimisation des géométries. Selon les paramètres de champ de force CGENFF (à gauche), la liaison NH ne pointe pas correctement vers l'atome d'oxygène de la chaîne latérale du résidu Gln33 alors que selon les calculs de chimie quantique (à droite), cette liaison NH pointe vers l'atome d'oxygène de la chaîne latérale de Gln33.

ands ont été extraites et minimisées avec le champ de force CHARMM36 ; (ii) des calculs d'énergie d'interaction ont été réalisés sur chacune des géométries avec la méthode de calcul HMScore (de la fonction de scoring XSCORE, p. 51) pour chaque ligand ; (iii) une moyenne de l'énergie d'interaction avec la protéine a été calculée à partir des 100 valeurs obtenues à l'étape (ii) et (iv) les énergies moyennes résultantes pour chacun des 8 ligands ont été comparées aux affinités expérimentales IC50.

Les deux méthodes de dynamique moléculaire LD et SBMD donnent de bons résultats avec respectivement des coefficients de corrélations de 0.83 et 0.91 (Figure 3.7). Cependant, on constate que le classement des affinités calculées pour les ligands **1** à **8** est moins bon lorsqu'on utilise la méthode de dynamique LD plutôt que la méthode de dynamique SBMD. En effet, le classement obtenu avec la méthode SBMD pour les ligands **1** à **8** est similaire au classement de ces mêmes ligands selon leur affinité expérimentale (ce qui n'est pas du tout le cas pour le classement obtenu avec la méthode LD), avec seulement une permutation du rang des ligands **1** et **5** et une surévaluation de l'affinité du ligand **7**. La méthode SBMD a donc été utilisée dans l'étape de relaxation du protocole de scoring

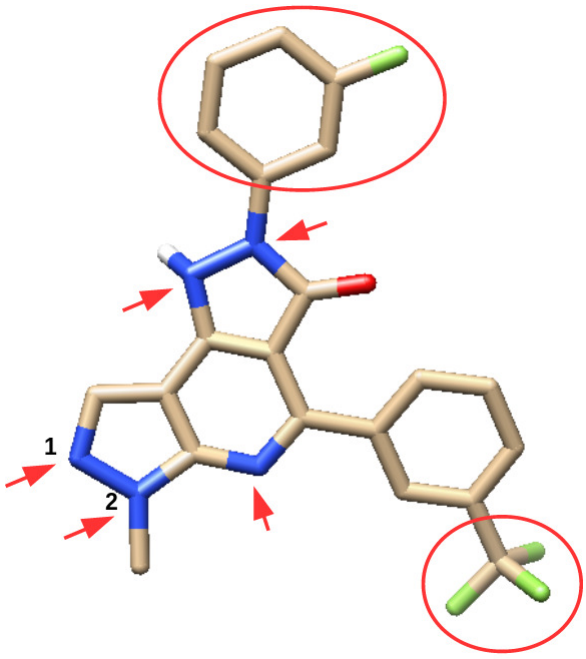
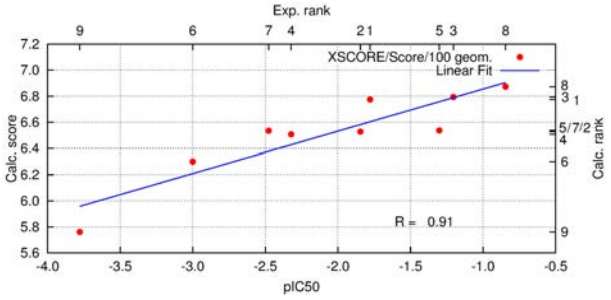


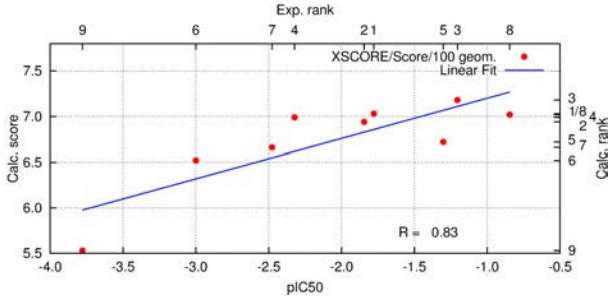
Figure 3.6: Les groupements et atomes pour lesquels des modifications de charges ont été réalisées en accord avec les résultats de chimie quantique sont entourés en rouge (illustré avec le ligand **5** pour lequel les atomes de fluor, d'azote et d'oxygène sont respectivement montrés en vert, bleu et rouge). Pour les ligands **14** et **15**, il faut remplacer le groupement CF_3 de la figure par un groupement NO_2. Les atomes d'azotes du cœur des ligands pour lesquels les charges ont aussi été modifiées sont pointés avec des flèches rouges. Pour le ligand **17**, les atomes d'azote 1 et 2 n'ont pas été modifiés puisqu'ils font partie du cycle pyrazole manquant dans le cœur de ce ligand.

(étape (c)).

Figure 3.7: Corrélation entre les affinités expérimentales et les affinités calculées avec la fonction de scoring secondaire XSCORE sur 100 géométries extraites de façon équidistante sur les trajectoires obtenues avec la méthode (a) SBMD et (b) LD. Les calculs ont été réalisés pour les ligands **1** à **8**.



(a)



(b)

Amélioration de la stabilité des ligands sur CD80 dans les simulations SBMD

Dans la sous-section ci-dessus, nous avons réalisé des modifications du champ de force pour le cœur des ligands uniquement. Afin d'améliorer la stabilité des trajectoires (obtenues avec la méthode SBMD) en considérant, cette fois-ci, les différentes parties des ligands, nous avons analysé et amélioré les paramètres du champ de force CGENFF pour les groupements qui constituent la partie nord et la partie sud des ligands. Pour cela, nous avons étendu notre étude aux ligands **9** et **17**.

Nous avons constaté que les paramètres initiaux du champ de force CGENFF sur les parties nord et sud des ligands n'étaient pas optimaux, ce qui conduisait à une instabilité des ligands sur la protéine pendant les simulations SBMD. Cette instabilité était due à : (i) une mauvaise qualité des charges des atomes des ligands qui constituent leurs parties nord et sud (Figure 3.1, p. 78) et (ii) des constantes de force de torsion incorrectes pour les angles dièdres qui font la liaison entre le cœur des ligands et leur partie nord.

Dans le cas (i) ci-dessus, les charges des ligands au niveau CGENFF n'étaient pas en accord avec celles calculées au niveau MP2/6-31G* pour notamment (1) le groupement CF_3 de la partie sud des ligands **1** à **12** et du ligand **17**, (2) le groupement NO_2 de la partie sud des ligands **14** et **15** et (3) le groupement fluorophényl de la partie nord du ligand **2**, des ligands **4** à **7** et des ligands **13** à **17** (Tableau 3.1, p. 77). Les groupements et/ou atomes pour lesquels les charges ont été modifiées peuvent être visualisés dans la Figure 3.6 (p. 88). Une mauvaise répartition des charges sur ces groupements empêchait la formation de liaisons hydrogène entre la partie sud des ligands et le résidu Thr41 (groupement CF_3 des ligands **1** à **12** et **17** et groupement NO_2 des ligands **14** et **15**) et conduisait des contacts insuffisants entre les groupements fluorophényl de la partie nord des ligands **2**, **4**, **5**, **6**, **7**, **13**, **14**, **15**, **16**, **17** et le résidu Tyr31.

Dans le cas du fluorophényl, nous avons en effet constaté que le champ de force CGENFF ne prenait pas en compte la polarisation dans l'espace que provoque l'atome de fluor, du fait de son électronégativité, sur les liaisons CH qui l'entourent. Les liaisons CH qui entourent l'atome de fluor doivent porter des charges absolues plus élevées que celles des liaisons CH plus éloignées. Afin de résoudre ces problèmes, nous avons ajusté les charges de chaque ligand de façon à respecter les tendances observées dans les résultats de chimie quantique (Figure 3.6 et section 1.3, p. 30). Ces charges ont été modifiées suivant une règle de trois. Pour cela, nous nous sommes basés sur la charge de l'hydrogène aromatique qui est de +0.11 dans le champ de force CGENFF et de +0.21 selon les calculs des chimie quantique (charges Mülliken). Par exemple pour ajuster la charge de l'atome de fluor du groupement fluorophényl qui constitue la partie du ligand **2**, nous avons fait l'opération suivante : $Charge_F$ * (0.11 / 0.21) avec $Charge_F$ qui est la charge de ce fluor dans le champ de force CGENFF et doit être ajusté. La même opération a été réalisée pour les groupements CF_3 et chlorophényl. Pour le groupement NO_2, nous avons utilisé les mêmes charges que celles observées pour des ligands déjà paramétrés et qui ont un groupement similaire (c'est-à-dire un groupement qui ressemble à la partie

sud des ligands **14** et **15**). Pour le groupement COO^- qui se trouve dans la partie nord des ligands **2** - **4** et **8**, nous avons gardé la charge classique donnée dans le champ de force CGENFF.

Enfin, nous avons constaté que la partie nord des ligands avait des mouvements de forte amplitude pendant des simulations SBMD. Ce phénomène était dû à des constantes de force de torsion incorrectes pour les angles dièdres $\angle DELP$ et $\angle AELM$ qui font la jonction entre le cœur et la partie nord des ligands (Figure 3.4). Nous avons donc modifié les constantes de force de torsion de ces angles dièdres.

L'ensemble des modifications du champ de force présentées dans cette sous-section pour les différentes parties des ligands (cœur, nord et sud) ont permis d'améliorer les contacts des ligands avec la protéine CD80 au cours des simulations DM en présence d'eau explicite et d'améliorer ainsi la stabilité des trajectoires qui ont été utilisées dans l'étape (d) de calcul de scores (ou étape d'évaluation d'énergies d'interaction).

3.2.2 Évaluation des énergies d'interaction

Lorsque les trajectoires respectent toutes les conditions optimales de simulations DM (définies p. 80), 100 géométries minimisées avec CHARMM sont extraites de façon équidistante sur la plus stable des trajectoires de chaque ligand. Puis les énergies de ces géométries sont évaluées par les différentes méthodes XSCORE [27], DrugScore [105], ITScore [30], ID-Score [29], Cyscore [28] et MM-PBSA [32]. Pour chaque méthode testée et pour chaque ligand, nous avons calculé un score moyen obtenu avec le score de chacune des 100 géométries extraites de la trajectoire. Si ces scores moyens sont corrélés aux affinités expérimentales des ligands, alors le protocole de scoring est validé avec la méthode de calcul (fonction de scoring) correspondante. Les résultats obtenus avec chaque méthode testée seront présentés et commentés dans le paragraphe qui suit à l'exception de ceux obtenus avec les méthodes Cyscore et Drugscore pour lesquels les

affinités calculées ne sont pas corrélées avec les affinités expérimentales (coefficient de corrélation inférieur à 0.4).

Avant de commenter les résultats obtenus avec chaque méthode de calculs, revenons sur les structures de quelques ligands. Les ligands **2**, **3**, **4**, **5**, **8**, **13**, **14** et **17** (Figure 3.1, p. 77) peuvent adopter une conformation syn ou anti par rapport au résidu Gln33. En effet, ils possèdent des atomes de fluor, de chlore ou un groupement COO^- en position méta du cycle phényl qui constitue leur partie nord et qui peuvent pointer vers le résidu Gln33 (conformation syn) ou de l'autre côté (conformation anti) (Figure 3.8). Les ligands **6** et **7** peuvent aussi adopter une conformation syn ou anti selon que leur atome de fluor placé en ortho sur le cycle de leur partie nord pointe respectivement vers le résidu Gln33 ou de l'autre côté. Lors de l'analyse des résultats qui seront présentés plus loin dans cette sous-section, nous avons constaté que les scores calculés avec toutes les méthodes citées ci-dessus étaient mieux corrélés avec les expériences lorsque les ligands adoptaient une conformation syn plutôt qu'une conformation anti. En effet, dans la conformation syn, on observe des liaisons hydrogène et des liaisons halogène entre le résidu Gln33 et respectivement les groupements COO^- (Figure 3.8, b) et les halogènes fluor et chlore des ligands (Figures 3.8, d). En revanche, peu d'interaction sont observés entre ces groupements et la protéine lorsque les ligands adoptent une conformation anti (Figure 3.8, a et c).

Les ligands **9**, **10** et **11** (Figure 3.1 de la page 77) sont de plus haut poids moléculaire que les autres ligands du fait de la taille de leur groupement R qui constitue leur partie nord. Ces groupements R sont liés au cycle phényl de la partie nord de ces ligands grâce à une liaison amide se trouvant en position méta de ce cycle. Du fait de la présence de R en position méta, ces ligands peuvent aussi adopter une conformation syn ou anti par rapport au résidu Gln33. De plus, deux autres conformations peuvent être adoptées par ces ligands de haut poids moléculaire : la conformation cis lorsque les liaisons CO et NH de la liaison amide pointent dans la même direction et la conformation trans lorsque

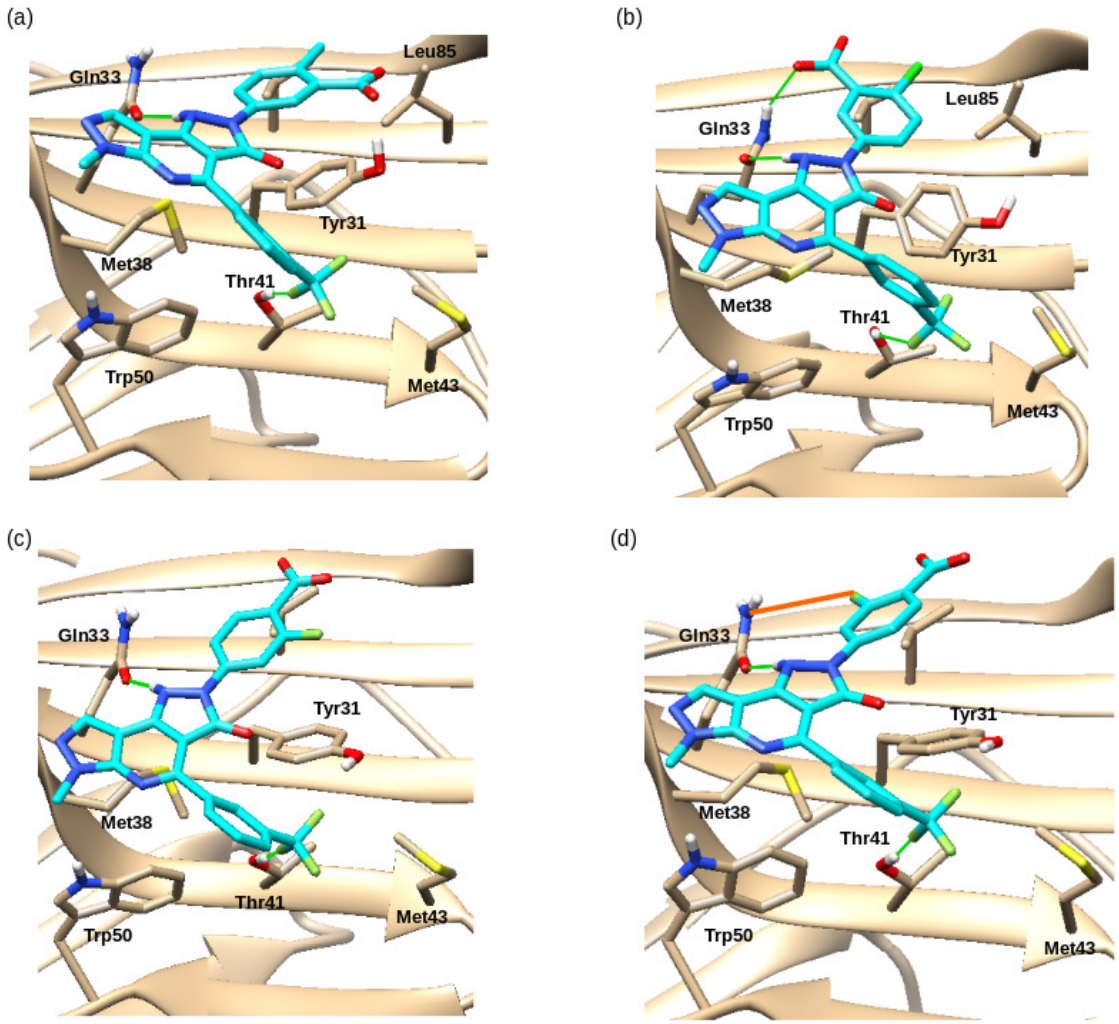


Figure 3.8: Conformations syn et anti des ligands **4** et **8**. Les figures (a) et (b) montrent le ligand **8** respectivement dans les conformations anti et syn. Les figures (c) et (d) montrent le ligand **4** respectivement dans les conformations anti et syn. Dans la conformation syn, le groupement COO^- du ligand **8** (Figure (b)) et l'atome de fluor de la partie nord du ligand **4** (Figure (d)) pointent vers le résidu Gln33 alors que dans la conformation anti ces groupements pointent de l'autre côté (a et c). Les liaisons hydrogène possibles dans chacune de ces conformations sont représentées en trait vert. La liaison halogène formée entre le fluor de la partie nord du ligand **4** dans la conformation syn et Gln33 est montré en trait orange (Figure (d)).

ces mêmes liaisons pointent dans des directions opposées (Figure 3.9). Dans les calculs réalisés avec toutes les méthodes citées précédemment, il n'y avait pas de différences significatives en terme d'énergie comparée ou en terme de corrélation comparée entre les scores obtenus à partir des géométries dans lesquelles les ligands **9**, **10** et **11** adoptent une conformation cis et les scores obtenus à partir des géométries dans lesquelles ces ligands adoptent une conformation trans. Nous avons donc décidé de comparer les affinités expérimentales aux affinités calculées (avec les différentes méthodes de calcul testées) à partir des géométries issues des simulations dans lesquelles les ligands **9**, **10** et **11** adoptent une conformation cis d'une part et les ligands **2**, **3**, **4**, **5**, **6**, **7**, **8**, **13**, **14** et **17** adoptent une conformation syn par rapport au résidu Gln33 d'autre part. Les corrélations entre les affinités calculées et les affinités expérimentales sont présentées dans les Figures 3.10 et 3.12.

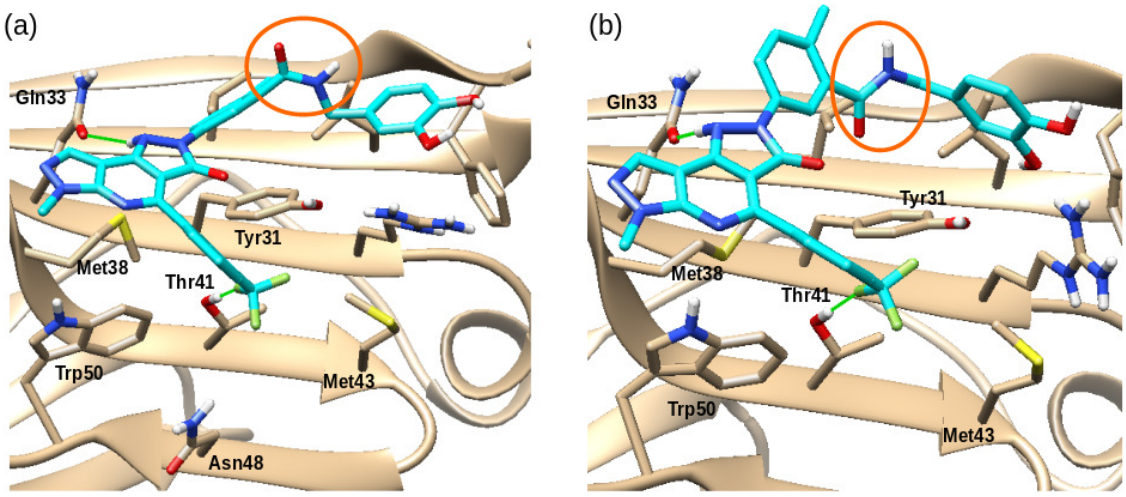


Figure 3.9: Conformations cis (a) et trans (b) de la liaison amide de la partie nord du ligand **9**. La liaison amide est entourée en orange.

Avant d'analyser ces figures, il est à souligner que, pour les méthodes de calculs XSCORE et ID-Score, plus le score calculé pour un ligand est positif et grand, plus son

affinité est élevée. Pour les méthodes ITScore et MM-PBSA, plus le score d'un ligand est négatif et bas, plus son affinité est élevée. Par conséquent : on obtient une bonne corrélation entre les scores calculés et les affinités expérimentales lorsque : (i) le coefficient est proche de 1 avec les méthodes XSCORE et ID-Score ; (ii) le coefficient est proche de -1 avec les méthodes ITScore et MM-PBSA.

La méthode de calcul XSCORE est composée de trois fonctions de scoring empiriques pouvant être utilisées ensemble ou séparément : HPScore, HSScore et HMScore. La moyenne des résultats obtenus avec ces trois fonctions permet d'avoir le score global noté X-Score. Nous avons testé chacune de ces fonctions de scoring sur les complexes formés par les ligands **1** à **17**. Puisque les résultats obtenus avec les fonctions HPScore et HSScore étaient très insatisfaisants, seuls ceux obtenus avec HMScore et X-Score seront présentés.

Les graphes de corrélation entre les affinités expérimentales des ligands et les scores obtenus avec HMScore et X-Score sont montrés dans la Figure 3.10. Les affinités des ligands de haut poids moléculaire **9**, **10** et **11** sont surévaluées par les fonctions HMScore et X-Score, et cela peu importe la conformation adoptée par ceux-ci (conformation cis ou trans de la liaison amide qui lie les groupements R au reste de la partie nord des ligands **9**, **10** et **11**) (Tableau 3.1, p. 77 et Figure 3.9). En effet, dans la Figure 3.10, on constate que ces ligands se trouvent tous très éloignés de la masse de points représentant les ligands de faible poids moléculaire.

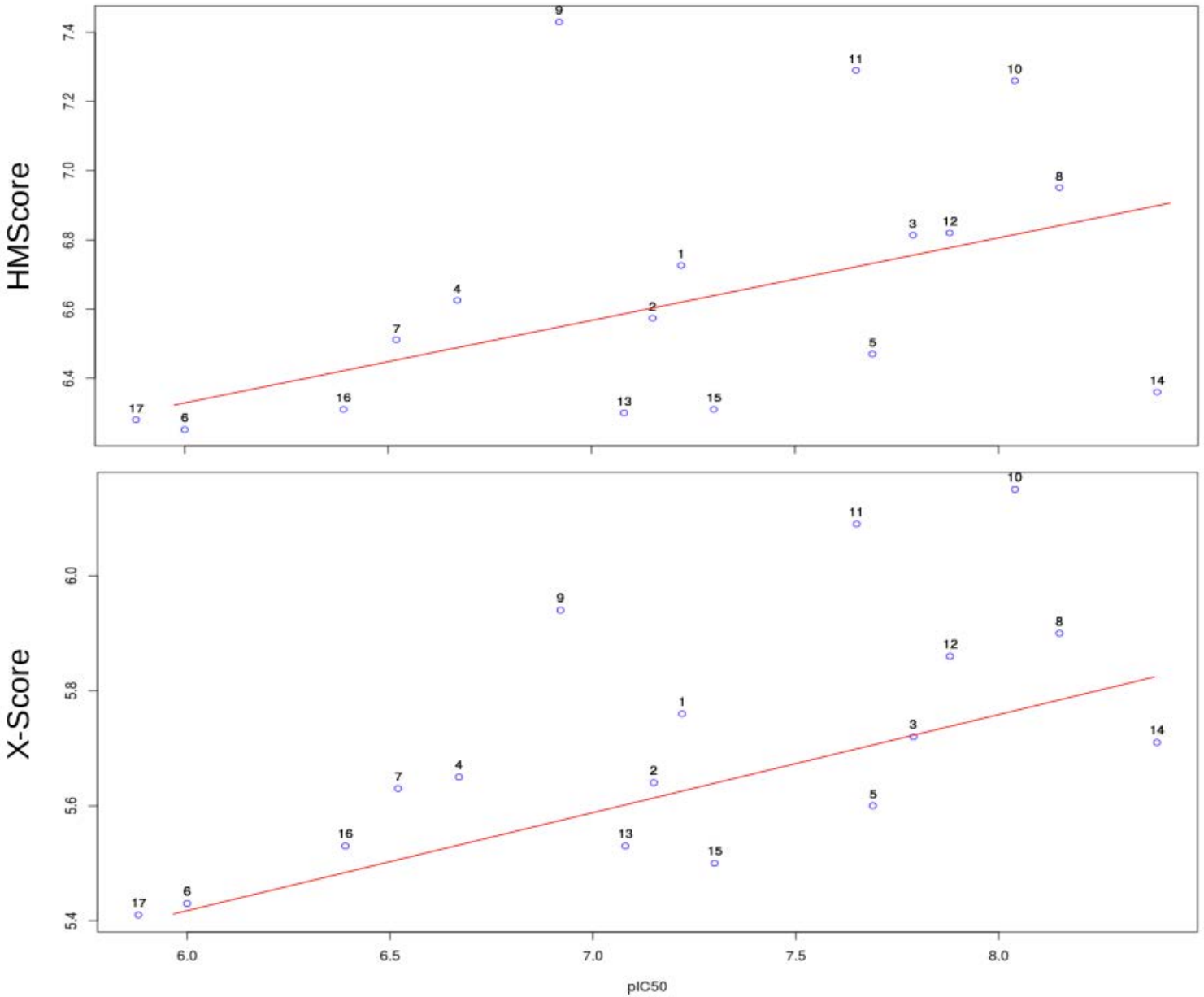


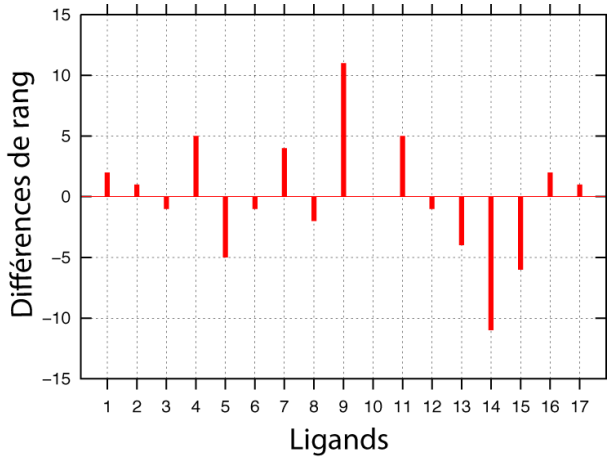
Figure 3.10: Corrélation entre les affinités expérimentales pIC50 et les scores HMScore (en haut avec R = 0.75) et X-Score (en bas, R = 0.79). Les coefficients de corrélation R ont été calculés sans les ligands de haut poids moléculaire **9**, **10** et **11** et le ligand **14**.

Les résultats obtenus avec HMScore et X-Score diffèrent peu sauf pour le ligand **14** pour lequel le score est sous-évalué par HMScore, et dans une moindre mesure, les ligands **12** et **13** pour lesquels les affinités sont respectivement surévaluée par X-Score et sous-évaluée par HMScore. Si on ne prend pas en compte les ligands de haut poids moléculaire (**9**, **10** et **11**) et le ligand **14**, on obtient des coefficients de corrélation de 0.75 et de 0.79 entre les affinités expérimentales et les scores respectivement calculés avec HMScore et X-Score. Ces deux méthodes de calcul semblent donc reproduire les affinités expérimentales de façon similaire. La comparaison du classement des ligands selon leurs affinités expérimentales au classement obtenu selon leur affinité calculée montre que peu de ligands sont correctement classés par HMScore et X-Score (Figure 3.11, a et b). Les différences de classement obtenues avec HMScore et X-Score sont semblables (Figure 3.11, c et d). Lorsqu'on écarte les ligands de haut poids moléculaire (**9**, **10** et **11**) et le ligand **14**, on constate que les différences entre le classement expérimental et les classements calculés sont semblables (Figure 3.11, c et d), avec toutefois une surévaluation ou une sous-évaluation de l'affinité de la plupart des ligands par HMScore et X-Score. Seul le classement des ligands **8**, **12** et **3** est correctement évalué par HMScore d'une part et ceux des ligands **8**, **12**, **6** et **17** sont correctement évalués par X-Score d'autre part. Dans les deux cas, les différences de rangs varient entre 1 et 4 (valeurs absolues) à l'exception du ligand **15** qui est fortement sous-évalué par les deux méthodes de calculs et dont le rang calculé diffère de 6 du rang expérimental. En d'autres termes, si on utilise les méthodes de calculs HMScore ou X-Score pour identifier les ligands qui se lient avec une forte probabilité à CD80 ou à un système similaire (Lu), il faut être conscient de cette marge d'erreur lors des analyses. Néanmoins, étant donné les coefficients de corrélation obtenus, cette méthode de scoring (HMSCore ou X-Score) permet de discriminer les ligands potentiels des autres molécules qui ne se lient pas ou peu.

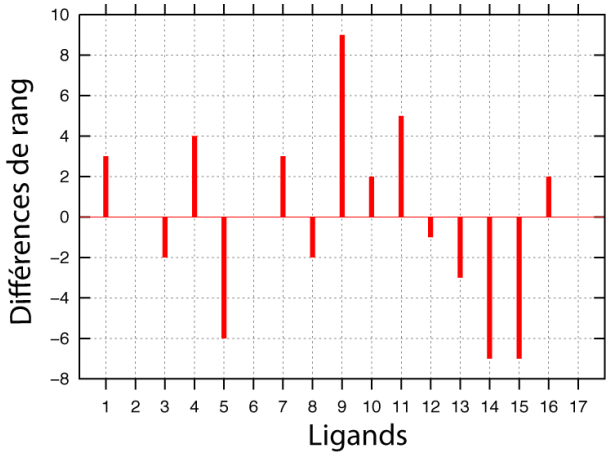
La méthode de calcul ID-Score décrit plusieurs types d'interaction : les interactions de van der Waals, les liaisons hydrogène, les interactions électrostatiques, les interactions

$\pi - \pi$, les effets de solvatation, les effets de perte d'entropie ainsi que la complémentarité entre la structure du ligand et celle du site de liaison lors de la complexation [29]. Bien que cette fonction de scoring offre une description plutôt complète des interactions protéine-ligand, on n'obtient qu'une faible corrélation entre les scores calculés et les affinités expérimentales (Figure 3.12, a). On constate que les affinités des ligands **1**, **4**, **6**, **7** et **12** (Tableau 3.1, p. 77) sont surévaluées par la méthode ID-Score, ce qui donne un coefficient de corrélation de 0.53.

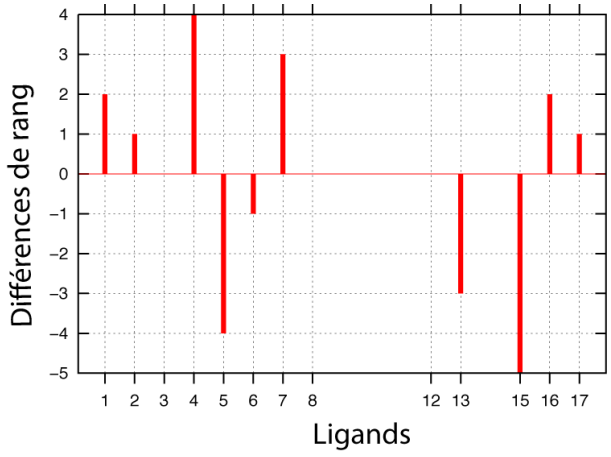
Figure 3.11: Différence entre le classement des ligands selon leurs affinités expérimentales et selon leurs affinités calculées par HMScore (a et c) et par X-Score (b et d) pour les 17 ligands (a et b) et pour l'ensemble des ligands à l'exception des ligands **9**, **10**, **11** et **14** (c et d). Les différences de rangs sont en ordonnées. Les valeurs de différences entre les rangs expérimentaux et les rangs calculés sont nulles, négatives et positives lorsque les énergies de liaison sont respectivement identiques, sous-évaluées et surévaluées par HMScore ou X-Score.



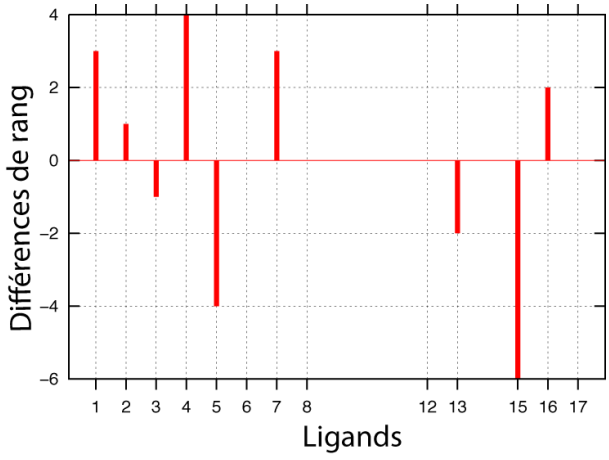
(a)



(b)



(c)



(d)

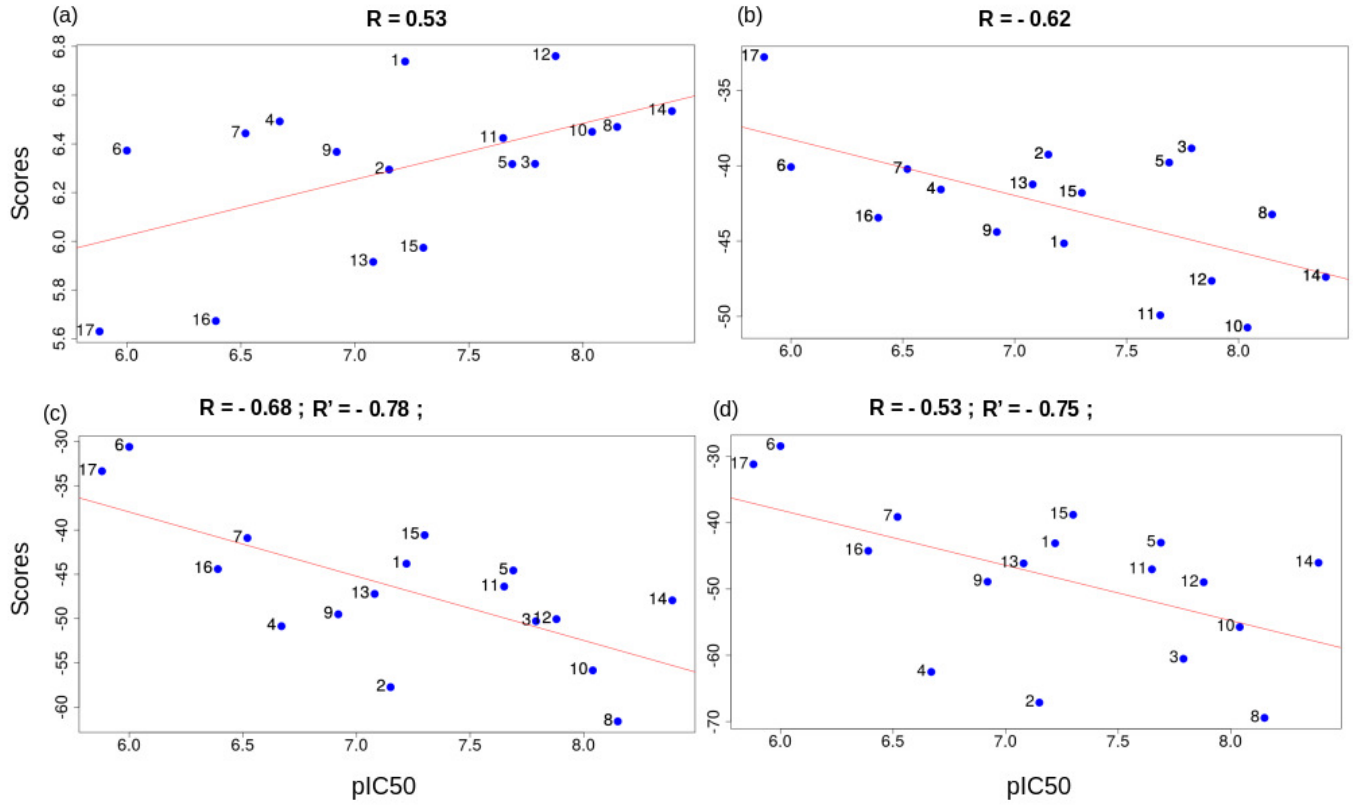


Figure 3.12: Corrélation entre les affinités expérimentales et les scores calculés par ID-Score (a), ITScore (b), MM-PBSA (c) et MM (d). Les valeurs R correspondent aux coefficients de corrélation obtenus avec la méthode de Spearman pour l'ensemble des ligands. Si on ne prend pas en compte les ligands chargés **2**, **3**, **4** et **8** dans le calcul du coefficient pour MM-PBSA, on obtient R'.

ITScore est une méthode de calcul itérative basée sur des études statistiques de complexes protéine-ligand [30]. Cette méthode ne semble pas non plus convenir à notre système CD80-ligand. En effet, bien qu'on obtienne une meilleure corrélation avec ITScore (R = -0.62) qu'avec ID-Score (R = 0.53), celle-ci reste insuffisante (Figure 3.12, b).

Enfin, nous avons comparé les affinités expérimentales aux énergies calculées avec la méthode MM-PBSA. Cette méthode de calcul combine la méthode de mécanique moléculaire (MM) à la méthode PBSA (Poisson-Boltzmann Surface Area) [32]. La partie MM permet de calculer les énergies liées (liaisons covalentes, angles de valence, angles dièdres), les interactions électrostatiques et les interactions de van der Waals alors que la partie PBSA permet de calculer l'énergie de solvatation (contributions polaires et non polaires). Les énergies calculées avec la méthode MM-PBSA sont bien corrélées avec les affinités expérimentales avec un coefficient proche de -0.70 (Figure 3.12, c). Cette corrélation est cependant meilleure si on écarte les ligands chargés **2**, **4** et **8**. En effet, on constate que les énergies de solvatation des ligands chargés sont surévaluées par la méthode PBSA, ce qui donne des énergies MM-PBSA trop basses par rapport à celles calculées pour les ligands neutres. Cette remarque a aussi été faite dans une étude de Wang et al [106]. Ainsi, si on ne considère que les ligands neutres dans le calcul du coefficient de corrélation, on obtient un coefficient proche de -0.80. Toutefois, il est à noter que cette bonne corrélation est due en grande partie à la modification des charges que nous avons réalisée sur la partie cœur des ligands. En effet, lorsqu'on compare les énergies calculées avec la méthode MM (sans l'énergie de solvatation) aux affinités expérimentales pour les ligands neutres, on obtient déjà à ce stade une bonne corrélation de -0.75 (Figure 3.12, d). Lors de la comparaison de classement des ligands neutres selon leur affinité calculée (avec la méthode MM-PBSA) avec celui basé sur leur affinité expérimentale, une différence de classement maximale de 6 est observée.

Les meilleures corrélations entre les affinités calculées (scores) et les affinités expérimentales ont été obtenues avec les méthodes XSCORE (HMScore et X-Score) et MM-

PBSA. Cependant, ces méthodes n'évaluent pas de manière précise les interactions avec les atomes d'halogènes (présents dans les 17 ligands de CD80 étudiés ici) et la protéine. En effet, dans ces méthodes de scoring, les liaisons halogène ainsi que les liaisons hydrogène faibles pouvant être formées à partir des atomes de fluor sont décrites comme de simples contacts de van der Waals. De plus, les contributions des nombreuses interactions $\pi - \pi$ entre les ligands et les résidus Tyr31 et Trp50 sont souvent sous-estimées par ces méthodes de calculs empiriques. Afin de mieux décrire ces interactions et de reproduire les affinités expérimentales de ces ligands, nous avons réalisé des calculs de chimie quantique sur les complexes formés par les ligands **1** à **8** et les ligands **14**, **15** et **16**.

Chapitre 4

Calcul des énergies d'interaction avec des méthodes de chimie quantique

Comparées aux techniques de mécanique moléculaire, les techniques de chimie quantique permettent d'évaluer avec une meilleure précision les interactions entre la protéine et le ligand. En effet, alors que la mécanique moléculaire décrit les atomes d'une molécule comme des sphères qui sont reliées entre elles par des ressorts et qui interagissent entre elles selon les principes de mécanique classique, la chimie quantique décrit les molécules au niveau électronique : on résout l'équation de Schrödinger qui décrit les mouvements des électrons en considérant les noyaux comme fixes. Il existe différentes méthodes qui permettent de résoudre cette équation : les méthodes *ab initio* dans lesquelles toutes les intégrales sont calculées (Hartree-Fock (HF) et post-Hartree-Fock telle que la méthode MP2) et les méthodes semi-empiriques dans lesquelles certaines intégrales sont estimées à partir de données expérimentales (PM6, CNDO, etc ...). Ces méthodes de chimie quantique nécessitent un temps de calcul plus long que les techniques de mécanique moléculaire : en chimie quantique, le temps de calcul augmente à la puissance 4 avec le nombre d'orbitales atomiques de la base.

Dans le chapitre précédent, nous avons validé la fonction de scoring secondaire XSCORE qui a permis de reproduire les affinités expérimentales de la plupart des ligands (ligands **1** à **8** et **12**, **13**, **15** et **16**) dans la pose que nous avons choisie (Figure 3.2, p. 79). Bien qu'un coefficient de corrélation de 0.79 (sans les ligands **9** à **11** et **14**) ait été obtenu avec XSCORE, cette méthode de calcul ne décrit pas les interactions halogène entre les ligands et la protéine. Dans ce chapitre, nous avons donc voulu (i) reproduire les affinités expérimentales des ligands **1** à **8** et des ligands **14**, **15** et **16** avec des calculs de chimie quantique, ce qui nous permettra aussi de confirmer la pose choisie (Figure 3.2, p. 79) et (ii) analyser les interactions entre CD80 et ces ligands dans la pose sélectionnée.

Les ligands cités ci-dessus ont été choisis parce que : (i) ce sont des ligands de petite taille, ce qui nous permettra d'avoir un temps de calcul raisonnable, (ii) chacun de ces ligands possède un atome de fluor ou de chlore dans leur partie nord (Figure 3.1, p. 78 et Tableau 3.1, p. 77) qui peut faire une liaison halogène avec la protéine (interactions mal décrites par XSCORE) dans la pose sélectionnée, et (iii) chacun de ces ligands possède un groupement CF_3 (ligands **1** à **8**) ou NO_2 (ligands **14** et **15**) ou NH_2 (ligand **16**) qui fait une liaison hydrogène avec le résidu Thr41 (Figure 3.2, p. 79). En choisissant ces ligands, nous voulions décrire plus précisément les liaisons hydrogène, les liaisons halogène et les interactions $\pi - \pi$ établies entre CD80 et les ligands dans la pose sélectionnée et ainsi obtenir une corrélation plus précise entre les affinités calculées et les affinités expérimentales.

Deux méthodes de calcul de chimie quantique ont été testées dans ce chapitre : (i) PM6-DH2X (dans MOPAC) qui permet de réaliser des calculs semi-empirique et (ii) FMO qui permet de faire des calculs *ab initio*.

La méthode FMO permet de faire des calculs *ab initio* sur des systèmes constitués de milliers d'atomes. Pour des systèmes d'une telle envergure, les interactions sont difficilement décrites par des méthodes de chimie quantique classiques pour lesquelles le temps de calcul augmente à la puissance 4 avec le nombre d'orbitales atomiques de la

base. La méthode FMO permet de contourner ce problème en faisant des calculs *ab initio* sur des fragments du système plutôt que sur le système entier, ce qui réduit considérablement le temps de calcul. Un fragment peut être par exemple un "résidu" d'acide aminé[1], une molécule d'eau ou encore un ligand non protéique. L'énergie du système entier s'obtient par la somme (i) de l'énergie des fragments I seuls (monomères) et (ii) de l'énergie d'interaction entre chaque paire de dimères IJ noyée dans un champ électrostatique résultant de la densité électronique des autres monomères [33] (Figure 4.1). Cette méthode permet donc de faire des calculs qui restent très précis sur de larges systèmes dans un intervalle de temps raisonnable.

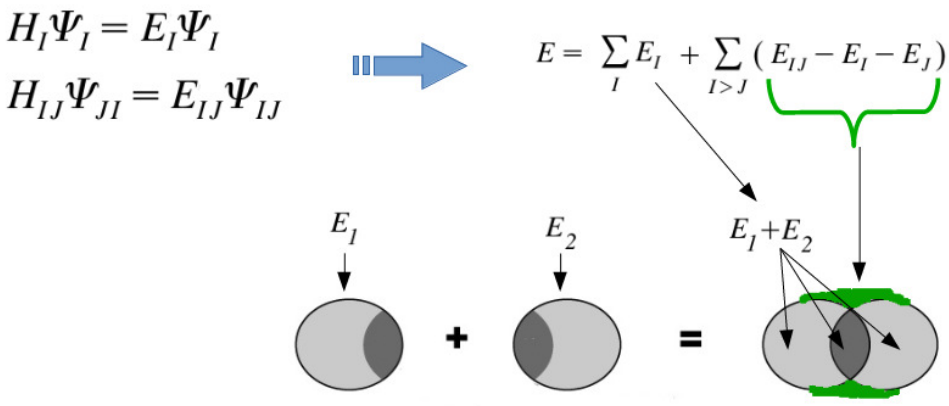


Figure 4.1: Les équations à gauche de la Figure sont les équations de Shrödinger pour le monomère I et le dimère IJ. L'énergie E s'obtient en calculant la somme des énergies des monomères (premier terme) et la somme des énergies d'interaction entre tous les dimères (deuxième terme).

Dans ce qui suit, nous comparerons d'abord, et très brièvement, les énergies de complexation calculées avec la méthode PM6-DH2X pour les ligands **1** - **8** et **14** - **16** aux

[1]La fragmentation de la chaîne polypeptidique se fait par rupture de la liaison entre le C^{α} et le C du carbonyle de la liaison peptidique.

affinités expérimentales correspondantes. Puis, nous nous concentrerons sur les résultats obtenus avec la méthode de calcul FMO.

4.1 Calcul des affinités des ligands avec la méthode PM6-DH2X

Afin d'évaluer les affinités des ligands, nous avons d'abord réalisé des calculs avec la méthode semi-empirique PM6-DH2X. Pour chacun des ligands étudiés ici (ligands **1** - **8** et **14** - **16**), 10 géométries ont été extraites de façon équidistante. Le protocole suivi pour obtenir ces géométries correspond au protocole 2 qui est expliqué dans la section suivante, protocole pour lequel les calculs réalisés à l'étape 3 sont remplacés par les calculs PM6-DH2X. Cet hamiltonien PM6-DH2X permet de décrire plusieurs liaisons non covalentes, dont les liaisons hydrogène ainsi que les liaisons halogène. Deux types de calculs ont été réalisés : des calculs dans le vide et des calculs en présence d'eau implicite en utilisant le modèle de solvatation COSMO [107]. Les résultats obtenus dans les deux cas sont très insatisfaisants. En effet, on obtient des coefficients de corrélation très faibles de -0.28 et -0.32 lorsque les calculs sont respectivement faits en présence d'eau implicite (Figure 4.2a) et dans le vide (Figure 4.2b). Dans la suite du manuscrit, la méthode PM6-DH2X ne sera pas utilisée pour des calculs d'énergie de complexation, mais pour minimiser les complexes protéine-ligand, on parlera de minimisation MOPAC.

4.2 Calcul des affinités des ligands avec la méthode FMO

Dans cette section, nous présenterons d'abord les différents protocoles testés afin de reproduire les affinités expérimentales avec différentes méthodes de calculs FMO. Puis, nous justifierons la base utilisée pour l'ensemble de ces calculs FMO avant de comparer

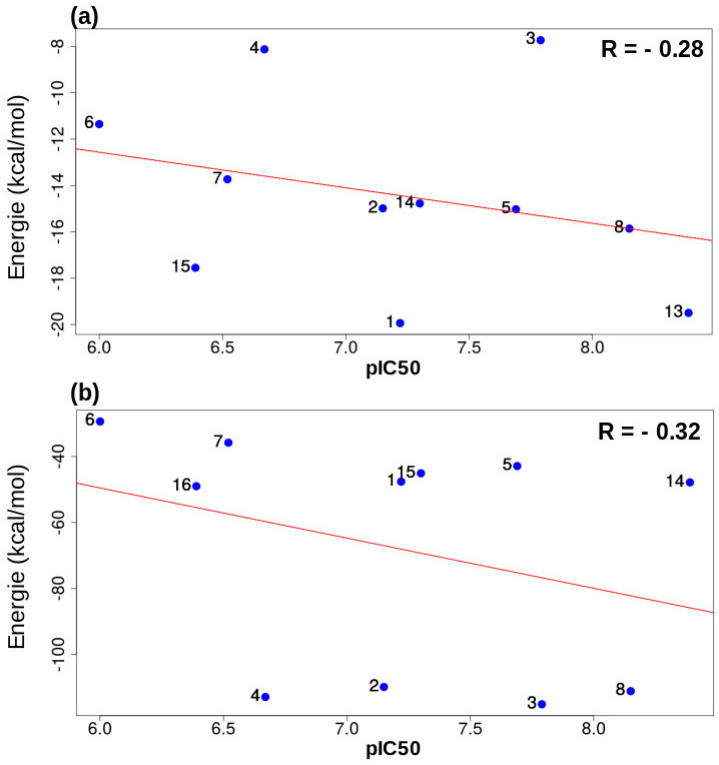


Figure 4.2: Corrélation entre les affinités expérimentales et les affinités calculées avec la méthode PM6-DH2X (a) en présence d'eau implicite et (b) dans le vide. Les calculs ont été réalisés sur 10 géométries extraites de façon équidistante des trajectoires des complexes formés entre CD80 et chacun des ligands **1** - **8** et **14** - **16**. Les coefficients de corrélation R sont montrés dans chacun des graphes.

les affinités calculées (avec les différentes méthodes de calculs FMO) aux affinités expérimentales des ligands **1** à **8** et des ligands **14**, **15** et **16** dans la pose choisie (Tableau 3.1, p. 77 et Figure 3.2, p. 79).

4.2.1 Description des différents protocoles testés

Afin de reproduire les affinités expérimentales des ligands cités ci-dessus, trois protocoles ont été testés et sont présentés dans la Figure 4.3.

Dans les protocoles 1 et 2 de la Figure 4.3, 10 géométries sont extraites de façon équidistante sur la trajectoire stable de chaque ligand (étape 1) puis celles-ci sont minimisées avec CHARMM (protocole 1, étape 2) ou MOPAC (protocole 2, étape 2') en présence d'eau explicite. Dans les minimisations MOPAC, nous utilisons l'hamiltonien PM6-DH2X. Dans le protocole 3, nous avons décidé de faire un échantillonnage moins aléatoire que ceux réalisés dans les protocoles 1 et 2 : un échantillonnage ciblé qui consiste à extraire les géométries les plus représentatives de la trajectoire de chaque ligand. Le but de cet échantillonnage ciblé est de faire un minimum de calcul sur les géométries les plus représentatives uniquement pour chacun des complexes protéine-ligand.

Pour sélectionner les géométries les plus représentatives de la trajectoire de chacun des ligands, nous avons procédé à une étape de clusterisation de la trajectoire qui consiste à regrouper dans un même cluster les géométries pour lesquelles la position du ligand et celles des chaînes latérales des résidus Tyr31, Met38, Gln33, Thr41 et Tpr50 sont similaires. Quatre de ces cinq derniers résidus (Met38, Gln33, Tyr31 et Tpr50) sont fortement impliqués dans la liaison des ligands **1**-**8** et des ligands **14**-**16** selon les calculs FMO-RI-MP2 (résultats présentés dans la section 4.3, p. 124). Cette étape de clusterisation a été réalisée avec le programme WORDOM [108]. Pour chaque ligand, nous avons sélectionné les clusters qui contiennent les plus grands nombres de géométries, c'est-à-dire les clusters les plus peuplés de la trajectoire. Le nombre de clusters retenu par ligand

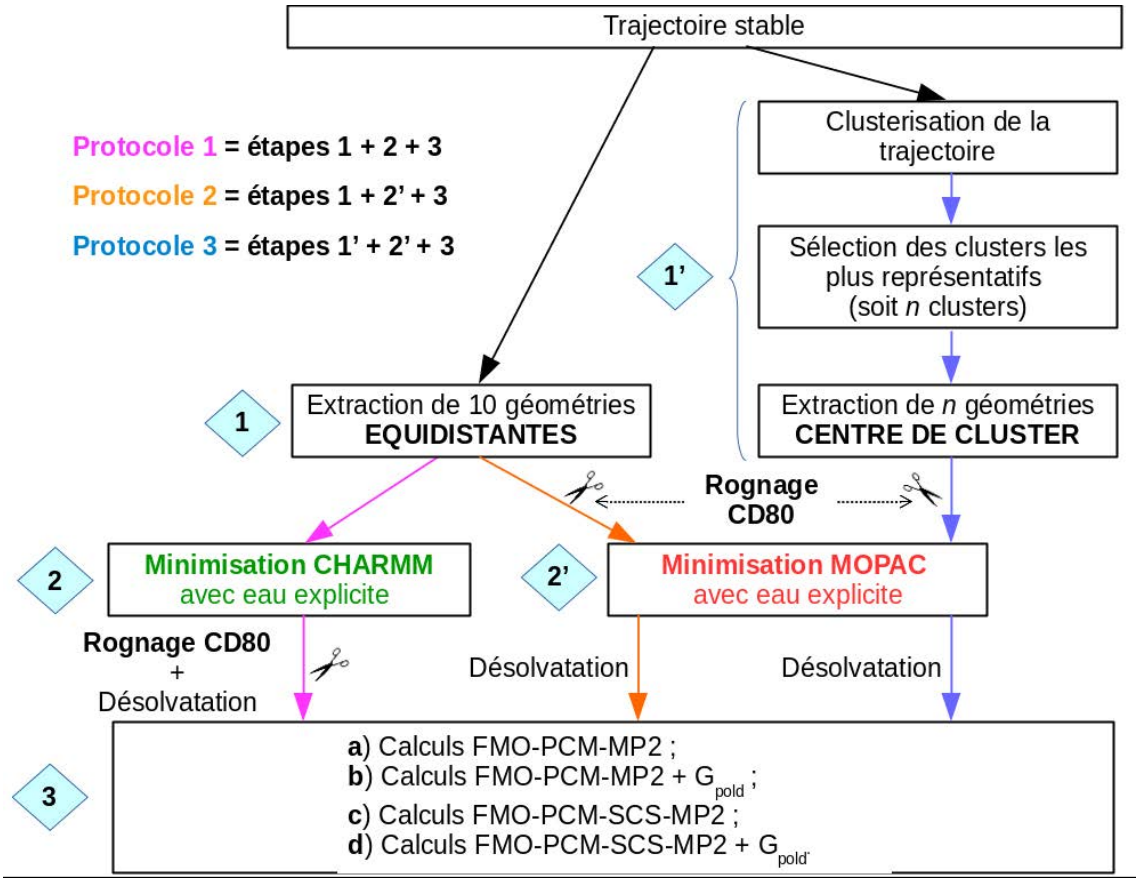


Figure 4.3: Trois protocoles ont été testés pour évaluer les énergies d'interaction CD80-ligand avec la méthode FMO : les protocoles 1, 2 et 3 qui sont respectivement formés des étapes 1 +2 +3, 1 +2'+3 et 1'+2'+3. Les étapes 1 et 1' correspondent respectivement aux étapes d'échantillonnage systématique et par clusterisation de la trajectoire. L'étape 1' de clusterisation de la trajectoire consiste à regrouper les géométries qui se ressemblent en un cluster. Les clusters retenus pour chaque ligand sont les clusters contenant le plus grand nombre de géométries. À l'intérieur de chaque cluster, une seule géométrie est retenue et correspond à la géométrie la plus représentative de ce cluster appelée *géométrie centre de cluster*. Le nombre de clusters *n* (et donc le nombre de géométries centres de cluster extraites) varie entre 5 et 10 selon le ligand. Les étapes 2 et 2' correspondent respectivement aux étapes de minimisation avec CHARMM et avec MOPAC en présence d'eau explicite. Dans l'étape 2, la minimisation CHARMM a été réalisée sur des systèmes entiers alors que dans l'étape 2', la minimisation MOPAC a été réalisée sur des systèmes réduits (protéine tronquée afin de ne garder que la face qui interagit avec les ligands). La stratégie de minimisation MOPAC utilisée à l'étape 2' est montrée dans la Figure 4.5. Enfin, l'étape 3 consiste à calculer les énergies des complexes préalablement désolvatés (et préalablement tronqués également dans le protocole 1) suivant quatre méthodes de calcul FMO : (i) avec le niveau MP2 sans et avec la contribution supplémentaire à l'énergie de solvatation G_{pold} (**a** et **b**), (ii) avec le niveau SCS-MP2 sans et avec la contribution supplémentaire à l'énergie de solvatation G_{pold} (**c** et **d**).

(n) varie entre 5 et 10. Pour chaque cluster, le programme WORDOM identifie un centre (centre de cluster ; étape 1') : il s'agit de la géométrie la plus représentative du cluster considéré (Figure 4.4, p. 111). Ces géométries seront appelées géométries centres de cluster dans le reste du manuscrit. Les géométries centres de cluster ont ensuite été minimisées avec MOPAC en présence d'eau explicite. Puis, dans tous les protocoles, les géométries minimisées ont été désolvatées avant de calculer leurs énergies d'interaction associées en utilisant différentes variantes de calculs FMO (**a**, **b**, **c** et **d** ; étape 3).

Pour chaque ligand, tous les calculs FMO-MP2 ont été réalisés sur un système réduit dans lequel la protéine a été tronquée afin de ne garder que la face qui interagit avec le ligand. La réduction de taille de chacun des systèmes a été réalisée après l'étape 2 de minimisation CHARMM dans le protocole 1 et avant l'étape 2' de minimisation MOPAC dans les protocoles 2 et 3. Par conséquent, les minimisations CHARMM (protocole 1) ont été réalisées sur des systèmes entiers alors que les minimisations MOPAC (protocoles 2 et 3) ont été réalisées sur des systèmes de taille réduite. Il était important de réduire la taille des systèmes avant la minimisation MOPAC puisque celle-ci prend beaucoup plus de temps (environ 2 heures sur un système de taille réduite) qu'une minimisation CHARMM (environ 1 minute sur un système entier). Ce temps de calcul est d'autant plus grand lorsqu'on rajoute des molécules d'eau explicite puisqu'on augmente le nombre d'électrons à traiter. Ainsi, dans l'optique de réaliser des minimisations MOPAC en présence d'eau explicite dans des temps raisonnables, nous avons employé la stratégie suivante (Figure 4.5): (i) seuls les résidus qui sont en contact avec le ligand considéré sont laissés mobiles lors de la minimisation et le reste de la protéine tronquée a été gardé fixe, (ii) l'équivalent de deux couches d'eau mobiles a été utilisé (couche d'eau interne jusqu'à 6 Å autour du ligand) et (iii) une couche d'eau externe entre les rayons de 6 et 9 Å autour du ligand a été utilisée et gardée fixe lors de la minimisation afin d'éviter le phénomène d'évaporation.

Dans toutes les variantes de calcul FMO testées (**a**, **b**, **c** et **d** ; étape 3, Figure 4.3,

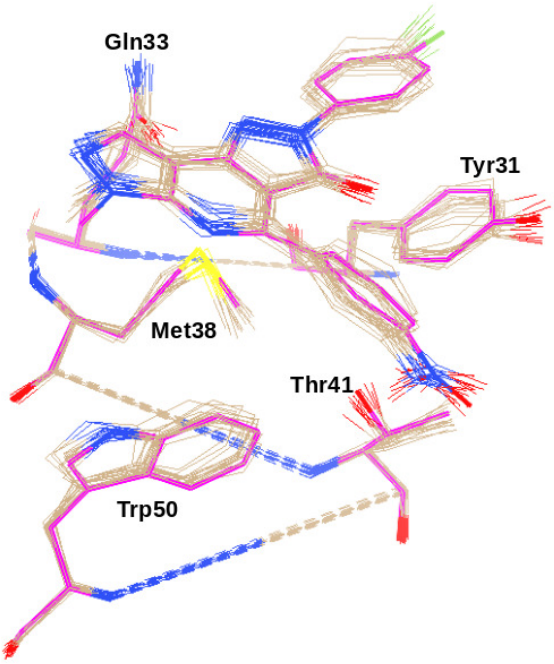


Figure 4.4: Exemple du neuvième cluster issu de la clusterisation de la trajectoire choisie pour le ligand **15**. Ce cluster contient 18 géométries (en marron) dont l'une est la géométrie centre de cluster (en magenta). Le cluster a été obtenu en considérant les conformations adoptées par le ligand et les résidus Tyr31, Gln33, Met38, Thr41 et Trp50 tout le long de la trajectoire.

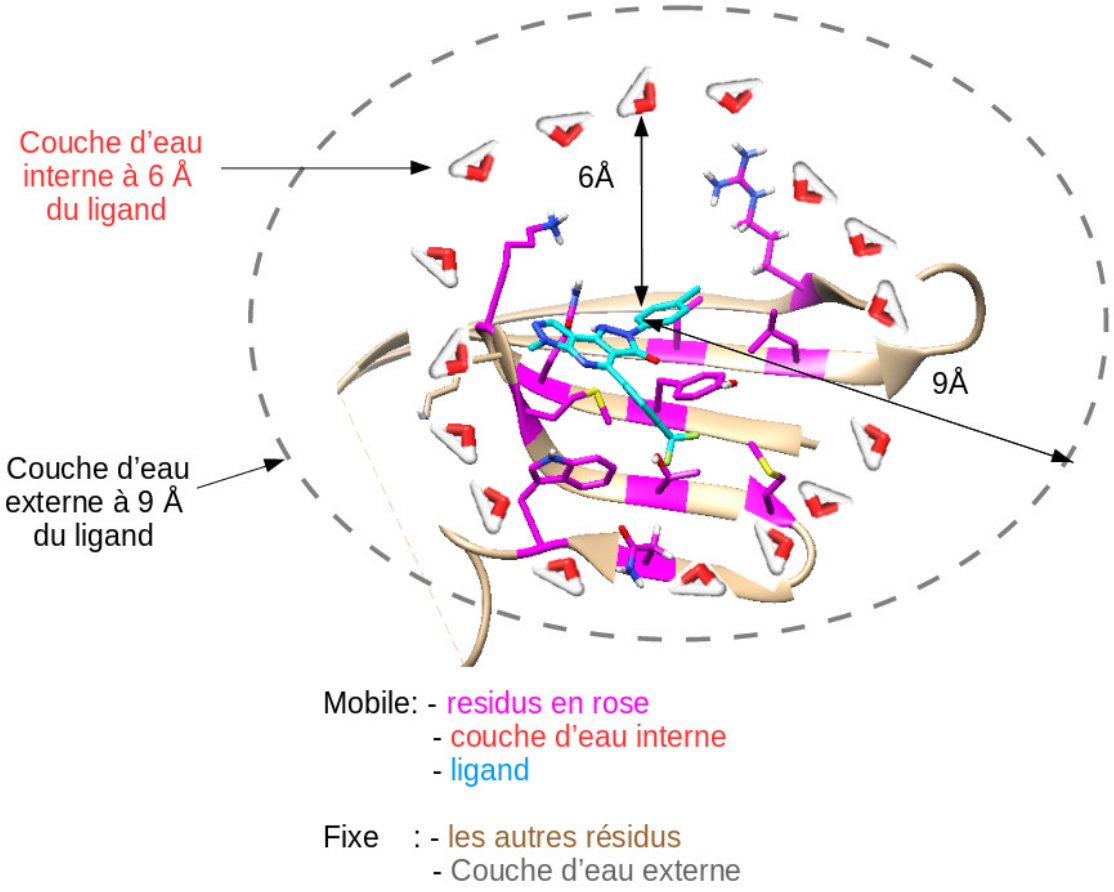


Figure 4.5: Système de taille réduite sur lequel la minimisation MOPAC (en utilisant la méthode PM6-DH2X) de l'étape 2' des protocoles 2 et 3 (Figure 4.3) a été réalisée. Le système est constitué d'une seule face de la protéine (face qui interagit avec le ligand), du ligand (en bleu), d'une couche d'eau interne située à 6 Å du ligand et d'une couche d'eau externe située entre 6 et 9 Å du ligand (trait pointillé gris). Pendant la minimisation, les résidus en rose, le ligand et la couche d'eau interne sont laissés mobiles alors le reste de la protéine tronquée (marron) et la couche d'eau externe sont gardés fixes. Pour des raisons de clarté, toutes les molécules d'eau n'ont pas été représentées.

p. 109), l'effet du solvant a été pris en compte grâce à la méthode PCM (Polarizable Continuum Model) [109]. La méthode PCM permet de solvater le système de façon implicite : le solvant est modélisé par un continuum diélectrique polarisable dans lequel le soluté crée une cavité. D'autre part, dans l'ensemble de ces calculs FMO, nous avons utilisé la base 6-31G pour décrire les orbitales atomiques des atomes des complexes. Nous discuterons du choix de cette base dans la sous-section suivante 4.2.2.

Les énergies de complexation FMO-PCM sont calculées selon la méthode usuelle :

$$\Delta G_{FMO-PCM} = G_{FMO-PCM,complexe} - (G_{FMO-PCM,proteine} + G_{FMO-PCM,ligand}) \quad (4.1)$$

avec pour les variantes **a** et **c** :

$$G_{FMO-PCM} = G_{elec} + G_{cav} + G_{disp} + G_{rep} \quad (4.2)$$

et pour les variantes **b** et **d** :

$$G_{FMO-PCM+G_{pold}} = G_{interne} + G_{elec} + G_{cav} + G_{disp} + G_{rep} \quad (4.3)$$

et

$$G_{interne} = E_{gaz} + G_{pold} \quad (4.4)$$

Le terme entropique ($-T\Delta S$) n'a pas été considéré dans tous les calculs.

E_{gaz} : énergie du soluté dans le vide ;

G_{pold} : énergie de déstabilisation due à la polarisation mutuelle solvant-soluté [110] ;

G_{elec} : énergie électrostatique prenant en compte la polarisation des électrons du soluté par le solvant (pas de terme équivalent dans méthode PB puisque le champ de force n'est pas polarisable) ;

G_{cav} : énergie nécessaire à la création d'une cavité dans le solvant pour l'immersion du soluté ;

G_{disp} : énergie de dispersion soluté-solvant (équivalent au terme $\frac{-B}{r}^6$ dans l'expression de Lennard-Jones) ;

G_{disp} : énergie d'échange (ou de répulsion) soluté-solvant (équivalent au terme $\frac{A}{r}^{12}$ dans l'expression de Lennard-Jones) ;

Afin de prendre en compte la contribution $G_{interne}$ dans les variantes de calcul **b** et **d**, les calculs d'énergie de complexation nécessitent les deux types suivants de calcul par soluté (*i.e.* pour chacun des complexes, ligands et protéines) : (1) dans le vide/gaz et (2) dans le modèle d'eau PCM. Néanmoins, la façon la plus habituelle de calculer l'énergie de complexation avec le modèle d'eau PCM est de négliger la contribution $G_{interne}$, ce qui permet d'éviter les calculs dans le vide pour les trois solutés à considérer.

Les niveaux de calcul MP2 et SCS-MP2 permettent tous les deux de prendre en compte la corrélation électronique d'un système. La corrélation électronique décrit les mouvements concertés entre les électrons d'un système. L'énergie de corrélation électronique se décompose en deux termes : (i) la contribution à l'énergie de corrélation due aux spins parallèles et (ii) celle due aux spins antiparallèles. Dans le niveau de calcul MP2, ces deux contributions sont traitées de façon égale alors que dans le niveau SCS-MP2 (Spin Component Scaled-MP2), on considère que l'énergie de corrélation due aux spins parallèles est différente de celle qui est due aux spins antiparallèles. Le niveau SCS-MP2 est donc une variante du niveau MP2 qui va calibrer les contributions de l'énergie de corrélation dues respectivement aux spins parallèles et antiparallèles : celle due aux spins parallèles va être réduite et celle due aux spins antiparallèles (sous-estimée dans le niveau MP2) va être augmentée dans le niveau SCS-MP2 [111].

Les affinités calculées de chaque ligand résultent de la moyenne des affinités calculées sur les géométries issues de chaque protocole, soit une moyenne des affinités calculées

sur les 10 géométries issues du protocole 1 ou 2, et une moyenne des affinités calculées sur les *n* géométries issues du protocole 3.

4.2.2 Choix de la base utilisée dans les calculs FMO-MP2

Dans tous les calculs FMO décrits ci-dessus, nous utilisons la base 6-31G. En chimie quantique, une base constitue un ensemble de fonctions mathématiques (gaussiennes) qui décrit les orbitales atomiques des atomes des molécules. Une orbitale atomique est définie comme l'espace dans lequel gravite au plus une paire d'électrons d'un atome. Plus le nombre de fonctions gaussiennes utilisées pour décrire chacune des orbitales atomiques sera élevé, plus les orbitales atomiques seront précisément décrites mais, en contre partie, plus le temps de calcul sera élevé. Il est donc intéressant de savoir si l'utilisation de la base 6-31G est suffisamment précise pour décrire notre système ou si une base plus lourde telle que 6-31G* est nécessaire. Pour cela, nous avons également calculé les affinités des deux ligands **8** et **14** avec la méthode FMO-PCM-MP2 et la base 6-31G* (6-31G*/FMO-PCM-MP2) sur les 10 géométries issues du protocole 2 (Figure 4.3). Puis les résultats ont été comparés à ceux obtenus avec la variante de calcul **a** (FMO-PCM-MP2 et la base 6-31G notée 6-31G/FMO-PCM-MP2, Figure 4.3) pour les mêmes géométries de ces mêmes ligands. Nous avons choisi de faire ces calculs sur les ligands **8** et **14** puisqu'ils sont respectivement chargé et neutre et représentent donc les deux types de ligands présents dans le set étudié (Tableau 3.1, p. 77).

La base 6-31G permet de décrire la densité électronique des atomes de la façon suivante : une orbitale de cœur (1s par exemple pour les atomes C, N et O) est décrite par une contraction de 6 gaussiennes et chacune des orbitales de valence est décrite par une contraction de 3 gaussiennes plus une gaussienne individuelle. La base 6-31G* décrit les orbitales atomiques de la même façon que la base 6-31G mais possède une fonction supplémentaire de type d qui permet de décrire la polarisation. Cette dernière permet donc

une description *a priori* plus précise des systèmes étudiés mais requiert plus de temps : un calcul de l'énergie $G_{FMO-PCM,proteine}$ de l'équation 4.1 (page 113) pour la protéine seule tronquée (soit 1026 atomes) avec la méthode 6-31G*/FMO-PCM-MP2 prend environ 23 heures contre environ 5 heures avec la méthode 6-31G/FMO-PCM-MP2.

La comparaison des affinités des ligands **8** et **14** obtenues avec les bases 6-31G et 6-31G* pour chacune des 10 géométries est montrée dans la Figure 4.6 (p. 117). On constate que les affinités calculées avec la base 6-31G* sont, pour chacune des 10 géométries, plus négatives (affinités plus élevées) que celles calculées avec la base 6-31G, que ce soit pour le ligand **8** chargé ou le ligand **14** neutre. Les affinités moyennes obtenues à partir de 10 géométries du ligand **8** sont de -1.59 kcal/mol et -7.02 kcal/mol respectivement avec les bases 6-31G et 6-31G*. Pour le ligand **14**, les affinités moyennes sont de -0.12 kcal/mol et -5.53 kcal/mol respectivement avec les bases 6-31G et 6-31G*. Ainsi, et de manière surprenante, la différence d'énergie entre l'affinité moyenne calculée avec la base 6-31G et celles calculées avec la base 6-31G* est d'environ -5.4 kcal/mol pour chacun de ces deux ligands (soit Δ = -1.59-(-7.02) = 5.43 kcal/mol pour le ligand **14** et Δ = -0.12-(-5.53) = 5.41 kcal/mol pour le ligand **8**) . Puisque seules les valeurs d'énergies absolues obtenues pour ces deux ligands sont différentes et que les valeurs d'énergies relatives sont les mêmes, nous pouvons donc considérer que la base 6-31G représente un compromis judicieux en termes de temps et de précision pour décrire les interactions entre ces deux types de ligands (chargé et neutre) et CD80. Les résultats qui suivent sont donc ceux qui ont été obtenus avec la base 6-31G uniquement.

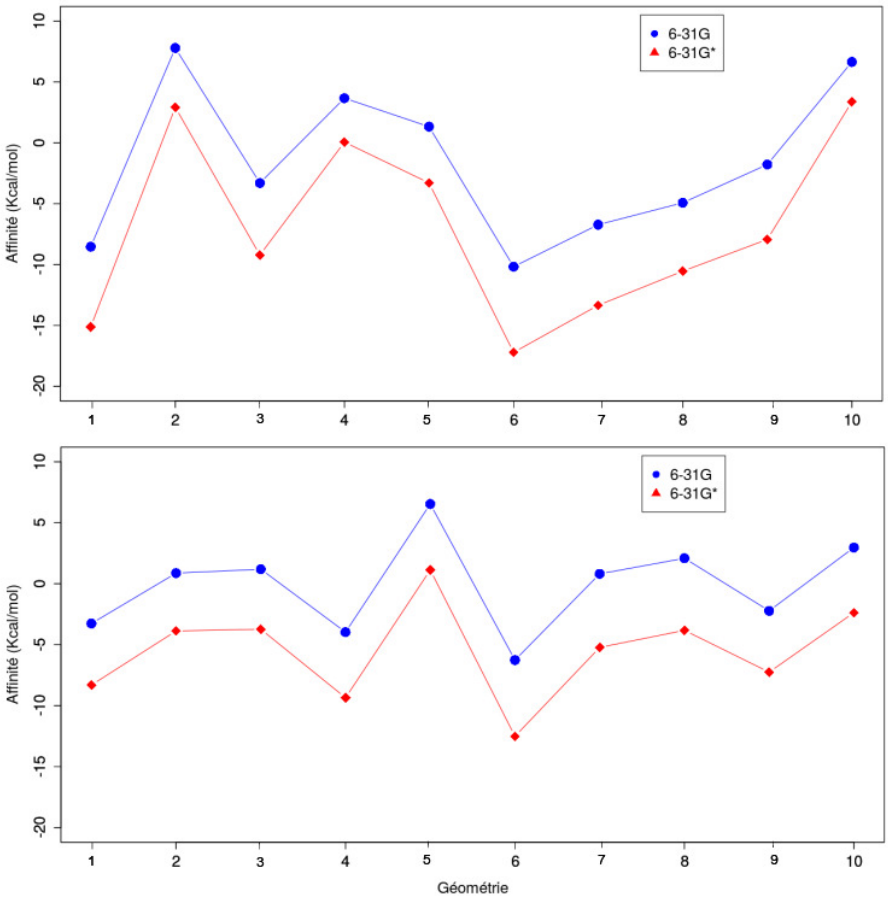


Figure 4.6: Affinités des ligands **8** (en haut) et **14** (en bas) calculés avec la méthode FMO-PCM-MP2 et la base 6-31G (en points bleus) ou la base 6-31G* (en losange rouge) pour chacune des 10 géométries issues du protocole 2 de la Figure 4.3, p. 109.

4.2.3 Comparaison des affinités calculées à partir des différentes variantes FMO avec les affinités expérimentales pour les ligands 1 - 8 et 14 - 16

La figure 4.7 montre les corrélations obtenues entre les affinités calculées avec les variantes de calcul FMO testées (**a**, **b**, **c** et **d**) dans les différents protocoles (protocoles 1, 2 et 3 de la Figure 4.3, p. 109) et les affinités expérimentales des ligands **1** à **8** et des ligands **14**, **15** et **16**. Pour la méthode FMO, tout comme pour la méthode MM-PBSA, plus l'énergie calculée est négative, plus l'affinité du ligand est grande : un bon coefficient de corrélation doit donc être proche de -1.

Seule la variante **a** a donné de bonnes corrélations entre les affinités calculées sur les géométries issues des protocoles 1 et 2 (respectivement R' = -0.80 et R' = -0.93) et les affinités expérimentales (Figure 4.7, **A1** et **A2**). La seule différence entre les protocoles 1 et 2 est l'étape 2 ou 2' de minimisation des géométries extraites de la trajectoire stable de chaque ligand : les géométries ont été minimisées avec la méthode de mécanique moléculaire (CHARMM, protocole 1 de la Figure 4.3) ou de chimie quantique (MOPAC, protocole 2 de la Figure 4.3). En revanche, alors que les protocoles 2 et 3 ne diffèrent que par la méthode d'échantillonnage, les graphes **A2** et **A3** associés indiquent une corrélation légèrement meilleure dans le cas **A2** et, surtout, une nette amélioration de cette corrélation dans le cas **A2** lorsque les ligands **1** et **16** sont exclus (R' = -0.93). Il est à noter d'ailleurs que, quelque soit la variante utilisée (**a** à **d**), la corrélation mesurée est meilleure pour le protocole 2 (différence mineure de 0.1 à 0.15 entre les coefficients, mais notable dans tous les cas). Il semblerait donc que le protocole 2 dans lequel l'échantillonnage se fait de manière équidistante dans la trajectoire soit le plus approprié.

Cette constatation est également validée par la comparaison des graphes **A2**/**A4**, **B2**/**B4**, **C2**/**C4** et **D2**/**D4**. Ces couples de graphes montrent en effet des coefficients

de corrélation très similaires et des énergies absolues de complexation pour chacun des ligands qui sont quasiment identiques, sauf pour le ligand **15** et, dans une moindre mesure, également pour le ligand **3**. La comparaison de ces graphes suggère donc que le gain en précision pour le protocole 2 qui serait apporté par l'utilisation d'énergie de complexation calculées à partir de géométries centre de cluster (protocole 3) est faible voir inexistant.

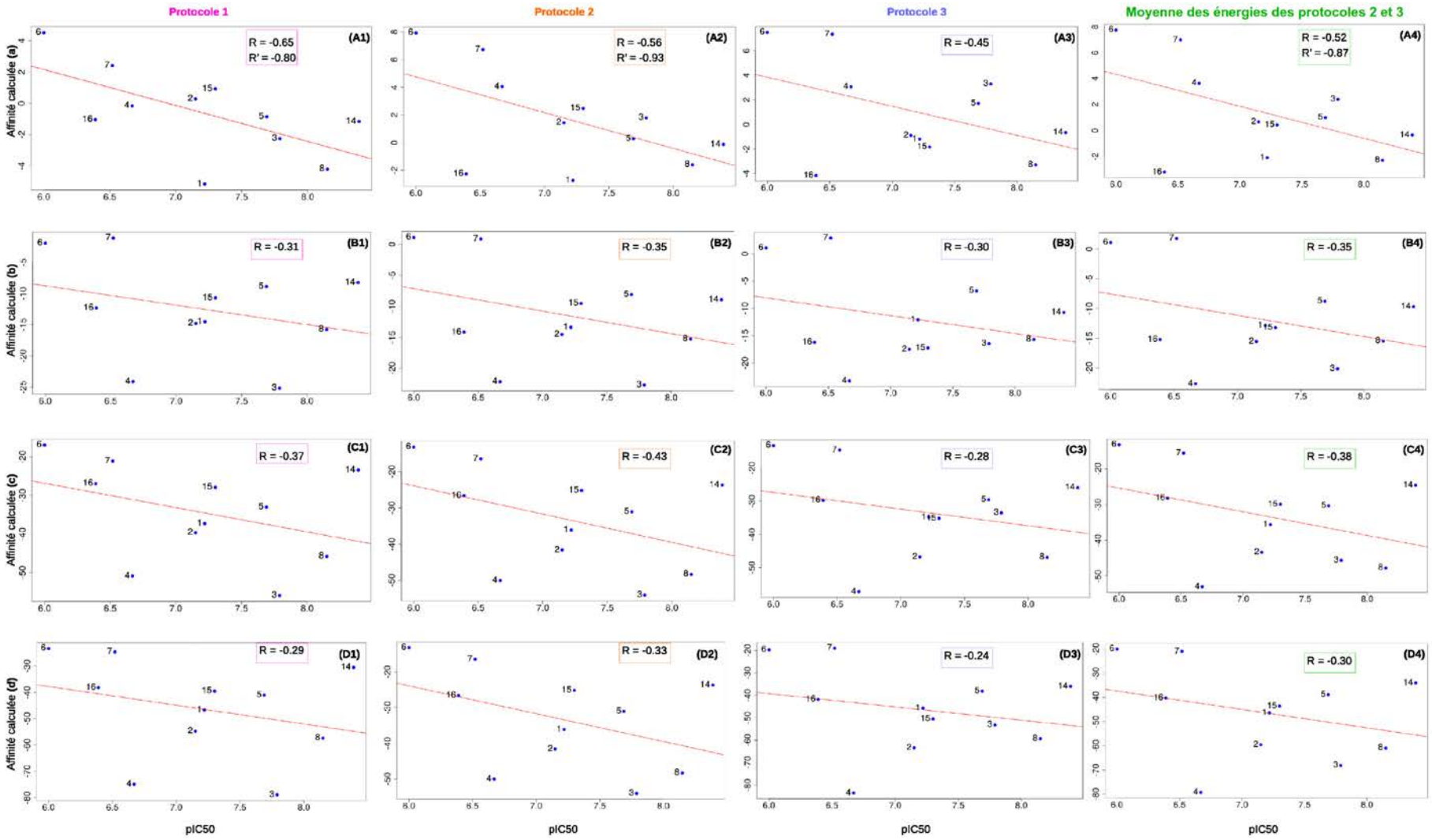


Figure 4.7: Corrélations entre les affinités calculées obtenues par les différentes variantes de calculs testés (**a**, **b**, **c** et **d**) et les affinités expérimentales (pIC50). Les graphes de corrélation **AX**, **BX**, **CX** et **DX** représentés sur chaque ligne correspondent respectivement aux résultats obtenus avec la variante de calcul **a** (FMO-PCM-MP2), **b** (FMO-PCM-MP2 + G_{pold}), **c** (FMO-PCM-SCS-MP2) et **d** (FMO-PCM-SCS-MP2 + G_{pold}) et les protocoles X (avec X = 1, 2, 3) présentés dans la Figure 4.3 (p. 109). Dans les graphes **A4**, **B4**, **C4** et **D4** l'énergie associé à chaque ligand a été obtenue en faisant la moyenne sur les 10 énergies calculées sur les 10 géométries issues du protocole 2 et sur les *n* énergies calculées sur les *n* géométries issues du protocole 3 (noté protocole 2+3 dans le texte du manuscrit). Le coefficient de corrélation R a été obtenu sur les 17 ligands et le coefficient de corrélation R' a été obtenu en écartant les ligands **1** et **16**.

Afin de mieux cerner l'amplitude des variations d'énergie en fonction des différentes conformations qu'adoptent les ligands et la protéine au cours de la trajectoire, nous avons réalisé des calculs FMO-PCM-MP2 sur deux géométries issues d'un même cluster (géométries intra-cluster) et sur deux géométries issues de deux clusters différents (géométries inter-cluster) pour le ligand **14**. Pour la comparaison des énergies calculées sur les deux géométries intra-cluster, nous avons choisi une géométrie centre d'un cluster de 109 géométries et une géométrie dont le RMSD par rapport à la géométrie centre de cluster est de 0.40 Å. Les énergies d'interaction ont été calculées avec la variante FMO-PCM-MP2 après avoir minimisé ces géométries en présence d'eau explicite avec MOPAC. Les calculs révèlent une différence d'énergie de 5.39 kcal/mol (en valeur absolue) entre ces deux géométries intra-cluster. Pour la comparaison des énergies calculées sur les géométries inter-cluster, nous avons choisi la géométrie centre d'un cluster de 214 géométries et la géométrie centre d'un cluster de 24 géométries. Des calculs du même type que précédemment sur ces deux géométries révèlent une différence d'énergie de 8.56 kcal/mol en valeur absolue.

Dans les deux cas de comparaison d'énergie (intra-cluster et inter-cluster), on observe de très grandes différences d'énergie, ce qui montre que les énergies d'interaction calculées avec FMO-PCM-MP2 sont très sensibles aux variations de conformations du complexe protéine-ligand au cours de la trajectoire puisque des variations importantes sont observées entre géométries d'un même cluster. Ainsi, une petite variation de la conformation du complexe peut donc significativement changer la valeur de l'énergie d'interaction calculée avec la méthode FMO-MP2. La variation d'énergie de complexation entre conformations intra- ou inter-cluster étant du même ordre, la stratégie qui consiste à utiliser des clusters de géométries représentatives n'est pas adaptée, en tout cas, à cette méthode de calcul FMO-MP2. La méthode d'échantillonnage équidistante semble donc bien être la plus adaptée puisqu'à elle seule (protocoles 1 et 2, variantes **a**), elle permet d'obtenir une corrélation satisfaisante avec les affinités expérimentales sans

même la contribution apportée par l'autre type d'échantillonnage.

Les meilleures corrélations ont été obtenues avec la variante de calcul **a** appliquée aux géométries issues des protocoles 1, 2 et 2+3 correspondant respectivement aux graphes de corrélation **A1** (R' = -0.80), **A2** (R' = -0.93) et **A4** (R' = -0.87) de la Figure 4.7 et sans considérer les ligands **1** et **16** pour lesquels les énergies ont été surestimées. Ces deux ligands ont la particularité d'avoir pour seul groupement, sur le cycle phényl de leur partie nord, un atome de chlore (ligand **1**, Figure 4.8a) ou un atome de fluor (ligand **16**, Figure 4.8b) en position para. Il est possible que les interactions de dispersion et/ou de polarisation établies entre ces atomes et les résidus Val83 et Leu85 aient été mal évaluées du fait de la base 6-31G utilisée qui doit être insuffisante dans ces cas (particulièrement pour décrire les effets de polarisation dus à ces atomes d'halogène et/ou la polarisation de l'atome de chlore).

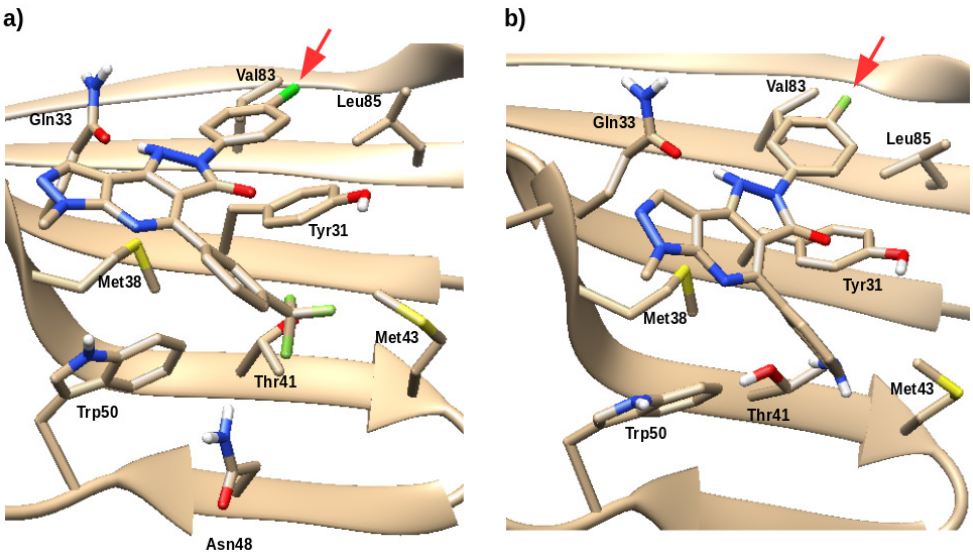


Figure 4.8: Complexes entre CD80 tronqué et les ligands **1** (a) et **16** (b). Ces ligands ont la particularité d'avoir un atome de chlore (ligand **1**, a) ou de fluor (ligand **16**, b) en position para sur le cycle phényl qui constitue leur partie nord. Ces atomes sont pointés par une flèche rouge.

Plusieurs conclusions peuvent être tirées de ces analyses : (i) étant donné que les résultats sont similaires lorsque les géométries extraites des trajectoires sont minimisées avec CHARMM ou MOPAC, il est préférable de privilégier la minimisation avec la méthode de mécanique moléculaire qui est beaucoup plus rapide qu'avec la méthode de chimie quantique (soit respectivement un temps de calcul de 10 minutes et de 24 heures pour la minimisation de 10 géométries d'un seul ligand) ; (ii) un échantillonnage systématique est plus approprié qu'un échantillonnage par clusterisation lorsque les calculs d'énergie de complexation sont réalisés avec la méthode FMO-MP2 ; (iii) au-delà d'un certain nombre de géométries pour le calcul de l'affinité moyenne, la corrélation avec les affinités expérimentales semble être quasiment constante puisque celles-ci sont similaires que les calculs d'affinités soient faits sur 10 ou sur 10 + n géométries par ligand (respectivement le protocole 2 et la combinaison des résultats issus des protocoles 2 et 3), ce qui laisse penser que, dans le cas de ce système en tout cas, la limite de 10 géométries reste un bon compromis.

Enfin lorsqu'on effectue des calculs d'affinités avec les variantes **b**, **c** ou **d** sur les géométries issues des protocoles 1, 2 ou 3 pour chaque ligand, on constate que les affinités sont plus élevées (énergies très négatives) que celles calculées avec la variante **a**. Cependant, avec les variantes de calcul **b**, **c** et **d**, les corrélations entre les affinités expérimentales et les affinités calculées n'excèdent pas -0.43.

Bien que la variante de calcul **a** permette de décrire correctement les interactions CD80-ligand, nous ne pouvons pas l'utiliser dans le criblage virtuel prévu pour Lu étant donné que le calcul d'énergie d'un seul complexe avec la méthode FMO-PCM-MP2 prend plusieurs heures. Le fait qu'une bonne corrélation soit obtenue avec les méthodes de calculs XSCORE, MM-PBSA et FMO ne peut être dû au hasard. Les résultats de ce chapitre ainsi que ceux du précédent nous permettent donc de valider le site de liaison ainsi que la pose des ligands sur CD80.

4.3 Analyse des interactions par résidu

Dans la section 3.1 (p. 75), nous avons présenté la pose que nous avons sélectionnée pour les ligands **1** à **17** sur CD80. Grâce à la méthode de mécanique moléculaire, nous avions décrit les interactions établies dans les complexes CD80-ligand issus de DOCK6 : les ligands font des interactions avec plusieurs résidus, dont des liaisons hydrogène avec les résidus Gln33 et Thr41, des interactions π-π avec les résidus Trp50 et Tyr31 notamment.

Afin d'évaluer la force des interactions entre ces résidus et les ligands dans la pose sélectionnée, nous avons calculé les énergies d'interaction entre chaque résidu de CD80 impliqué dans des interactions avec les ligands **1** - **8** et **14** - **16** grâce à la méthode FMO-RI-MP2 avec le programme PAICS [67] en utilisant la même base 6-31G. La méthode RI-MP2 est une méthode environ 10 fois plus rapide que MP2 avec une précision qui est équivalente à MP2 (la différence en énergies absolues pour une protéine est de l'ordre de $\sim$0.005 %) [66]. L'idée générale de cette méthode est de remplacer les intégrales très nombreuses à quatre centres (c'est-à-dire impliquant quatre orbitales atomiques) par une expression les liants aux intégrales à trois et deux centres qui sont moins nombreuses. Les énergies par résidu dans chacun des complexes ont été obtenues selon les étapes suivantes : (1) extraction de 10 géométries équidistantes sur la trajectoire stable de chaque ligand (les mêmes que celles des sections précédentes) ; (2) minimisation des géométries avec MOPAC en utilisant la méthode PM6-DH2X en présence d'eau explicite ; (3) calculs des énergies avec la méthode FMO-RI-MP2 pour les géométries dépourvues d'eau explicite ou implicite (dans le vide) afin de ne s'intéresser qu'aux interactions entre la protéine et le ligand. L'étape (2) de minimisation et l'étape (3) de calculs d'affinités citées ici ont été réalisées sur un système de taille réduite dans lequel CD80 a été tronqué, tel qu'indiqué dans les protocoles 2 et 3 de la Figure 4.3 (p. 109). La stratégie utilisée pour la minimisation de ces géométries est la même que celle qui est présentée dans la

Figure 4.5 (p. 112).

Le Tableau 4.1 montre les énergies d'interaction par résidu qui ont été obtenues en faisant la moyenne de ces énergies par résidu calculées pour chaque géométrie extraite de la trajectoire de chaque ligand (soit 10 géométries par ligand). Seuls les résidus qui ont des énergies d'interaction inférieures à -1 kcal/mol ont été représentés dans le tableau : Tyr31, Gln33, Lys37, Met38, Trp50, Val83 et Thr41. Ces énergies ont été représentées graphiquement dans la Figure 4.10.

Ligand	Tyr31	Gln33	Lys37	Met38	Trp50	Val83	Thr41
1	-8.84 ± 1.64	-4.27 ± 0.47	-3.07 ± 0.74	-6.02 ± 1.26	-2.62 ± 1.22	-2.07 ± 0.38	_
2	-3.14 ± 0.97	-4.97 ± 1.53	-4.79 ± 1.10	-10.26 ± 1.15	-2.78 ± 0.54	-1.25 ± 0.63	_
3	-4.80 ± 1.56	-4.71 ± 0.60	-3.59 ± 0.71	-5.80 ± 1.49	-1.61 ± 0.71	-2.50 ± 1.36	_
4	-4.13 ± 1.30	-3.89 ± 1.18	-3.97 ± 0.50	-8.65 ± 1.38	-3.26 ± 1.10	-1.79 ± 0.63	_
5	-7.54 ± 1.08	-4.89 ± 0.40	-3.04 ± 0.41	-4.96 ± 0.51	-2.32 ± 0.53	-1.96 ± 0.47	_
6	-9.01 ± 1.19	_	_	-3.40 ± 1.10	-3.51 ± 0.77	-1.07 ± 0.41	_
7	-4.54 ± 2.99	-3.45 ± 1.26	-3.53 ± 1.23	-8.66 ± 1.57	-3.38 ± 1.46	-1.89 ± 0.43	_
8	-3.22 ± 1.55	-5.96 ± 1.30	-4.89 ± 0.81	-10.67 ± 0.58	-2.90 ± 1.11	-1.62 ± 0.88	_
14	-7.10 ± 1.46	-4.86 ± 0.76	-3.99 ± 1.05	-10.05 ± 1.18	-4.26 ± 1.25	-1.85 ± 0.55	-2.15 ± 0.76
15	-6.35 ± 2.14	-4.45 ± 0.41	-1.88 ± 0.57	-5.15 ± 1.48	-3.24 ± 1.13	-2.20 ± 0.41	-1.47 ± 0.98
16	-8.37 ± 1.88	-3.98 ± 1.55	-2.34 ± 1.38	-6.60 ± 2.17	-4.67 ± 1.20	-2.03 ± 0.46	-2.66 ± 0.20

Tableau 4.1: Énergies d'interaction entre les résidus Tyr31, Gln33, Lys37, Met38, Trp50, Val83 et Thr41 et les ligands **1** à **8** et les ligands **14**, **15** et **16**.

D'après la Figure 4.10 et le Tableau 4.1, nous pouvons remarquer que les valeurs d'énergie d'interaction comparées entre ligands sont relativement similaires pour les chaînes latérales des résidus Val83, Trp50 et Gln33 et le squelette peptidique du résidu Lys37. Celles calculées pour les résidus Asn48 et Met43, restent faibles (Figure 4.9 ; valeurs supérieures à -0.6 kcal/mol) et ne sont pas présentées dans le Tableau 4.1. Les énergies d'interaction pour les ligands **1** à **8** et le résidu Thr41 sont faibles et n'excèdent pas -0.9 kcal/mol. Les deux résidus Thr41 et Asn48 sont tour à tour impliqués dans des liaisons hydrogène très faibles avec les atomes de fluor du groupement CF_3 se trouvant dans la partie sud des ligands **1** à **8** (Figure 3.1, p. 78 et Figure 4.9, à gauche). Une liaison hydrogène avec un atome de fluor a typiquement une énergie de -1.6 kcal/mol, soit une énergie de liaison de moitié plus faible que celle de la liaison hydrogène classique (environ -3.2 kcal/mol) [112]. Pour les ligands **1** à **8** les faibles énergies que nous observons pour les liaisons hydrogène avec les atomes de fluor du groupement CF_3 sont probablement dues à des distances non optimales entre ces atomes de fluor et les atomes d'hydrogène des groupements OH et NH_2 des résidus respectifs Thr41 et Asn48. C'est le cas par exemple du ligand **2** pour lequel les distances entre un atome de fluor du groupement CF_3 et l'azote du groupement NH_2 du résidu Asn48 varient entre 2.93 et 4.5 Å dans les géométries sur lesquelles les calculs FMO-RI-MP2 ont été réalisés. En revanche, pour les ligands **14** à **16**, une forte liaison hydrogène est formée entre le résidu Thr41 et le groupement NO_2 des ligands **14** et **15**, d'une part, et entre ce même résidu et le groupement NH_2 du ligand **16** d'autre part (Figure 4.9, à droite). Les valeurs d'énergie d'interaction correspondantes sont inférieures à celles observées pour les ligands **1** à **8**, soit une énergie moyenne entre -1.5 et -2 kcal/mol pour les ligands **14** à **16** contre une énergie moyenne entre -0.5 et -0.9 kcal/mol pour les ligands **1** à **8**.

On constate aussi que les résidus Tyr31 et Met38 forment les interactions les plus fortes avec les ligands et, pour ces résidus, nous observons des variations d'énergie entre ligands plus importantes que pour les résidus Val83, Trp50, Lys37 et Gln33 (Figure 4.10

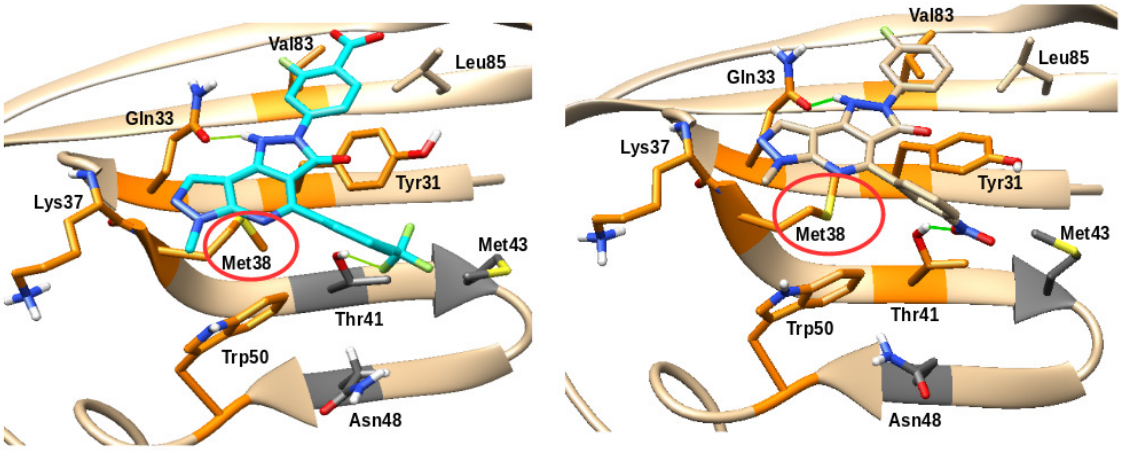


Figure 4.9: Résidus pour lesquels une forte énergie d'interaction (résidus en orange) et une faible énergie d'interaction (résidus en gris) ont été calculées avec la méthode FMO-RI-MP2. Deux exemples de ligands sont pris ici : le ligand **4** (à gauche) pour lequel une faible liaison hydrogène est établie entre son groupement CF_3 et le résidu Thr41 (tout comme pour les ligands **1** à **8**) et le ligand **14** (à droite) pour lequel une forte liaison hydrogène est établie entre son groupement NO_2 et ce même résidu (tout comme pour le ligand **15**). Dans le cas du ligand **16**, c'est le groupement NH_2 qui fait une forte liaison hydrogène avec le résidu Thr41. Les liaisons hydrogène sont montrées en trait plein vert. Le résidu Met48 (entouré en rouge) interagit très fortement avec le cœur des ligands dans les deux exemples, mais c'est la conformation qu'il adopte avec le ligand **14** (à droite) qui est connu comme étant la plus stabilisante.

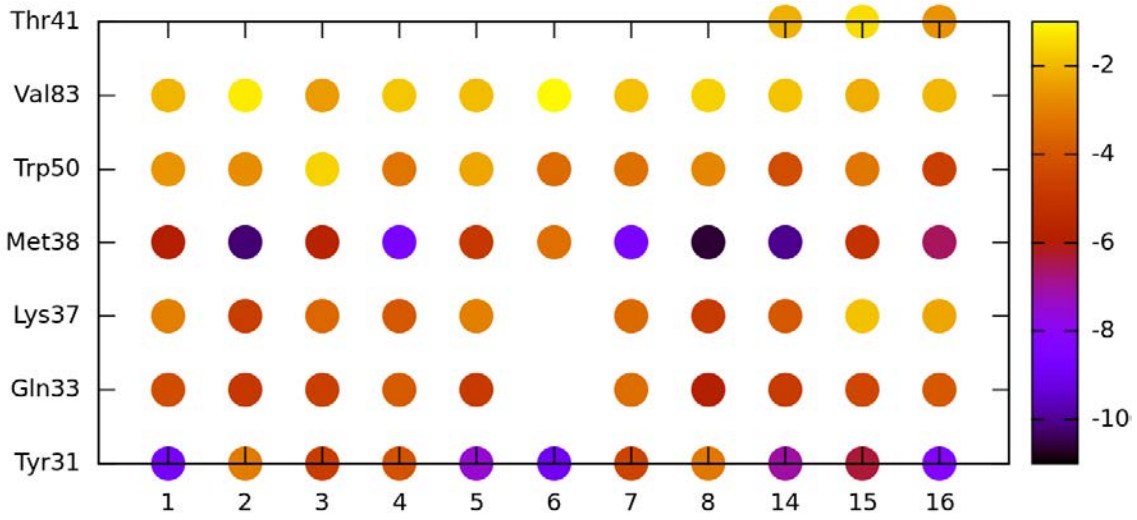


Figure 4.10: Représentation graphique des énergies d'interaction calculées entre chaque résidu et les ligands **1** - **8** et **14** - **16** (en abscisse). L'échelle à droite de la Figure indique les valeurs d'énergie en kcal/mol.

et Tableau 4.1). Le résidu Met38 favorise la liaison des ligands en interagissant avec leur cœur. Dans la Figure 4.9, deux conformations du résidu Met38 peuvent être observée : (i) une conformation dans laquelle l'atome de soufre est plus proche du cœur (ligand **4**, à gauche de la Figure 4.9) et (ii) une conformation dans laquelle l'atome de soufre est orienté vers l'intérieur de la protéine (ligand **14**, à droite de la Figure 4.9). Bien que ces deux conformations de la Met38 favorisent la liaison des ligands étudiés ici, la conformation (ii) mentionnée ci-dessus est celle qui est connue (c'est-à-dire celle qui est le plus souvent observée dans les structures cristallographiques de complexes protéine-ligand) comme étant de plus forte énergie et qui participe fortement à la stabilité des complexes [113]. Pour tous les ligands étudiés ici, les complexes initialement utilisés pour les simulations DM contiennent un résidu Met38 qui adopte la conformation (ii) citée ci-dessus (conformation montrée à droite de la Figure 4.9). Puis, le résidu Met38 change de conformation au cours des simulations DM. Dans les complexes formés par

les ligands **2**, **8** et **14**, le résidu Met38, qui a une valeur d'énergie d'interaction de -10 kcal/mol dans ces complexes, adopte toujours la conformation (ii) alors que dans ceux formés par les ligands **1**, **3**, **5**, **6**, **15** et **16** (valeurs d'énergie d'interaction du résidu Met38 entre -3 et -6 kcal/mol), le résidu Met38 adopte la conformation (i) (soit la conformation montrée à gauche de la Figure 4.9) dans les géométries sélectionnées pour les calculs FMO-RI-MP2. Dans les géométries sélectionnées pour les calculs FMO-RI-MP2 pour les ligands **4** et **7**, on constate que le résidu Met38 adopte les deux conformations (i) et (ii) ce qui explique la valeur d'énergie intermédiaire de -8 kcal/mol pour ce résidu dans les complexes formés par ces ligands.

Le résidu Tyr31 interagit selon des interactions π-π avec le cycle de la partie nord de l'ensemble des ligands. Ce cycle est décoré de différents substituants qui influencent différemment la densité électronique de ce cycle du ligand. Ainsi, la présence d'un substituant influence de deux manières ces interactions π-π, à la fois par un déplacement latéral dû à une interaction supplémentaire apportée par ce substituant et aussi par la force des interactions π-π qui s'en trouve modifiée.

Étant donné que la surface est plane, une substitution sur le ligand peut facilement entraîner un mouvement de translation sur la surface, favorisant une nouvelle interaction ou une interaction plus forte avec un résidu (Figure 4.11 où l'extrémité 1 ou 2 entraînera un déplacement spécifique et une interaction respectivement vers le résidu X ou Y). Ce déplacement, même faible, en surface entraîne des variations de l'énergie d'interaction par résidu telles que l'illustrent les variations d'énergie pour les résidus Tyr31 et Met38 qui sont trouvés "en dessous" de la surface du ligand. Une surface plane telle que celle trouvée généralement aux interfaces protéiques permet plus facilement un déplacement translationnel pour un ligand inhibiteur lié à une des surfaces protéiques qu'une poche d'interaction de type enzymatique qui offre plus de contraintes stériques autour du ligand. Cette observation étayée par nos calculs des énergies d'interaction permet de supposer que la relation structure–affinité dans le cas des inhibiteurs de PPI est plus difficile

à appréhender, du moins lorsque la surface liant l'inhibiteur est effectivement plane.

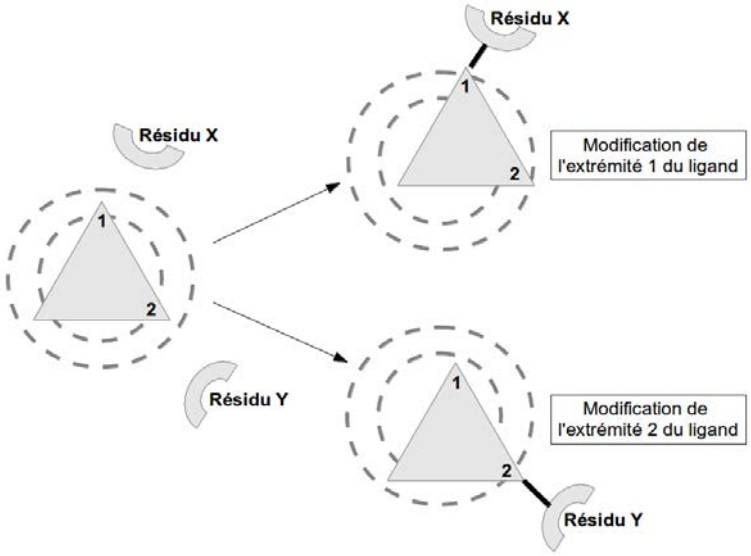


Figure 4.11: Déplacements du ligand (triangle) à partir d'une position de référence (partie gauche) suite à une substitution du côté 1 ou 2 par un groupement qui développera une interaction respectivement avec le résidu X ou Y. Ce déplacement entraîne des variations d'énergie d'interaction avec le résidu X ou Y mais aussi avec les résidus sous la surface du ligand (Tyr31 et Met38) qui, eux, n'interagissent pas nécessairement avec les groupements substitués du côté 1 ou 2.

Les calculs de chimie quantique réalisés dans ce chapitre apportent des informations quantitatives à l'analyse qualitative que nous avons réalisée après l'étape de docking (section 3.1, p. 75) pour les interactions établies entre CD80 et les ligands **1** - **8** et **14** - **16** dans la pose sélectionnée (Figure 3.2, p. 79). En effet, grâce à ces calculs, nous avons constaté que le résidu Met38 interagit de façon plus importante que ce que peut décrire la méthode de mécanique moléculaire : la méthode MP2 utilisée dans les calculs FMO permet de mieux décrire les interactions de dispersion entre le résidu Met38 et les cycles aromatiques du cœur des ligands au niveau électronique ainsi que les effets

de polarisation de la chaîne latérale de Met38 et/ou des cycles du ligand. Bien que les interactions de dispersion soient décrites par la fonction de Lennard-Jones en mécanique moléculaire, cette fonction décrit les interactions entre chaque atome du résidu et des ligands et non entre les atomes du résidu et la densité électronique du système π des cycles aromatiques du cœur des ligands. L'interaction entre un résidu Methionine et un cycle aromatique est une interaction plus forte que la somme des interactions de dispersion classiques et contribue fortement à la stabilité des complexes protéine-ligand et à la stabilisation des structures protéique [114]. En mécanique moléculaire, les cycles aromatiques sont "mimés" par l'utilisation du type atomique aromatique mais les électrons de type π n'existent pas réellement.

Chapitre 5

Recherche d'inhibiteurs d'interaction de Lu par criblage virtuel

Le protocole validé sur CD80 a été appliqué au domaine D2 de Lu afin de trouver de potentiels PPII (inhibiteurs d'interaction protéine-protéine) capables d'inhiber l'interaction Lu-Lnα5 (étape (e) de la Figure 3, p. 17). Nous avons réalisé un criblage de 1 295 678 molécules issues de la banque de molécules ZINC. Les étapes de ce criblage sont présentées dans la Figure 1.1 (p. 33). Les molécules ont été sélectionnées selon (i) leur classement calculé par les fonctions de scoring primaire implémentées dans DOCK6 et rDOCK durant l'étape de docking (étape (b)), (ii) le nombre de liaison hydrogène établi avec Lu, (iii) leur stabilité sur Lu au cours des différentes étapes de relaxation (étape (c), équilibration (c_{eq}) et production (c_{pr})) et (iv) leur classement calculé dans une étape (d) par la fonction de scoring secondaire XSCORE après les étapes d'équilibration (d_{eq}) et de production (d_{pr}) des simulations DM (Figure 1.1, p. 33). Le programme MATCH permet de paramétrer des milliers de ligands de façon automatique en générant des paramètres CGENFF pour les ligands ; certains ligands n'ont toutefois pas pu être paramétrés. Nous reviendrons sur ce point plus loin dans ce chapitre. Les critères de sélection (ii) et (iv)

ci-dessus ont été motivés par les résultats que nous avions obtenus suite à un premier criblage de 395 601 molécules sur Lu pour lequel les résultats seront brièvement présentés dans le paragraphe qui suit.

Dans un premier criblage virtuel de 395 601 molécules issues de ZINC sur Lu (voir les détails du protocole et des résultats dans l'article fourni en Annexe), nous avions utilisé le programme DOCK6 pour l'étape de docking et la fonction de scoring secondaire XSCORE pour l'étape d'évaluation d'affinité des ligands. De ce criblage, 12 ligands ont été identifiés parmi lesquels le top-3 a été testé expérimentalement par Wassim El Nemer et Sylvie Cochet, membre de l'équipe expérimentale de notre UMR. Les ligands classés en rangs 1, 2 et 3 (top-3) forment respectivement trois liaisons hydrogène avec les résidus Met167, Thr174 et Thr190, trois liaisons hydrogène avec les résidus Thr174, Thr190 et Tyr192 et trois liaisons hydrogène avec les résidus Thr174, Arg176 et Thr188. Pour des raisons de confidentialité liées aux demandes de brevets afférentes, nous ne pouvons présenter dans le manuscrit les structures de ces ligands en interaction avec Lu. Parmi les ligands qui ont été testés expérimentalement, les ligands de rang 2 et de rang 3, qui seront respectivement nommés ligA et ligB dans la suite du manuscrit, ont montré une inhibition de l'interaction Lu-Lnα5 (le détail des tests expérimentaux se trouve en Annexe).

Étant donné que les deux ligands actifs forment trois liaisons hydrogène avec Lu (selon nos observations dans les trajectoires de simulation DM correspondants), nous avons décidé de considérer ce critère pour sélectionner les ligands dans le protocole de criblage de 1 295 678 molécules sur Lu (critère de sélection (ii) ci-dessus). Les détails du protocole de criblage de 1 295 678 molécules sur Lu sont présentés dans la Figure 5.1 et dans le paragraphe qui suit.

Dans le criblage de 1 295 678 molécules, nous avons d'abord réalisé une étape de docking (étape (b) de la Figure 5.1) et nous avons sélectionné les ligands selon différents critères aux étapes **(1)** à **(3)** de cette Figure. Puis, nous avons réalisé une étape de

relaxation qui contient une étape de chauffage de 60 ps, une étape d'équilibration (c_{eq}) de 100 ps et une étape de production (c_{pr}) de 200 ps (Figure 5.1) dans les conditions données dans la section 1.4 (p. 31). Après chacune des étapes d'équilibration (c_{eq}) et de production (c_{pr}), nous avons sélectionné les ligands selon : (i) leur stabilité au cours de la trajectoire grâce à des calculs de RMSD (étapes **(4)** et **(6)**, Figure 5.1, p. 136) et (ii) leur énergie de liaison (score) calculée avec la fonction de scoring secondaire XSCORE (soit les fonctions X-Score et HMScore dans les étapes respectives (d_{eq}) et (d_{pr}) de la Figure 5.1). Aux étapes **(4)** et **(6)**, un ligand a été considéré comme instable lorsque la moyenne de RMSD calculée entre la pose initiale (pose de docking) et la pose correspondante à chacune des géométries extraites de la trajectoire était supérieure à 5 Å. Une géométrie par picoseconde a été extraite de chaque trajectoire étudiée (c_{eq}) et (c_{pr}). Les calculs de RMSD ont été réalisés sur les 50 dernières picosecondes de la trajectoire d'équilibration et sur les 200 picosecondes de la trajectoire de production. Par conséquent, nous avons utilisé 50 géométries à l'étape **(4)** et 200 géométries à l'étape **(6)** pour les calculs de RMSD. La valeur de 5 Å a été choisie afin de ne pas exclure des ligands pour lesquels un ou plusieurs groupements périphériques subissent une réorientation au cours des simulations.

Avant de présenter les résultats obtenus dans les différentes étapes de criblage citées ci-dessus (soit les étapes présentées dans la Figure 5.1) dans la section 5.2, nous étudierons d'abord une comparaison réalisée entre les résultats obtenus avec une fonction de scoring primaire et ceux obtenus avec une fonction de scoring secondaire dans la section 5.1. Puis, dans cette même section, nous ferons une comparaison entre les résultats obtenus sur des géométries directement issues du docking et ceux obtenus sur les géométries extraites des trajectoires de simulation DM. Enfin, nous analyserons le top-20 des ligands issus du protocole de criblage (Figure 5.1, p. 136) dans la section 5.4.

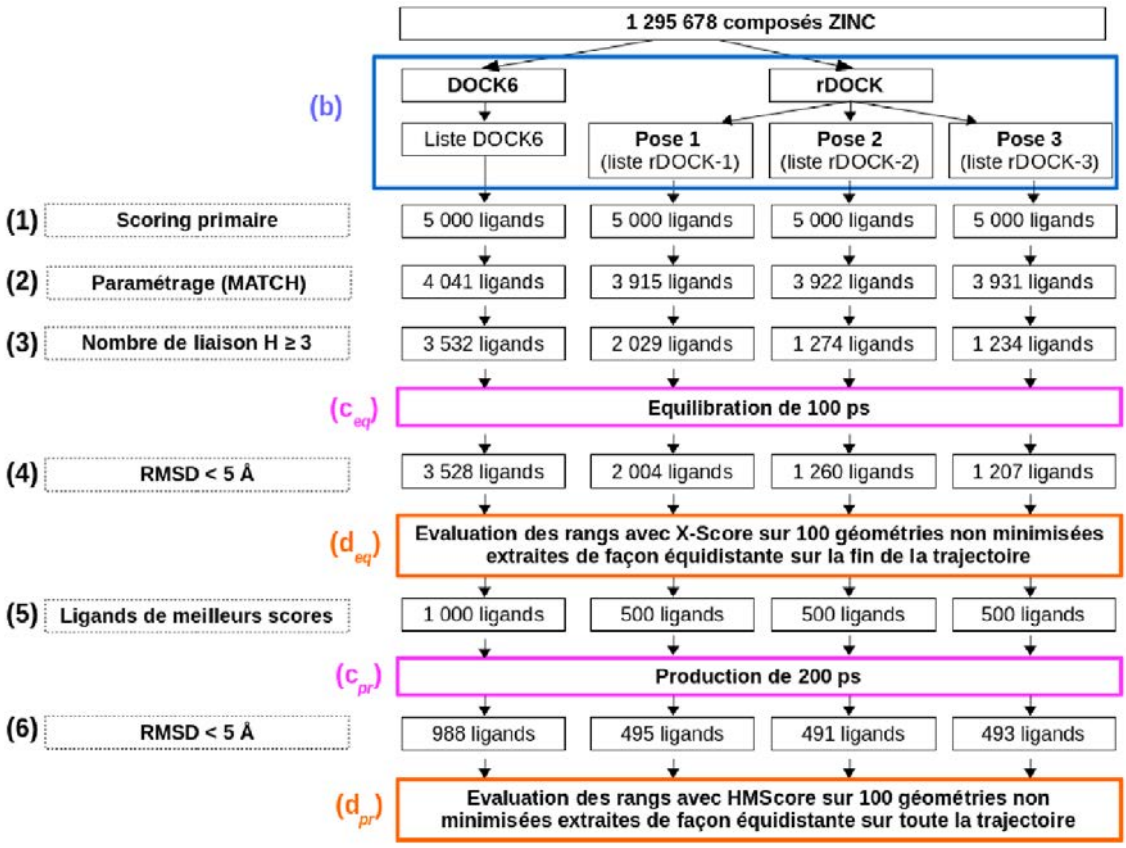


Figure 5.1: Dans le criblage de 1 295 678 de molécules, les deux programmes de docking DOCK6 et rDOCK ont été utilisés à l'étape **(b)** de docking. La liste DOCK6 est la liste de ligands dont les poses uniques ont été générées par DOCK6. Le programme rDOCK a généré trois poses par ligands (pose 1, pose, 2, pose 3). Les listes rDOCK-1, rDOCK-2 et rDOCK-3 sont les listes de ligands qui adoptent respectivement les poses 1, 2 et 3 générées par rDOCK pour chaque ligand. Dans chacune des listes, les ligands ont été sélectionnés suivant plusieurs critères: leur classement selon la fonction de scoring primaire à étape **(1)** dans lequel les 5 000 ligands de meilleur score ont été retenus ; la capacité de MATCH à pouvoir les paramétrer (étape **(2)**) ; le nombre de liaisons hydrogène qu'ils établissent avec Lu (seuls les ligands qui forment au moins trois liaisons hydrogène avec Lu sont sélectionnés à l'étape **(3)**) ; leur stabilité au cours des étapes c_{eq} d'équilibration et c_{pr} de production (étapes **(4)** et **(6)** dans lesquelles les ligands sont retenus si leur RMSD par rapport à la structure initiale est inférieur à 5 Å) ; leur rang calculé à l'étape d_{eq} et à l'étape d_{pr} sur 100 géométries extraites de façon équidistante respectivement sur la fin de la trajectoire d'équilibration et sur l'ensemble de la trajectoire de production. À l'étape **(5)**, seuls 1 000 ligands de la liste DOCK6 et 500 ligands de chacune des listes rDOCK pour lesquels les meilleurs scores ont été calculés avec X-Score à l'étape d_{eq} ont été retenus avant de passer à l'étape c_{pr} de relaxation.

.

5.1 Quel est l'intérêt d'utiliser une fonction de scoring secondaire et de réaliser une étape de relaxation dans un criblage virtuel ?

Dans le protocole de criblage appliqué à Lu pour rechercher de potentiels PPII capables d'inhiber l'interaction Lu-Lnα5, nous avons réalisé une étape de relaxation (étapes d'équilibration (c_{eq}) et de production (c_{pr}) de la Figure 5.1, p. 136) et deux étapes de calcul des énergies de liaison des ligands avec la fonction de scoring primaire XSCORE (étapes (d_{eq}) et (d_{pr}) de la Figure 5.1, p. 136). Dans cette section, nous présenterons d'abord l'intérêt de réévaluer les énergies des complexes issus du docking avec une fonction de scoring secondaire. Puis nous verrons l'importance de prendre en compte la flexibilité de ces complexes protéine-ligand (grâce à une étape de relaxation) dans les calculs d'énergie.

5.1.1 Quel est l'intérêt d'utiliser une fonction de scoring secondaire dans un criblage virtuel ?

Un programme de docking a pour fonction de générer des poses pour un à plusieurs milliers de ligands puis de les classer grâce à leur fonction de scoring (fonction de scoring primaire). Il doit donc associer efficacité et rapidité. Cela étant, il est difficile pour un programme de docking d'allier une fonction de scoring qui décrit de façon précise les interactions protéine-ligand et qui possède une grande vitesse de calculs : plus la fonction de scoring prendra des contributions énergétiques en compte, plus le temps de calcul sera long. Il est donc toujours pertinent de reclasser les ligands avec une fonction de scoring secondaire, toujours plus élaborée qu'une fonction de scoring primaire.

Sur un panel de 100 ligands choisi arbitrairement, nous avons voulu comparer les

rangs calculés à partir de la fonction de scoring primaire de DOCK6 avec ceux calculés, pour ces mêmes ligands dans les mêmes géométries respectives, avec la fonction de scoring secondaire HMScore de XSCORE (p. 51). Seule une géométrie a été utilisée dans les deux cas, laquelle n'a subi aucune minimisation supplémentaire à l'aide de CHARMM ou relaxation par DM. Les résultats sont présentés dans la Figure 5.2.

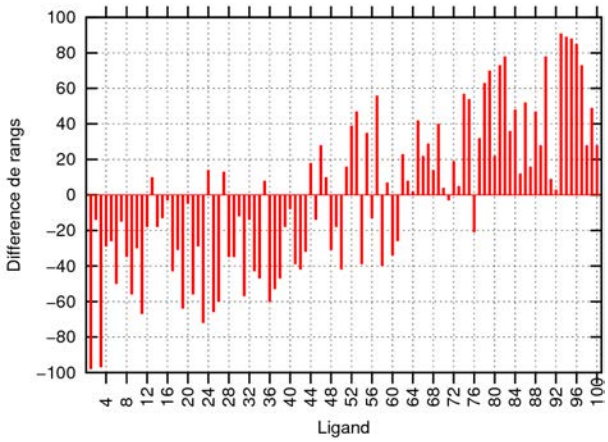


Figure 5.2: Différences entre les rangs obtenus avec la fonction de scoring primaire de DOCK6 et la fonction de scoring secondaire HMScore. Les calculs HMScore ont été appliqués sur les géométries de 100 ligands directement issus du docking avec DOCK6 (soit une seule géométrie par ligand non minimisée avec CHARMM et non relaxée avec une simulation DM).

Les différences entre les rangs attribués par la fonction de scoring primaire et ceux par la fonction de scoring secondaire sont en moyenne de 36.36 avec un écart-type de 43.66. On observe qu'un ligand classé en première position par une fonction de scoring primaire peut se retrouver en position 100 par une fonction de scoring secondaire et vice-versa (cas du ligand 1, Figure 5.2). Cela pourrait s'expliquer par les contributions prises en compte respectivement par chacune des fonctions de scoring. En effet, DOCK6 ne calcule que les contributions de van der Waals et électrostatiques alors que HMScore calcule

aussi les liaisons hydrogène, les contributions entropiques et les effets hydrophobes (voir p. 51).

Dans le cas de Lu, les ligands doivent se lier sur une surface plane. La présence de liaisons hydrogène entre ces ligands et Lu semblent jouer un rôle important dans la liaison des ligands. Il est donc pertinent de pouvoir calculer avec la fonction de scoring secondaire la contribution énergétique apportée. C'est pourquoi la fonction de scoring HMScore est bien adaptée au système étudié.

5.1.2 Quel est l'intérêt de prendre en compte la flexibilité des complexes dans les calculs de score ?

Dans le protocole de criblage, une étape de dynamique moléculaire (relaxation, soit les étapes (c_{eq}) et (c_{pr}) de la Figure 5.1, p. 136) permet de prendre en compte la flexibilité des complexes protéine-ligand issus du docking. Puis, des calculs de scores sont réalisés, par exemple, sur 100 géométries non minimisées extraites de façon équidistante de la trajectoire de production obtenue à l'étape (c_{pr}) de la Figure 5.1 (p. 136). Dans les simulations DM, des liaisons hydrogène se font et se défont, des contacts de van der Waals peuvent être présents à une picoseconde donnée puis être absents à la picoseconde suivante. Ces variations de contacts entre protéine et ligand dans la dynamique moléculaire montrent bien qu'une seule géométrie ne peut être représentative des différentes conformations qu'adopte le complexe pendant la dynamique.

Afin de montrer l'importance de la dynamique moléculaire et de son échantillonnage dans les calculs de scoring secondaire, nous avons réalisé, pour 100 complexes choisis arbitrairement, des calculs HMScore sur la géométrie initiale générée par DOCK6 (étape **(1)**, Figure 5.1, p. 136) et celui calculé sur 100 géométries extraites de façon équidistante de la trajectoire de production de 200 ps (obtenue à l'étape (c_{pr}), Figure 5.1, p. 136) pour chacun des 100 complexes protéine-ligand. La Figure 5.3 (p. 140) montre, pour chacun

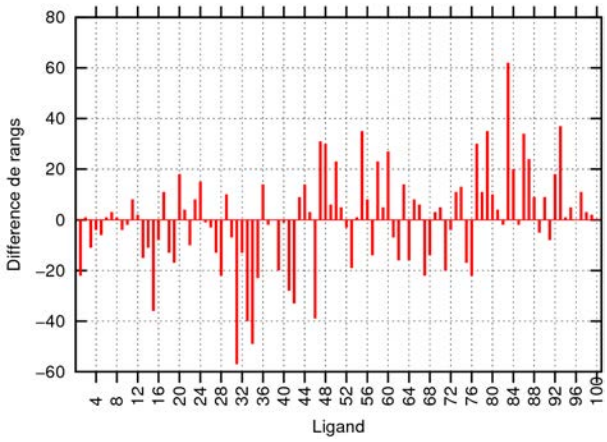


Figure 5.3: Différences entre les rangs HMScore des géométries générées par DOCK6 et les rangs HMScore des géométries issus des simulations. Dans ce dernier cas, le rang a été obtenu en faisant la moyenne des scores HMScore de 100 géométries non minimisées extraites de façon équidistante sur l'ensemble de la trajectoire de production.

des 100 complexes, la différence entre le rang calculé avec HMScore sur la géométrie initiale et sur les 100 géométries extraites de la trajectoire de production. Les différences de rangs sont en moyenne de 14.02 avec un écart-type de 18.95. Les différences de rangs calculés avec HMScore peuvent aller jusqu'à 57 et 62 (en valeurs absolues). Sur un panel de 100 ligands, ces différences ne sont pas négligeables et montrent l'importance de considérer les changements de conformation d'un complexe protéine-ligand dans un criblage virtuel. Cette idée a été évoquée par Merz et al dans d'autres études [115].

5.2 Résultats obtenus sur les différentes étapes du criblage

Dans cette section, nous présenterons les résultats dans chacune des étapes du protocole de criblage de la Figure 5.1 (p. 136).

Dans le criblage de 1 295 678 molécules sur Lu et dans l'idée de réaliser un consensus

docking, nous avons utilisé les deux programmes DOCK6 [34] et rDOCK [35] à l'étape (b) de la Figure 5.1 (p. 136). Contrairement à DOCK6, rDOCK est capable de prendre en considération une liste de molécules d'eau donnée lors des calculs de docking. La structure cristallographique de Lu présente des molécules d'eau avec de faibles B-factors (facteurs de température). En particulier, les B-factors des molécules d'eau 244, 257 et 264 sont respectivement de 20, 24 et 25 $Å^2$. De plus, les deux premières molécules d'eau citées forment des liaisons hydrogène avec Lu : HOH244 forme des liaisons hydrogène avec les résidus Thr140 et Glu137 ; HOH257 forme des liaisons hydrogène avec Gln136, Thr127, HOH264 et HOH414 (Figure 5.4). Enfin, la molécule d'eau HOH264 forme des liaisons hydrogène avec HOH343, HOH257 et l'azote du squelette peptidique du résidu Glu137 (code PDB 2PET, Figure 5.4). De faibles B-factors associés à un grand nombre de liaisons hydrogène semblent suggérer que les molécules d'eau ne seraient pas faciles à remplacer par un ligand sauf si celui-ci serait capable de reproduire le même réseau de liaisons hydrogène. Dans la suite, nous avons considéré cette hypothèse dans le cas du docking avec rDOCK.

Le programme rDOCK a permis de générer trois poses par ligands (pose 1, pose 2 et pose 3) et de prendre en compte les trois molécules d'eau cristallographique HOH244, HOH257 et HOH264 lors de la formation des complexes protéine-ligand afin d'étudier leurs contributions dans la liaison des ligands. En revanche, DOCK6 a permis de ne garder que la meilleure pose par ligand et n'offre pas la possibilité de prendre en compte les molécules d'eau citées ci-dessus. Dans la suite du manuscrit, les ligands dont les poses ont été générées par DOCK6 seront appelés «ligands DOCK6 » et on dira qu'ils font partie de la « liste DOCK6 ». Les ligands dont les poses ont été générées par rDOCK seront appelés « ligands rDOCK-x » et on dira qu'ils font partie de la « liste rDOCK-x » avec x qui prend la valeur de 1, 2 ou 3 si le ligand adopte respectivement la pose 1 (meilleure énergie), la pose 2 (deuxième meilleure énergie) ou la pose 3 (troisième meilleure énergie). Ces listes de ligands (liste DOCK6, liste rDOCK-1, liste rDOCK-2

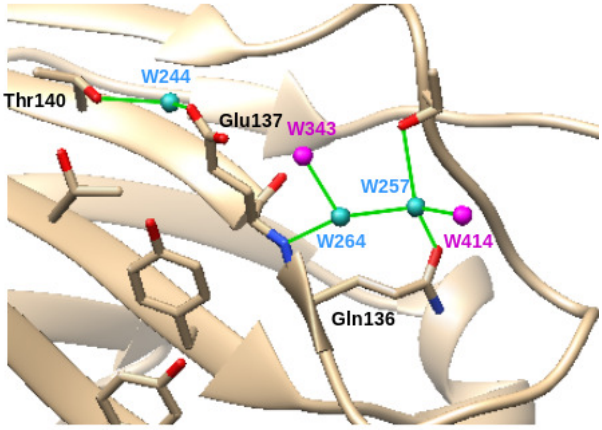


Figure 5.4: La structure cristallographique de Lu possède des molécules d'eau HOH244 (W244), HOH257 (W257) et HOH264 (W264) qui ont des valeurs de B-factor inférieures à 25 $Å^2$ (en cyan). Les molécules d'eau partenaires HOH343 (W343) et HOH414 (W414) sont aussi montrées en magenta.

et liste rDOCK-3) ont été analysées parallèlement dans les étapes suivantes du criblage (Figure 5.1, p. 136).

Une première étape de sélection (étape **(1)**, Figure 5.1, p. 136) consistait à ne retenir que 5 000 ligands de chaque liste (listes DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3) pour lesquels les meilleurs scores ont été calculés par les fonctions de scoring primaire implémentées dans les programmes de docking correspondants (voir p. 42 et 47).

Puis, afin de relaxer les complexes et prendre en compte la flexibilité de la protéine et du ligand, nous avons réalisé une étape de simulation par dynamique moléculaire (voir section 1.4, p. 31) dans laquelle des ligands ont été sélectionnés après l'étape d'équilibration et après l'étape de production (étapes (c_{eq}) et (c_{pr}), Figure 5.1, p. 136). Pour réaliser cette étape de dynamique moléculaire, nous avons paramétré les ligands (étape **(2)**, Figure 5.1, p. 136) avec le champ de force CGENFF en utilisant le programme MATCH [68] qui n'a pu paramétrer que 4 041 ligands DOCK6, 3 915 ligands rDOCK-1,

3 922 ligands rDOCK-2 et 3 931 ligands rDOCK-3. Cela tient probablement de types atomiques qui ne sont pas pris en compte ou d'états de protonation incorrects pour ces ligands. Puisque plusieurs milliers de ligands n'ont pas pu être paramétrés avec MATCH, nous avons décidé de ne pas traiter ces ligands manquants dans la suite du criblage.

Les étapes de simulation DM (c_{eq}) et (c_{pr}) de la Figure 5.1 (p. 136) ont été réalisées sur des ligands qui font au moins trois liaisons hydrogène avec Lu. Notre choix de ne retenir que les ligands qui respectent ce critère (étape **(3)**, Figure 5.1, p. 136) s'explique pour deux raisons. D'abord, la présence de liaisons hydrogène entre un ligand et le site de liaison présentant une surface plane semble être importante pour la stabilité du ligand (c'est ce que nous avons pu observer dans le cas de CD80 et ses ligands). En effet, un tel site de liaison ne possède pas de poche profonde dans laquelle les mouvements du ligand seraient contraints (comme c'est le cas pour les enzymes et la plupart des récepteurs) [116]. Puis, pour les deux ligands validés par les tests in vitro, le nombre de liaisons hydrogène avec Lu est de cinq pour l'un et quatre pour l'autre [1]. Nous avons donc décidé de continuer la suite du criblage avec les ligands formant au moins trois liaisons hydrogène avec la protéine afin de ne pas éliminer précocement les ligands qui peuvent former au moins une liaison hydrogène supplémentaire lors de l'étape de relaxation. Suite à l'étape de sélection **(3)**, il reste donc respectivement 3532, 2029, 1274 et 1234 ligands dans les listes DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3.

Aux étapes **(4)** et **(6)**, nous avons évalué la stabilité des ligands respectivement sur la fin de la trajectoire d'équilibration (les 50 dernières picosecondes) obtenue à l'étape (c_{eq}) et sur l'ensemble de la trajectoire de production obtenue à l'étape (c_{pr}) (Figure 5.1 (p. 136). Pour cela, nous avons réalisé des calculs de RMSD (Root Mean Square Deviation) et d'écart-type sur les géométries extraites sur chacune de ces portions de trajectoire. Les ligands ont été considérés comme instables lorsqu'ils présentaient des valeurs de RMSD supérieures à 5 Å par rapport à la pose initiale (pose de docking avant les simulations DM) et associées à des écarts types supérieurs à 1 Å. En ne gardant que les ligands qui

ont un RMSD < 5 Å, on estime que les algorithmes de docking DOCK6 et rDOCK ont correctement placé les ligands sur Lu sans qu'un repositionnement complet des ligands pendant la simulation DM ne soit nécessaire, tout en accordant une certaine flexibilité au ligand en particulier pour un ou plusieurs de ses groupements périphériques qui pourraient subir une réorientation complète suite à la torsion d'un ou plusieurs angles. À l'étape **(4)** (Figure 5.1, p. 136), seuls 4, 25, 14 et 27 ligands provenant des listes respectives DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3 se sont montrés instables sur les 50 dernières picosecondes de la trajectoire d'équilibration.

Puis, à l'étape (d_{eq}) et pour chacun des ligands restants, 100 géométries non minimisées ont été extraites de façon équidistante sur les 30 dernières picosecondes de la trajectoire d'équilibration obtenue à l'étape (c_{eq}) afin d'y appliquer les calculs X-Score. Les choix d'extraire des géométries (i) non minimisées et (ii) sur les 30 dernières picosecondes d'équilibration sont expliqués dans le paragraphe qui suit. Les calculs de scores réalisés à l'étape (d_{eq}) avaient pour but de réaliser une première sélection des ligands grâce à la fonction de scoring X-Score (implémenté dans la méthode de calcul XSCORE) afin de diminuer le nombre de ligands à traiter dans la suite du criblage. En effet, l'étape suivante est une étape de production de simulation DM de 200 ps pour lequel un temps de calcul d'environ 2 h est nécessaire pour chaque complexe protéine-ligand. Il était donc important de diminuer le nombre de ligands à traiter dans cette étape.

L'extraction de 100 géométries suivies de leur minimisation prend environ 15 minutes sur 16 cœurs contre 10 secondes pour l'extraction de 100 géométries non minimisées. Étant donné que cette opération devait se réaliser 7999 fois (somme de l'ensemble des ligands et/ou pose de ligands restants après l'équilibration), cela aurait pris environ 83 jours si l'on avait choisi de minimiser les géométries contre environ 22 heures pour extraire des géométries non minimisées. Il était donc important d'évaluer s'il y avait de grandes différences de résultats entre les calculs XSCORE appliqués sur les géométries minimisées et ceux appliqués sur les géométries non minimisées.

L'étape d'équilibration est l'étape pendant laquelle un système, désordonné par une augmentation de température lors de la phase de chauffage qui la précède, va s'adapter à son nouvel environnement et trouver un état d'équilibre. Le système est donc toujours moins stable au début d'une simulation d'équilibration qu'à sa fin. Dans cette logique, les géométries ont été extraites à la fin des simulations d'équilibration. Cependant, la portion de la trajectoire à extraire ne doit pas être choisie au hasard : il faut choisir la portion la plus pertinente du système équilibré. Ici encore, il a donc fallu évaluer si XSCORE (fonction de scoring X-Score ici) donnait des résultats différents lorsqu'il était appliqué sur une portion de la trajectoire plutôt qu'une autre. L'ensemble des tests réalisés est présenté dans la Figure 5.5.

Les tests ont été réalisés sur un panel de 100 ligands. Pour chacun des ligands, deux séries de tests ont été réalisées afin de répondre aux questions suivantes : (i) la minimisation des géométries a-t-elle un impact sur les résultats obtenus avec X-Score ? (Figure 5.5, partie orange) et (ii) dans quelle portion de la trajectoire doit-on extraire les géométries ? (Figure 5.5, partie jaune).

Pour répondre à la première question, 100 géométries (minimisées ou non) ont été extraites sur les 50 dernières picosecondes de la trajectoire d'équilibration (Figure 5.5, partie orange) pour chacun des 100 ligands. Les trajectoires d'équilibration pour ces 100 ligands sont celles qui ont été obtenues à l'étape (c_{eq}) de la Figure 5.1 (p. 136). Puis, des calculs de scores avec la fonction de scoring X-Score ont été réalisés sur chacune des 100 géométries équidistantes issues de la trajectoire du complexe formé par chaque ligand. Les 100 valeurs de score obtenues pour chaque complexe protéine-ligand ont été utilisées pour déduire un score moyen. Deux listes ont donc été obtenues : (i) une liste de scores calculés avec X-Score sur les géométries minimisées et (ii) une liste de scores calculés avec X-Score sur les géométries non minimisées pour chaque complexe protéine-ligand. Après avoir classé les ligands selon leur score moyen, des rangs associés à chaque ligand ont été obtenus. La Figure 5.6 montre les différences de rangs obtenus entre le

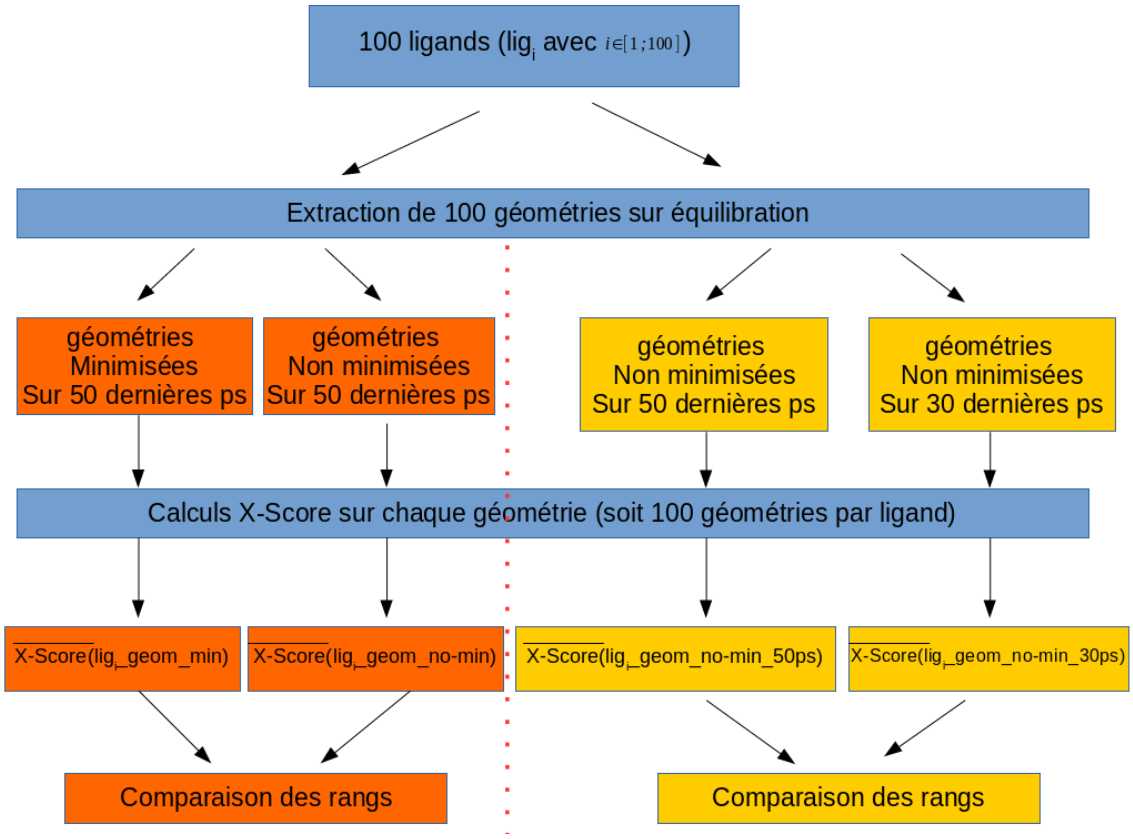


Figure 5.5: Étapes qui ont permis d'étudier l'impact de la minimisation (partie orange) et de l'échantillonnage (partie jaune) sur les résultats XSCORE pour un panel de 100 ligands choisis aléatoirement. Pour chaque ligand lig$_i$, 100 géométries ont été extraites sur les 50 dernières picosecondes de la trajectoire d'équilibration (parties orange et jaune) ou sur les 30 dernières picosecondes de l'équilibration (partie jaune). Les scores ont ensuite été calculés avec la fonction de scoring X-Score (implémentée dans la méthode de calcul XSCORE) sur les géométries minimisées ou non. Dans la partie orange, les scores moyens $\overline{X-Score}$(lig$_i$_geom_min) et $\overline{X-Score}$(lig$_i$_geom_no-min) ont été respectivement obtenues sur les 100 géométries minimisées et non minimisées extraites sur les 50 dernières picosecondes de la trajectoire d'équilibration. Dans la partie jaune, les scores moyens $\overline{X-Score}$(lig$_i$_geom_no-min_50ps) et $\overline{X-Score}$(lig$_i$_geom_no-min_30ps) ont été obtenus sur les 100 géométries non minimisées respectivement extraites sur les 50 et 30 dernières picosecondes de la trajectoire d'équilibration. Les scores moyens ont été utilisés pour réaliser un classement. Ces classements ont été comparés.

rang calculé avec X-Score (i) sur les géométries minimisées et (ii) sur les géométries non minimisées pour chaque complexe protéine-ligand. On constate que les différences de rangs sont en moyenne de 1.52 avec quelques ligands ayant une différence de rang allant de 5 à 8. Ces différences étant relativement faibles, la minimisation des géométries ne semble pas avoir d'impact significatif sur les résultats obtenus avec X-Score. On a donc décidé de travailler avec des géométries non minimisées dans la suite du criblage.

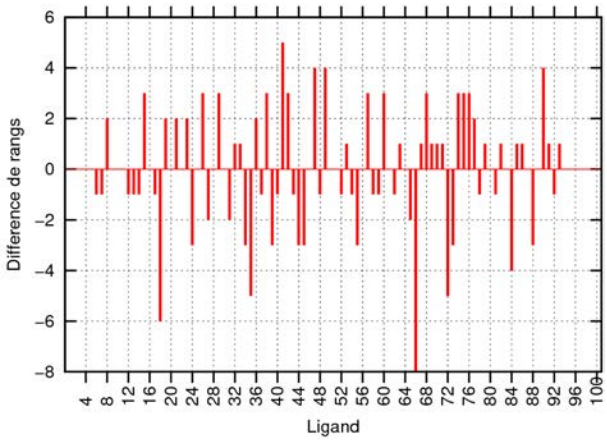


Figure 5.6: Différences entre les rangs X-Score obtenus sur des géométries non minimisées et ceux obtenus sur des géométries minimisées pour un panel de 100 ligands. Les géométries utilisées ont été extraites sur les 50 dernières picosecondes de l'équilibration. Les différences ont été obtenues en soustrayant les rangs des ligands respectifs.

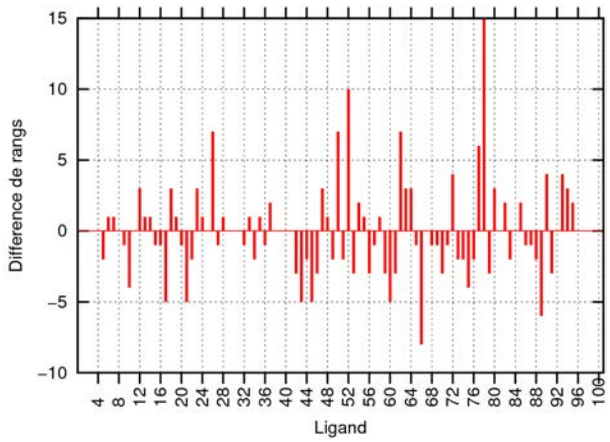
Le deuxième point consistait à choisir la bonne portion de la trajectoire à extraire. Pour cela, et toujours pour un panel de 100 ligands (soit 100 complexes protéine-ligand), 100 géométries non minimisées ont été extraites sur les 50 et les 30 dernières picosecondes de la trajectoire d'équilibration (Figure 5.5, partie jaune) obtenue à l'étape (c_{eq}) de la Figure 5.1 (p. 136). Puis, des calculs de scores avec la fonction de scoring X-Score ont été réalisés sur chaque géométrie afin d'avoir un score moyen pour chaque ligand. Deux listes ont été obtenues : (i) une liste de scores X-Score calculés sur les géométries

extraites sur les 50 dernières picosecondes, et (ii) une liste de scores X-Score calculés sur les géométries extraites sur les 30 dernières picosecondes d'équilibration pour chacun des 100 complexes protéine-ligand. La Figure 5.7a, montre les différences de rangs obtenus entre les rangs calculés avec X-Score (i) sur les géométries non minimisées extraites sur les 50 dernières picosecondes de la trajectoire d'équilibration et (ii) sur celles extraites sur les 30 dernières picosecondes de la trajectoire d'équilibration pour chaque complexe protéine-ligand. La différence moyenne est de 2.2, avec quelques ligands pour lesquels les différences de rangs varient de 5 à 15. Afin de faire un choix plus pertinent sur la portion de la trajectoire à extraire, les rangs calculés sur les géométries issus des 50 dernières picosecondes de la trajectoire d'équilibration ont été comparés aux rangs calculés sur les géométries issues de l'ensemble de la trajectoire de production (soit, pour chaque ligand, 100 géométries extraites de façon équidistante sur les 200 ps de la trajectoire de production) (Figure 5.7b). De la même façon, les rangs calculés sur les géométries issues des 30 dernières picosecondes de la trajectoire d'équilibration ont été comparés aux rangs calculés sur les géométries issues de l'ensemble de la trajectoire de production (Figure 5.7c).

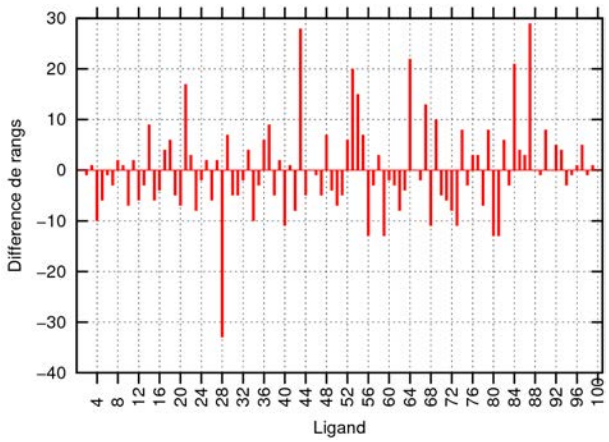
Pour la Figure 5.7b, les différences de rangs entre les géométries extraites sur les 50 dernières picosecondes de l'équilibration et les géométries extraites de la production sont en moyenne de 6.36 avec six ligands pour lesquels les rangs sont supérieurs à 20. En revanche, les différences de rangs entre les géométries extraites sur les 30 dernières picosecondes de l'équilibration et celles extraites de la production sont en moyenne de 6.44 avec quatre ligands pour lesquels les rangs sont supérieurs à 20 (Figure 5.7c). Étant donné qu'il n'y a pas de différences significatives entre les résultats obtenus sur les géométries extraites sur les 30 et sur les 50 dernières picosecondes de l'équilibration, nous avons choisi d'échantillonner sur les 30 dernières picosecondes de l'équilibration.

Après avoir obtenu un score moyen pour chaque ligand, score qui a été calculé sur 100 géométries non minimisées extraites de façon équidistante sur les 30 dernières pi-

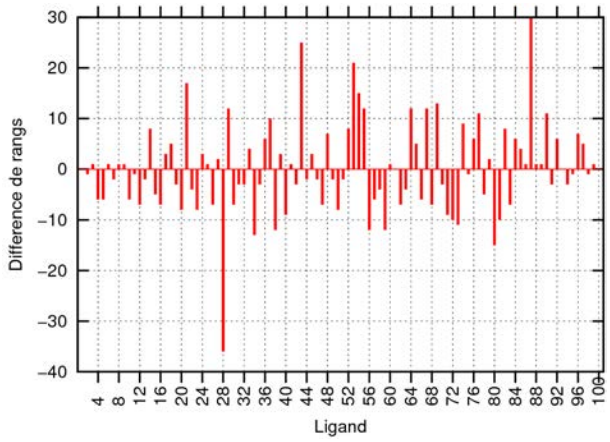
Figure 5.7: Différences entre (a) le rang calculé sur les géométries extraites sur les 30 dernières picosecondes et le rang calculé sur les géométries extraites sur les 50 dernières picosecondes de la trajectoire d'équilibration pour chaque complexe protéine-ligand ; (b) le rang calculé sur les géométries extraites sur les 50 dernières picosecondes de l'équilibration et le rang calculé sur les géométries extraites sur les 200 ps de la production pour chaque complexe protéine-ligand ; (c) le rang calculé sur les géométries extraites sur les 30 dernières picosecondes de l'équilibration et le rang calculé sur les géométries non minimisées extraites sur les 200 ps de la production pour chaque complexe protéine-ligand. Les calculs ont été réalisés pour un panel de 100 ligands choisi aléatoirement (soit 100 complexes protéine-ligand). Tous les calculs de scores ont été réalisés sur des géométries non minimisées.



(a)



(b)



(c)

cosecondes de la trajectoire d'équilibration (étape (d_{eq}), Figure 5.1, p. 136), nous avons sélectionné les 1 000 ligands DOCK6 et les 500 ligands de chacune des listes rDOCK (rDOCK-1, rDOCK-2 et rDOCK-3) de meilleurs scores (étape **(5)** de la Figure 5.1, p. 136).

Puis, dans la trajectoire de production de 200 ps obtenue à l'étape (c_{pr}) de la Figure 5.1, (p. 136), seuls un, cinq, neuf et sept ligands provenant des listes respectives DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3 ont présenté une instabilité et ont été éliminés à l'étape **(6)** de cette Figure. Enfin, pour chaque ligand restant, des calculs de scores avec la fonction de scoring HMScore (implémentée dans la méthode de calcul XSCORE) ont été réalisés sur 100 géométries non minimisées extraites de façon équidistante sur l'ensemble de la trajectoire de production. Le score moyen calculé sur les 100 géométries de chaque ligand a été utilisé pour obtenir un classement final. Les résultats issus de ce classement final seront analysés dans la section qui suit.

5.3 Analyse des résultats obtenus à la fin du criblage

Parmi les ligands restants à la fin du criblage, nous voulions savoir s'il nous était possible de réaliser une dernière étape de sélection qui consistait à choisir les ligands pour lesquels un consensus de docking est observé. Dans cette section, nous tenterons de répondre à cette question avant de présenter et d'analyser le top-20 des ligands issus du classement final.

Peut-on sélectionner les ligands selon un consensus de docking ?

Dans l'étape (b) de la Figure 5.1 (p. 136), nous avons utilisé les deux programmes de docking DOCK6 et rDOCK notamment afin de vérifier si une éventuelle sélection des ligands selon un consensus de docking était possible (c'est-à-dire un consensus de pose

et de rang calculé par la fonction de scoring primaire). Pour savoir si nous pouvions nous baser sur un consensus de docking pour sélectionner les ligands, nous avons recherché et comparé les poses et les rangs des ligands qui se retrouvaient dans plusieurs listes de résultats (DOCK6 et/ou rDOCK-1 et/ou rDOCK-2 et/ou rDOCK-3) après l'étape (b) de docking ; soit à l'étape **(1)** de la Figure 5.1 (p. 136) ainsi qu'à la fin du criblage. Si un consensus de docking est observé pour un ligand, la probabilité que ce ligand soit un bon candidat est plus élevée [117]. En effet, chaque programme de docking a son propre algorithme qui lui permet de rechercher la position du ligand sur une protéine et sa propre fonction de scoring (fonction de scoring primaire) qui lui permet d'évaluer les poses des ligands. C'est le cas des programmes de docking DOCK6 et rDOCK : la fonction de scoring primaire implémentée dans le programme rDOCK calcule des contributions énergétiques que la fonction de scoring primaire implémentée dans DOCK6 ne prend pas en compte (section 1.6, p. 41).

À l'étape **(1)** de la Figure 5.1 (p. 136), 12 ligands étaient classés à la fois dans le top-5 000 des ligands DOCK6 et dans le top-5 000 des ligands rDOCK. Pour chacun de ces 12 ligands, nous avons réalisé des calculs de RSMD entre la pose générée par DOCK6 (pose DOCK6) et chacune des trois poses générées par rDOCK (poses rDOCK-1, rDOCK-2, rDOCK-3). En d'autres mots, pour les ligands qui étaient à la fois dans les quatre listes (DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3), trois valeurs de RMSD ont été calculées et seront nommées RMSD-1, RMSD-2 et RMSD-3. La valeur RMSD-1 est calculée entre la pose DOCK6 et la pose rDOCK-1. La valeur RMSD-2 est calculée entre la pose DOCK6 et la pose rDOCK-2. Enfin, la valeur RMSD-3 est calculée entre la pose DOCK6 et la pose rDOCK-3. Bien qu'on retrouve chacun de ces 12 ligands dans au moins deux listes de ligands (DOCK6 et rDOCK-1 ou rDOCK-2 ou rDOCK-3), aucun consensus de pose n'a été constaté. En effet, pour chacun des 12 ligands, on constate que la pose générée par DOCK6 est différente de celle générée par rDOCK, qu'il s'agisse de la pose rDOCK-1, rDOCK-2 ou rDOCK-3 puisque les valeurs de RSMD-1, RMSD-2 et RMSD-3

Tableau 5.1: Moyenne des différences des rangs calculée avec la fonction de scoring primaire DOCK6 et avec la fonction de scoring primaire rDOCK d'une part, et différence entre les RMSD moyens d'autre part pour les 12 ligands qui sont à la fois dans les listes DOCK6 et rDOCK (rDOCK-1, rDOCK-2, rDOCK-3) à l'étape **(1)** du protocole de criblage (Figure 5.1, p. 136).

	DOCK6 / rDOCK-1	DOCK6 / rDOCK-2	DOCK6 / rDOCK-3
Moyenne des différences de rang	1 826 ± 1 005	1 530 ± 800	1 435 ± 1 200
Différence de RSMD moyen	9.94 ± 1.61 (RMSD-1)	10.93 ± 2.5 (RMSD-2)	10.54 ± 1.54 (RMSD-3)

sont grandes (toutes supérieures à 7 Å). Les valeurs moyennes de RSMD-1, RMSD-2 et RMSD-3 calculées à partir des valeurs individuelles des 12 ligands sont présentées dans le Tableau 5.1 : on constate qu'elles sont autour de 10 Å. Nous avons aussi, pour chacun des 12 ligands, comparé le rang calculé par la fonction de scoring primaire de DOCK6 (rang DOCK6) et celui calculé par la fonction de scoring primaire de rDOCK dans chacune des poses rDOCK-1, rDOCK-2 et rDOCK-3 (dénommée respectivement par rang rDOCK-1, rDOCK-2 et rDOCK-3). Dans le Tableau 5.1, on constate que la moyenne des différences de rangs calculée sur les 12 ligands varie beaucoup : (i) moyenne de 1826 entre les rangs DOCK6 et les rangs rDOCK-1 ; 1530 entre les rangs DOCK6 et les rangs rDOCK-2 ; 1435 entre les rangs DOCK6 et les rangs rDOCK-3. Par conséquent, avant l'étape de relaxation (soit les étapes (c_{eq}) et (c_{pr}) de la Figure 5.1, p. 136), aucun consensus de pose ou de score n'a été observé.

Parmi les 12 ligands cités ci-dessus, seuls six ligands (ligand 1 à ligand 6, Tableau 5.2) ont été retenus à la fin du criblage, c'est-à-dire après l'étape (d_{pr}) de la Figure 5.1 (p. 136). Cependant, parmi ces six ligands, seul le ligand 1 du Tableau 5.2 se trouve à la fois dans deux listes de ligands : les listes DOCK6 et rDOCK-2. Les cinq autres ligands se

Tableau 5.2: Rangs calculés avec les fonctions de scoring primaire DOCK6 et rDOCK (R1) d'une part et avec la fonction de scoring secondaire HMScore (R2) d'autre part pour les six ligands qui se trouvent à la fois dans la liste DOCK6 et/ou rDOCK-1 et/ou rDOCK-2 et/ou rDOCK-3 avant la relaxation des complexes (étape **(1)**) et à la fin du criblage (soit le classement obtenu après d_{pr} de la Figure 5.1, p. 136). R1 a été calculé sur la géométrie issue du docking de chaque ligand (avec la fonction de scoring primaire rDOCK ou DOCK6). R2 a été calculé à l'étape (d_{pr}) sur 100 géométries extraites de façon équidistante sur la trajectoire de production pour chaque ligand. La première colonne du Tableau contient les numéros des ligands.

	Avant la relaxation des complexes (étape (1) du protocole de criblage)				**À la fin du criblage (étape (d_{pr}) du protocole de criblage)**			
	DOCK6	rDOCK-1	rDOCK-2	rDOCK-3	DOCK6	rDOCK-1	rDOCK-2	rDOCK-3
1	R1 : 2601	R1 : 369	R1 : 293	R1 : 449	R2 : 731	_	R2 : 235	_
2	R1 : 1601	_	R1 : 4080	R1 : 3076	R2 : 877	_	_	_
3	R1 : 127	R1 : 4718	_	_	R2 : 44	_	_	_
4	R1 : 4598	R1 : 4793	_	_	_	R2 : 58	_	_
5	R1 : 3390	R1 : 2312	R1 : 1954	R1 : 2282	_	_	_	R2 : 89
6	R1 : 3931	R1 : 3394	R1 : 4195	R1 : 2627	_	_	_	R2 : 340

trouvent seulement dans l'une des listes. Dans le Tableau 5.2, les résultats suivant sont présentés : rang HMScore (c'est-à-dire le rang de chaque ligand calculé avec la fonction de scoring secondaire HMScore à l'étape (d_{pr}) de la Figure 5.1, p. 136) et les rangs DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3 comme définis plus haut pour chacun des six ligands cités ci-dessus. Ce tableau montre les résultats obtenus à l'étape **(1)** (c'est-à-dire avant la relaxation des complexes) et à la fin du criblage (c'est-à-dire après l'étape (d_{pr})) de la Figure 5.1 (p. 136).

À la fin du criblage, le ligand 1 est retrouvé en rang 731 sur 999 et 235 sur 491

respectivement dans les listes DOCK6 et rDOCK-2. De plus, ce même ligand adopte des poses très différentes dans chacune des listes (Figure 5.8). Il n'y a donc pas consensus de pose pour le ligand 1 et les scores associés à ces poses sont mauvais. Les ligands 3, 4 et 5 ont respectivement un rang de 44 sur 1000, 58 sur 500 et 89 sur 500 à la fin du criblage. Ces ligands sont dans le top-100, mais ils ne présentent pas de consensus de pose : la pose adoptée par le ligand 3 est très différente de celle du ligand 4 et du ligand 5 par exemple. Étant donné que les ligands ne sont pas bien classés ou que leurs poses ne sont pas reproduites pas les deux programmes de docking DOCK6 et rDOCK, on ne peut pas se baser sur un consensus de docking pour sélectionner les ligands dans notre cas. Après avoir fait un classement de l'ensemble des ligands (DOCK6, rDOCK-1, rDOCK-2 et rDOCK-3), les 20 ligands de meilleurs scores ont été analysés.

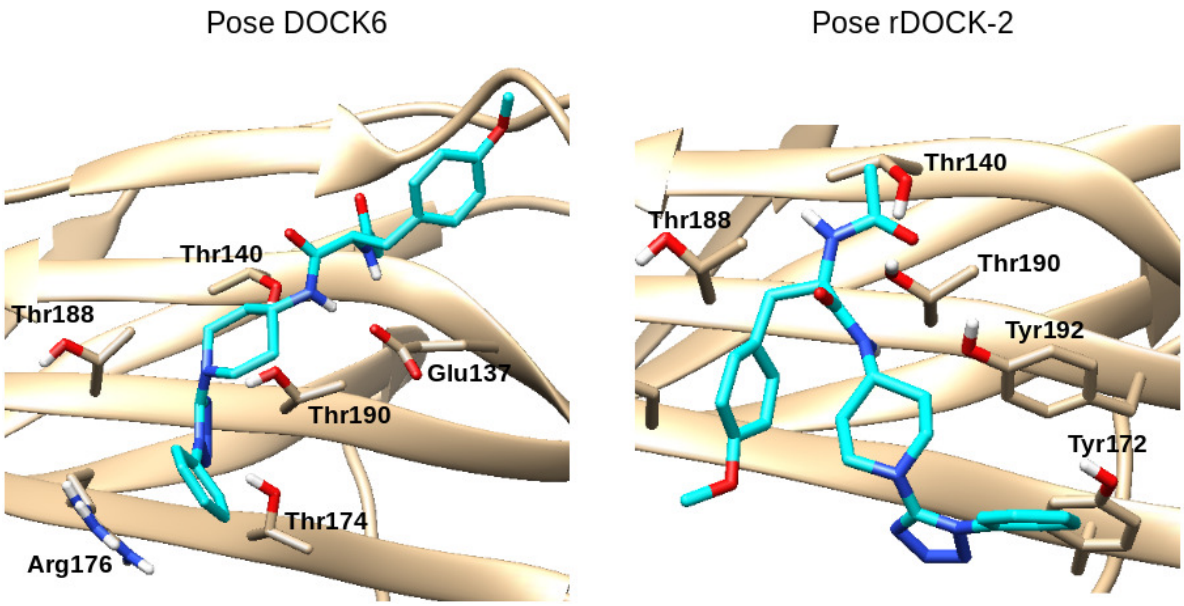


Figure 5.8: Snapshots du ligand 1 après criblage (c'est-à-dire à l'étape (d_{eq}) de la Figure 5.1, p. 136). Les poses générées par DOCK6 (pose DOCK6) et rDOCK (pose rDOCK-2) ainsi que les régions occupées sont très différentes pour le même ligand.

5.4 Analyse des 20 ligands de meilleurs scores

À l'étape (d_{pr}) de la Figure 5.1 (p. 136), des calculs de scores avec la fonction de scoring HMScore ont été appliqués aux complexes protéine-ligand restants pour obtenir un classement des ligands dans chacune des listes, soit une liste de 988 ligands DOCK6, une liste de 495 ligands rDOCK-1, une liste de 491 ligands rDOCK-2 et une liste de 493 ligands rDOCK-3. Puisqu'aucun consensus de pose ou de score n'a permis d'orienter le choix des ligands à analyser, nous avons décidé de fusionner les quatre listes de ligands afin d'avoir un classement final. De ce classement final, nous présenterons et analyserons le top-20.

Dans le top-20 issu du classement final, nous retrouvons les top-10, top-4, top-3 et top-3 respectivement des listes rDOCK-1, rDOCK-2, rDOCK-3 et DOCK6 obtenus à l'étape (d_{pr}) de la Figure 5.1 (p. 136). Les caractéristiques de ces ligands ainsi que leur structure 2D sont présentées dans le Tableau 5.3 et la Figure 5.9. Ce tableau sera commenté au fil des analyses des complexes.

La liaison de ces ligands à Lu est régie par les principes de thermodynamique : un ligand se lie à une protéine lorsque l'énergie libre de Gibbs du processus de liaison est négative. L'énergie libre de Gibbs est la somme des termes enthalpique et entropique. Les contributions enthalpiques peuvent être des liaisons hydrogène, ioniques ou halogènes, des interactions électrostatiques et de van der Waals, etc. L'entropie est une mesure de la dynamique de l'ensemble du système. Par exemple, les effets de solvatation tels que la réorganisation du solvant ou le déplacement de molécules d'eau étroitement liées à la protéine peuvent significativement contribuer au terme entropique de l'énergie libre d'association.

Ligand	Liste	MW (g/mol)	HBD	HBA	LogP	PSA (Å^2)	Nombre de liaison flexible	charge globale	Score
L1a	rDOCK-1	489.556	1	6	6.54	75	5	0	7.444
L2	rDOCK-1	494.422	0	5	6.08	47	7	0	7.274
L3	rDOCK-3	440.89	1	6	6.61	72	5	0	7.248
L1b	rDOCK-2	489.556	1	6	6.54	75	5	0	7.130
L5	DOCK6	522.069	4	7	3.49	92	12	+1	7.085
L6	rDOCK-3	449.554	0	5	6.71	62	7	0	7.025
L7	rDOCK-2	367.857	1	3	5.68	46	3	0	7.014
L8	DOCK6	501.559	2	8	5.57	116	10	-1	7.002
L9	rDOCK-1	414.553	2	5	6.73	62	5	0	6.977
L10	rDOCK-1	458.605	1	3	5.37	32	2	0	6.950
L11	rDOCK-1	437.33	0	5	4.53	43	3	0	6.930
L12	rDOCK-1	399.538	0	4	6.65	53	5	0	6.908
L13	rDOCK-1	337.423	0	5	3.77	66	5	0	6.888
L14	rDOCK-1	498.421	0	6	4.45	59	7	0	6.865
L15	rDOCK-3	429.495	0	3	6.71	31	4	0	6.864
L16	rDOCK-2	455.587	2	6	6.30	85	4	0	6.857
L17	rDOCK-2	524.645	0	8	6.75	99	13	0	6.829
L18	rDOCK-1	418.5	0	5	4.43	43	3	0	6.802
L19	DOCK6	491.038	2	7	5.44	93	8	0	6.779
L20	rDOCK-1	513.57	3	8	3.68	136	6	-1	6.750
LigA	DOCK6	474.3	2	8	1.83	112	6	-1	7.110
LigB	DOCK6	440.4	1	7	3.76	102	7	-1	6.884

Tableau 5.3: Caractéristiques physico-chimiques des ligands du top-20 du classement final obtenu avec HM-Score. Les scores de chaque ligand sont montrés dans la colonne Score. La deuxième colonne est la liste dans laquelle est issu chaque ligand. Le nombre de donneur de liaison hydrogène (HBD) et d'accepteur de liaison hydrogène (HBA) pour chaque ligand sont aussi montrés. Les deux dernières lignes du Tableau contiennent les scores et les caractéristiques physico-chimiques des ligands actifs ligA et ligB. Ces ligands n'ont pas été classés dans le top-20 dans ce Tableau.

L'entropie est une mesure de la dynamique de l'ensemble du système. Par exemple, les effets de solvatation tels que la réorganisation du solvant ou le déplacement de molécules d'eau étroitement liées à la protéine peuvent significativement contribuer au terme entropique de l'énergie libre d'association. Dans les analyses suivantes, les liaisons hydrogène et halogène ainsi que les interactions de van der Waals (contributions enthalpiques) seront prises en compte. Une liaison hydrogène est une liaison non covalente qui résulte de l'interaction attractive entre un hydrogène lié à un hétéroatome électronégatif (groupement donneur) et un autre hétéroatome électronégatif (groupement accepteur). Ces hétéroatomes peuvent être des atomes d'azote ou d'oxygène, mais aussi des halogènes tels que le fluor ou le chlore. Dans l'analyse de chacun des com-

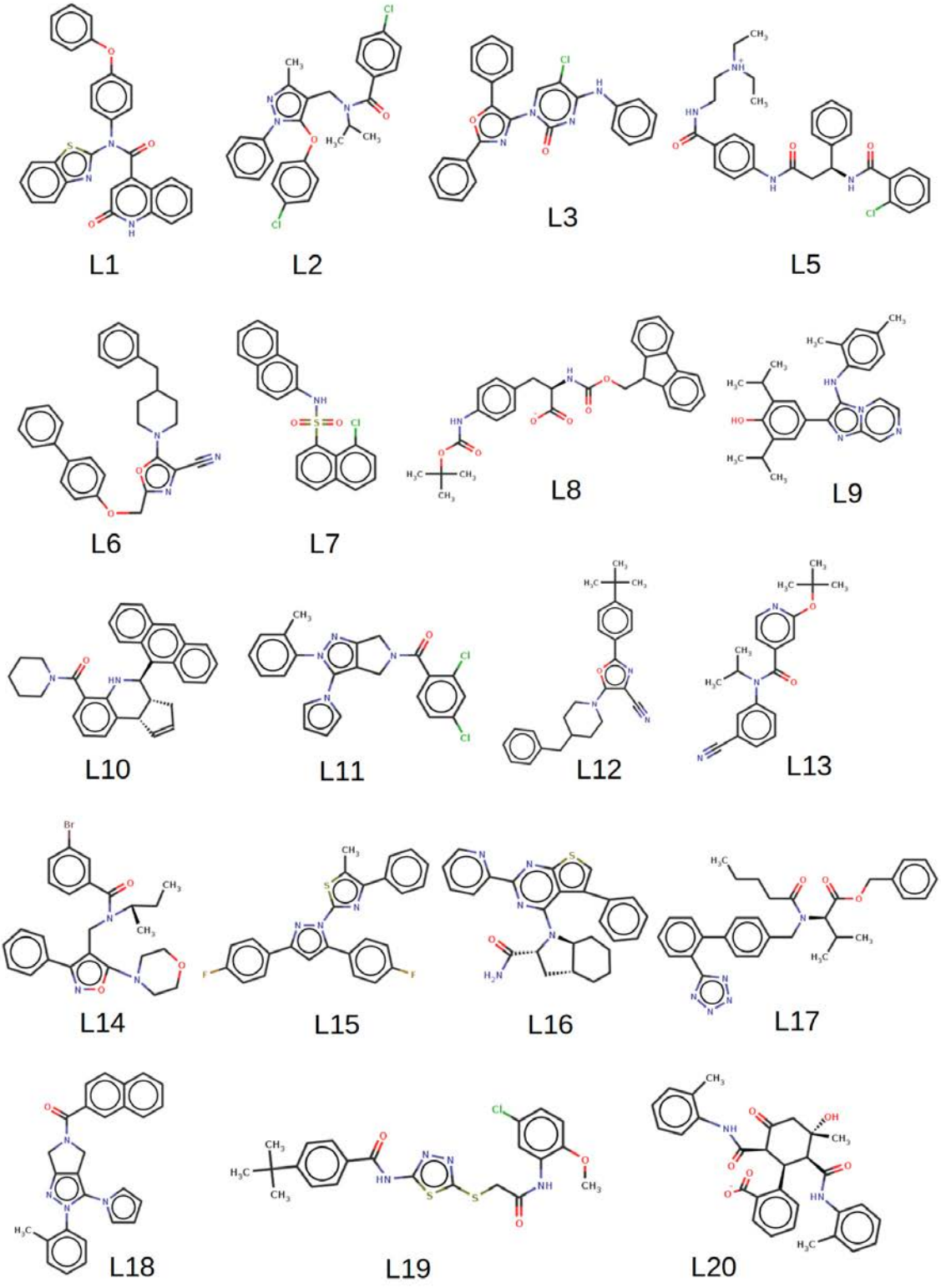


Figure 5.9: Structures 2D du top-20 du classement final. Le ligand L1 se trouve aux rangs 1 et 4 (dénommé L1a et L1b dans le Tableau 5.3). Seuls les ligands L5, L8 et L19 sont issus des listes DOCK6. Les autres ligands sont issus des listes rDOCK.

plexes formés entre Lu et les ligands du top-20, nous avons aussi étudié les réseaux de molécules d'eau cristallographique et leur remplacement par les ligands ainsi que le réseau de liaison hydrogène établi entre Lu et chacun des ligands.

5.4.1 Réseau de liaisons hydrogène

Liaisons hydrogène entre Lu et les ligands du top-20 du classement final

Comme expliqué à la page 134, les ligands actifs ligA et ligB, qui ont montré une inhibition de l'interaction Lu-Lnα5 selon les tests *in vitro*, se lient à Lu suivant un réseau de liaison hydrogène avec les résidus Thr174, Thr188, Thr190 ainsi que les résidus Tyr172 et/ou Tyr192 et/ou Arg176. Nous avons donc voulu, dans un premier temps, savoir si les ligands du top-20 du classement final (Figure 5.9, p. 157) se lient à Lu suivant un réseau de liaison hydrogène avec notamment les résidus cités ci-dessus. Pour cela, nous avons étudié (i) les liaisons hydrogène qui sont directement établies entre Lu et chacun des 20 ligands (dénommées liaison hydrogène directe dans la suite du manuscrit) et (ii) les liaisons hydrogène qui permettent aux ligands de se lier à Lu par l'intermédiaire de molécules d'eau (dénommées pont Lu-solvant-ligand dans la suite du manuscrit) dans la trajectoire de production obtenue à l'étape (c_{pr}) de la Figure 5.1 (p. 136). Une liaison hydrogène sera considérée comme pertinente lorsque (i) la distance entre deux hétéroatomes O et/ou N et/ou F et/ou Cl est inférieure ou égale à 3.4 Å et (ii) lorsque la liaison hydrogène est maintenue sur plus de 60 % de la trajectoire de production obtenue à l'étape (c_{pr}) de la Figure 5.1 (p. 136). Les liaisons hydrogène directes ainsi que les ponts Lu-solvant-ligand ont été calculés sur toutes les géométries de la trajectoire de production.

Lors de l'analyse de la trajectoire de production de six ligands du top-20 en complexe avec la protéine (L1a et L1b qui sont deux poses différentes du même ligand L1, L10, L13, L15, L18 et L19 de la Figure 5.9, p. 157), nous avons constaté que les groupements CO

et/ou NH du résidu Ala135 ou le groupement OH du résidu Tyr172 étaient enfouis : ni les ligands ni les molécules d'eau ne formaient de liaisons hydrogène avec ces groupements sur plus de 60 % des trajectoires. Ces conditions, qui correspondent à au moins deux donneur(s) ou accepteur(s) de liaison hydrogène enfouis, suggèrent que la liaison de ces ligands sur Lu serait défavorable. Les résultats pour ces six ligands ne sont donc pas présentés dans les analyses suivantes et dans les Tableaux 5.5, 5.4 et 5.6b.

Pour chacun des ligands restants qui, contrairement à ceux qui sont cités dans le paragraphe ci-dessus, se lient à Lu dans des conditions de liaison favorables (soit 13 ligands), nous avons listé les liaisons hydrogène directes établies entre Lu et le ligand dans (i) le complexe initial, c'est-à-dire le complexe issu du docking (étape **(1)** de la Figure 5.1, p. 136) et (ii) dans la trajectoire de production du complexe obtenue à l'étape (c_{pr}) de la Figure 5.1 (p. 136). Ces résultats sont respectivement montrés dans les Tableaux 5.4a et 5.5a (parmi les 13 ligands restants, seuls ceux qui forment bien ces liaisons hydrogène directes sont listés). À l'étape **(1)** (Figure 5.1, p. 136), 12 ligands forment entre une et trois liaisons hydrogène directes avec Lu, puis, dans la trajectoire de production, seuls 11 ligands forment entre une et trois liaisons hydrogène directes. En effet, le ligand L12 perd sa liaison hydrogène initialement formée avec la Tyr172 au cours de la simulation et aucune autre n'est formée. Ce ligand ne sera donc pas analysé par la suite. En revanche le ligand L2 qui ne formait pas de liaison hydrogène avec Lu dans le complexe initial, forme une liaison hydrogène qui est présente sur l'ensemble de la trajectoire de production (Tableau 5.5a). Bien qu'aucune liaison hydrogène directe n'est maintenue sur plus de 60 % de la trajectoire de production du complexe Lu-L6, nous avons constaté que le ligand L6 forme des liaisons hydrogène directes qui semblent jouer un rôle important dans sa stabilité (sur moins de 60 % de la trajectoire). Ce ligand sera donc retenu.

Dans le reste du manuscrit, seuls les ligands qui forment au moins une liaison hydrogène directe avec Lu et qui n'enfouissent pas les groupements CO et/ou NH du résidu

Ala135 ou le groupement OH du résidu Tyr172 ainsi que le ligand L6 ont été analysés, soit les 12 ligands : L2, L3, L5 à L9, L11, L14, L16, L17 et L20 (Figure 5.9, p. 157 et Figure 5.4.2 et p. 173). On les désignera dans la suite par les « 12 ligands restants ».

Dans le complexe initial généré à l'étape de docking pour chaque ligand, on constate que certaines liaisons hydrogène sont possibles suite à un léger réarrangement des chaînes latérales des résidus et/ou un léger déplacement du ligand permettant ainsi le rapprochement des groupements donneurs et/ou accepteurs respectifs. C'est ce qu'on observe dans les complexes initiaux formés avec les ligands L2, L8 et L20 pour lesquels des liaisons hydrogène sont possibles respectivement avec les résidus Tyr172, Thr190 et Thr174 (liaisons hydrogène non listées dans le Tableau 5.4a). Ces liaisons hydrogène, absentes dans le complexe initial de chacun de ces ligands, sont ensuite formées et maintenues sur plus de 60 % de la trajectoire de production des complexes correspondants (Tableau 5.5a).

Dans ce paragraphe, nous ne discuterons que des ligands qui sont présentés dans les Tableaux 5.4a et 5.5a. En comparant ces deux tableaux, on constate que les liaisons hydrogène directes présentes dans le complexe initial de chacun de ces ligands (Tableau 5.4) ne sont pas toujours maintenues sur plus de 60 % de la trajectoire de production de ce même complexe (Tableau 5.5a ; on parlera de complexe simulé). En effet, par exemple pour le ligand L9, on constate que parmi les liaisons hydrogène directes qui sont formées entre Lu et ce ligand dans le complexe initial (Tableau 5.4a), seule la liaison hydrogène établie avec le résidu Ala135 est maintenue sur plus de 60 % de la trajectoire de production (Tableau 5.5a). Dans les complexes formés par d'autres ligands, tel que pour le ligand L3, on constate qu'aucune liaison hydrogène directe formée dans le complexe initial n'a été maintenue dans la trajectoire de production, bien que d'autres liaisons hydrogène maintenues sur plus de 60 % de la trajectoire se soient formées (Tableaux 5.4 et 5.5). Ces résultats montrent que les liaisons hydrogène présentes dans les géométries issues du docking (complexe initial) ne peuvent être considérées comme représentatives

(a)

Ligand	Liaison
L3	TYR 172 OH <=> LIG O1 TYR 192 OH <=> LIG N1
L5	ALA 135 O <=> LIG N1 THR 174 OG1 <=> LIG O3
L6	THR 174 OG1 <=> LIG N3
L7	ALA 135 N <=> LIG O1
L8	GLU 137 OE2 <=> LIG N1 THR 174 OG1 <=> LIG O6
L9	ALA 135 O <=> LIG O1 TYR 192 OH <=> LIG N3
L11	ALA 135 N <=> LIG O1 TYR 172 OH <=> LIG N2
L12	TYR 172 OH <=> LIG O1
L14	SER 134 OG <=> LIG O3 GLU 137 N <=> LIG O1
L16	ALA 135 N <=> LIG N3 TYR 172 OH <=> LIG O1
L17	ASN 142 ND2 <=> LIG O1 THR 174 OG1 <=> LIG N3
L20	ARG 176 NH1 <=> LIG O2 THR 190 OG1 <=> LIG O6

(b)

Ligand	Liaison
ligA	THR 174 OG1 <=> LIG N1 THR 174 OG1 <=> LIG O6 THR 190 OG1 <=> LIG O6 TYR 192 OH <=> LIG O4
ligB	THR 174 OG1 <=> LIG O2 ARG 176 NH1 <=> LIG N1 ARG 176 NH1 <=> LIG N2 ARG 176 NH1 <=> LIG O2 THR 188 OG1 <=> LIG O4 THR 190 OG1 <=> LIG O1

Tableau 5.4: Liste des liaisons hydrogène directement formées entre Lu et les ligands dans les complexes initiaux (étape **(1)**, Figure 5.1, p. 136) (a) pour le top-20 du classement final, et (b) pour les ligands actifs ligA et ligB. Seuls les ligands qui n'enfouissent pas les groupements des résidus Ala135 et/ou Tyr172 et ceux qui forment des liaisons hydrogène directes avec Lu sont listés.

(a)

Ligand	Liaison	Fréquence (%)
L2	TYR 172 OH <=> LIG O2	100
L3	THR 174 OG1 <=> LIG O2	99.5
	TYR 192 OH <=> LIG O2	85
L5	ALA 135 O <=> LIG N1	99.5
	TYR 172 OH <=> LIG O2	95.5
	THR 174 OG1 <=> LIG O3	96
L7	ALA 135 N <=> LIG O1	99.5
Ł8	THR 190 OG1 <=> LIG O6	97.5
	THR 174 OG1 <=> LIG O6	100
	THR 140 OG1 <=> LIG N1	96
L9	ALA 135 O <=> LIG O1	98.5
L11	ALA 135 N <=> LIG O1	99
	TYR 172 OH <=> LIG O1	100
L14	TYR 172 OH <=> LIG O3	100
L16	TYR 192 OH <=> LIG N5	93.5
	TYR 172 OH <=> LIG O1	98.5
L17	THR 190 OG1 <=> LIG N4	95.5
	THR 174 OG1 <=> LIG N2	90
	THR 174 OG1 <=> LIG N3	97
L20	THR 190 OG1 <=> LIG N2	61.5
	ARG 176 NH1 <=> LIG O2	99
	THR 174 OG1 <=> LIG O3	98.5

(b)

Ligand	Liaison	Fréquence (%)
ligA	TYR 192 OH <=> LIG O4	99
	THR 190 OG1 <=> LIG O6	100
	THR 174 OG1 <=> LIG O6	100
ligB	THR 188 OG1 <=> LIG O4	88
	ARG 176 NH1 <=> LIG O2	100
	THR 174 OG1 <=> LIG O2	100

Tableau 5.5: Liste et fréquence des liaisons hydrogène directement formées entre Lu et (a) les ligands du top-20 et (b) les ligands actifs, pendant la trajectoire de production de chacun des ligands. Seuls les ligands qui n'enfouissent pas les groupements des résidus Ala135 et/ou Tyr172 et ceux qui forment des liaisons hydrogène directes avec Lu sont listés.

d'un complexe donné. En effet, certaines liaisons hydrogène du complexe initial peuvent être maintenues pendant les simulations, mais beaucoup d'entre elles ne sont que transitoires comme pour le cas du ligand L8 pour lequel deux liaisons hydrogène sont détectées dans le complexe initial parmi lesquelles seulement une est conservée sur l'ensemble de la trajectoire (liaison hydrogène entre l'atome OG1 de Thr174 et l'atome O6 du ligand ; Tableaux 5.4a et 5.5a). Nous pouvons souligner que tout comme pour le ligand ligA, les ligands L5, L8, L17 et L20 forment trois liaisons hydrogène avec Lu qui sont maintenues sur plus de 60 % de la trajectoire de production (Tableau 5.5, p. 162). La prise en compte de la flexibilité du complexe protéine-ligand par des simulations DM (aux étapes (c_{eq}) et (c_{pr}) de la Figure 5.1 (p. 136) permet donc d'obtenir les liaisons hydrogène les plus stables dans le temps et, vraisemblablement aussi, les plus représentatives du complexe. Ainsi, lors de l'étape suivante de calcul de scores sur ces complexes (étape (d_{pr}), Figure 5.1, p. 136), seules les liaisons hydrogène les plus fréquentes auront le plus de chance d'être comptabilisées par la fonction de scoring HMScore.

Pour les ligands actifs ligA et ligB, nous avons aussi constaté que les liaisons hydrogène directes formées dans leur complexe initial n'étaient pas toutes représentatives puisqu'elles ne sont pas toutes sont maintenues sur plus de 60 % de la trajectoire de production. En effet, alors que les ligands actifs ligA et ligB forment respectivement quatre et six liaisons hydrogène avec Lu dans leur complexe initial, seules trois liaisons hydrogène pour ces deux ligands ligA et ligB sont maintenues dans les complexes simulés (Tableaux 5.4b et 5.5b). Il est à préciser que les simulations DM obtenues pour les ligands actifs ligA et ligB ont été calculées dans les mêmes conditions que pour les 1 295 678 ligands criblés sur Lu (section 1.4, p. 31).

Enfin, l'analyse du tableau 5.5a montre que le nombre de liaisons hydrogène à lui seul n'est pas corrélé au classement des ligands. En effet, pour les ligands L2 et L3 (respectivement deuxième et troisième du top-20 ; Tableau 5.3), au plus deux liaisons hydrogène sont détectées par ligand alors que pour les ligands L17 et L20 (respective-

ment 17^{e} et 20^{e}), trois liaisons hydrogène sont comptées dans les complexes formés par chacun de ces ligands. Le nombre de liaisons hydrogène fréquentes à plus de 60 % sur la trajectoire ne semble donc pas, à lui seul, déterminer le classement des ligands.

Liaison des ligands sur Lu par l'intermédiaire de molécules d'eau

Dans la sous-section ci-dessus, nous avons présenté les liaisons hydrogène directement formées entre Lu et les ligands du top-20 du classement final (Figure 5.9, p. 157). Ces liaisons hydrogène participent directement à la liaison et à la stabilité des ligands sur Lu. En plus de ces liaisons hydrogène, la stabilité des ligands peut aussi être assurée par la présence de molécules d'eau qui favoriseraient la liaison des ligands sur la protéine. En effet, en plus de permettre la solvatation des complexes, certaines molécules d'eau peuvent être importantes dans la liaison du ligand, dans la stabilité du complexe et/ou dans la reconnaissance du site de liaison à travers un réseau de liaison hydrogène et d'interaction de van der Waals. Bien que dans notre système, nous n'ayons pas d'information expérimentale sur des molécules d'eau indispensables à la liaison des ligands, nous avons voulu savoir si la stabilité de ces ligands sur Lu pouvait aussi être due, en partie, à une ou plusieurs molécules d'eau. Pour cela, dans chacun des complexes formés entre Lu et les ligands du top-20 du classement final, nous avons analysé les ponts Lu-solvant-ligand qui sont maintenus par des liaisons hydrogène entre Lu et une molécule d'eau, d'une part, et entre cette molécule d'eau et le ligand d'autre part.

Les Tableaux 5.6a et 5.6b listent les ponts Lu-solvant-ligand respectivement formés avant et pendant la trajectoire de production des simulations DM (c'est-à-dire les étapes respectives **(1)** et (c_{pr}) de la Figure 5.1, p. 136). Pour les calculs de ponts Lu-solvant-ligand dans les complexes initiaux, seuls les ligands rDOCK ont pu être considérés puisque ce sont les seuls qui ont été complexés à Lu en présence d'eau explicite. En effet, le programme DOCK6 ne permettait pas de prendre en compte des molécules d'eau explicite pendant le docking. Les distances entre (i) Lu et ligand (ii) solvant et Lu et (iii) solvant et

ligand sont aussi présentés dans ces tableaux. Avant les simulations (étape **(1)** de la Figure 5.1, p. 136), des ponts Lu-solvant-ligand sont observés seulement dans les complexes formés entre Lu et quatre ligands sur 20 : L1b, L12, L13 et L18. Les ligands L1b, L13 et L18 font partie des ligands qui enfouissent les groupements CO et/ou NH du résidu Ala135 et/ou le groupement OH du résidu Tyr172 (Tableau 5.6a) et n'ont donc pas été présentés dans la suite de l'étude. Le ligand L12 qui ne forme aucune liaison hydrogène directe avec Lu n'a pas non plus été présenté dans la suite du manuscrit. Aucun des 12 ligands restants (ligands définis à la page 160) ne participe à un ou plusieurs ponts Lu-solvant-ligand à l'étape **(1)** (Figure 5.1, p. 136). En revanche, lors de la relaxation des complexes formés entre Lu et chacun des ligands L6, L8, L9, L14, L16, L17 et L20, on constate que des molécules d'eau favorisent et renforcent la liaison de ces ligands sur Lu en participant à des ponts Lu-solvant-ligand qui sont maintenus sur plus de 60 % de la trajectoire de production. Le plus souvent, les ponts Lu-solvant-ligand sont assurés par au moins deux molécules d'eau qui s'échangent tour à tour au cours de la simulation (on parlera de molécules d'eau intermédiaires). C'est par exemple le cas du ligand L9 qui participe à un pont Lu-solvant-ligand dans lequel quatre molécules d'eau s'échangent tour à tour au cours de la simulation pour assurer le maintien du pont sur 65 % de la trajectoire de production. Pour les ligands actifs ligA et ligB, on constate qu'un seul pont Lu-solvant-ligand renforce la liaison de ces ligands sur Lu avec respectivement une et quatre molécules d'eau intermédiaires. Les valeurs de distance $d_{prot-lig}$ calculées dans tous les cas listés du Tableau 5.6b indiquent qu'il n'existe pas de liaison hydrogène entre les atomes de Lu et ceux des ligands mentionnés dans la dernière colonne de ce Tableau.

(a)

Ligand	Liaison	$d_{prot-lig}$ (Å)	$d_{solv-lig}$ (Å)	$d_{solv-prot}$ (Å)
L1b	TYR 172 OH <=> LIG O2	2.56	2.85	2.93
	GLU 137 N <=> LIG O3	3.69	2.62	2.74
L12	TYR 192 OH <=> LIG N1	3.68	2.98	2.71
L13	ASN 124 ND2 <=> LIG N3	5.83	3.10	2.85
	THR 127 OG1 <=> LIG O2	4.52	2.83	2.86
L18	TYR 192 OH <=> LIG O1	4.55	2.93	3.07

(b)

Ligand	Liaison	Fréquence (%)	$d_{prot-lig}$ (Å)	$d_{solv-lig}$ (Å)	$d_{solv-prot}$ (Å)	Nombre de molécule d'eau intermédiaire
L6	ARG 176 NH1 <=> LIG O1	63.5	5.58 ± 0.30	3.12 ± 0.17	2.87 ± 0.14	1
L8	GLU 137 OE2 <=> LIG O2	89	3.79 ± 0.36	2.77 ± 0.18	2.78 ± 0.23	3
	SER 175 O <=> LIG O3	97	3.82 ± 0.30	2.64 ± 0.11	2.79 ± 0.14	1
	TYR 192 OH <=> LIG O4	93.5	4.24 ± 0.31	2.79 ± 0.17	2.92 ± 0.17	6
L9	SER 134 OG <=> LIG O1	65	3.51 ± 0.30	2.94 ± 0.18	2.92 ± 0.18	4
L14	TYR 192 OH <=> LIG N2	61	4.52 ± 0.35	3.07 ± 0.16	3.07 ± 0.18	5
L16	SER 134 OG <=> LIG N2	87	4.92 ± 0.24	3.09 ± 0.15	2.87 ± 0.16	2
L17	ARG 144 NE <=> LIG O2	97	5.40 ± 0.22	2.77 ± 0.18	2.86 ± 0.14	3
	ARG 144 NH2 <=> LIG O2	100	5.18 ± 0.24	2.81 ± 0.20	2.94 ± 0.19	5
L20	TYR 192 OH <=> LIG O3	77.5	5.11 ± 0.32	3.02 ± 0.17	2.91 ± 0.17	2
ligA	THR 174 OG1 <=> LIG O5	69.5	3.89 ± 0.20	2.77 ± 0.15	2.92 ± 0.18	1
ligB	SER 175 O <=> LIG O3	69.5	4.07 ± 0.34	3.03 ± 0.20	2.89 ± 0.20	4

Tableau 5.6: Liste et fréquences des ponts Lu-solvant-eau formés entre Lu et les ligands (a) dans les complexes initiaux, et (b) pendant les simulations DM. Seuls les ligands qui n'enfouissent pas les groupements des résidus Ala135 et/ou Tyr172 sont présentés. Pour chaque liaison, le nombre de molécules d'eau intermédiaire qui s'échangent tour à tour afin d'assurer le pont Lu-solvant-ligand pendant les simulations DM est donné. La distance est calculée entre les hétéroatomes impliqués.

Parmi les ligands restants, soit les 12 ligands qui n'enfouissent pas les groupements CO et/ou NH du résidu Ala135 et/ou le groupement OH du résidu Tyr172 et qui forment au moins une liaison hydrogène directe avec Lu (L2, L3, L5, L6 à L9, L11, L14, L16, L17 et L20, Figure 5.9, p. 157), on constate que sept de ces ligands (L6, L8, L9, L14, L16, L17 et L20, Figure 5.9, p. 157 et Tableau 5.6, p. 166) participent, en plus, à au moins un pont Lu-solvant-ligand. Il est possible que ces ponts Lu-solvant-ligand soient aussi importants que les liaisons hydrogène directes entre Lu et ces ligands pour assurer leur stabilité au cours de la simulation DM. Les ponts Lu-solvant-ligand ainsi que les autres interactions observées dans les complexes formés entre Lu et les 12 ligands restants sont présentés en détail dans la sous-section qui suit.

5.4.2 Analyse des complexes formés entre Lu et chacun des 12 ligands restants

Pour chacun des 12 ligands restants, le Tableau 5.7 montre le nombre de liaisons hydrogène directes établies entre Lu et le ligand ainsi que les résidus qui y sont impliqués, le nombre de ponts Lu-solvant-ligand observé dans les complexes, les résidus impliqués dans les interactions de van der Waals et les différentes observations particulières qui ont été faites dans les complexes. Un snapshot du complexe formé entre Lu et chacun des 12 ligands est montré dans la Figure 5.4.2. Ces snaphots sont issus de la trajectoire de production (obtenue à l'étape (c_{pr}) de la Figure 5.1, p. 136) de chacun des ligands. Dans cette section, les interactions Lu-ligand qui seront présentées et discutées sont celles qui sont observées dans la trajectoire de production de chaque ligand. Chacun des 12 ligands restants forme au moins une liaison hydrogène avec Lu et fait plusieurs contacts de van derWaals avec la protéine. En plus de ces interactions, les ligands L6, L8, L9, L14, L16, L17 et L20 participent à un ou plusieurs ponts Lu-solvant-ligand, ce qui renforce la stabilité de ces ligands au cours des simulations DM. Dans la suite de cette sous-section,

nous présenterons les complexes formés entre Lu et les ligands L2, L3, L5, L6, L7, L8, L11 et L20 pour lesquels nous notons quelques particularités.

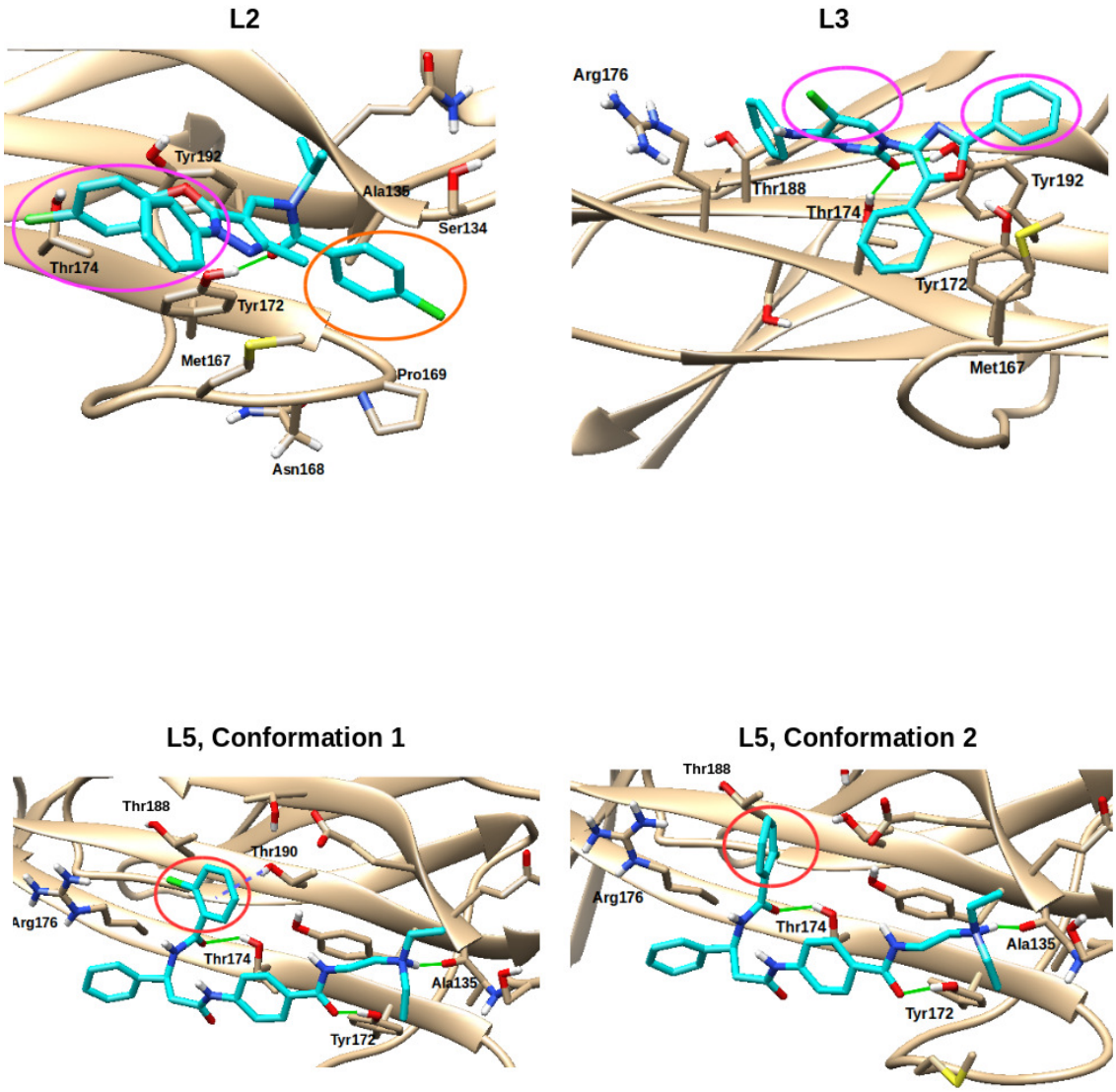
Le ligand L8 est un ligand intéressant puisqu'il se lie à Lu suivant un réseau de liaison hydrogène semblable à celui observé pour le ligand ligA. En effet, comme le ligand actif ligA, le ligand L8 se lie aux résidus Thr174 et Thr190 grâce à deux liaisons hydrogène directes et avec le résidu Tyr192 grâce à un pont Lu-solvant-ligand. Une autre particularité de ce ligand est qu'il remplace parfaitement la molécule d'eau HOH244 qui forme, dans le cristal de Lu, des liaisons hydrogène avec les atomes OG1 de Thr140 et OE2 de Glu137 (Figure 5.4, p. 142). Le ligand L8 remplace cette molécule d'eau grâce à la combinaison : (i) d'un pont Glu137-solvant-L8 et (ii) d'une liaison hydrogène directe entre Thr140 et L8.

Dans la trajectoire de production des complexes formés entre Lu et chacun des ligands L2, L3, L6, L7, L11 et L20, nous avons constaté que ces ligands font protubérance dans le solvant. Une telle protubérance est aussi observée dans le complexe formé par le ligand actif ligB et Lu. Un autre critère qui pourrait être utilisé pour la sélection des molécules à tester est la protubérance créée par le composé à la surface de Lu et qui peut gêner par encombrement stérique la liaison de la Laminine ; ces molécules devraient avoir au moins quelques atomes qui font saillie à la surface de Lu.

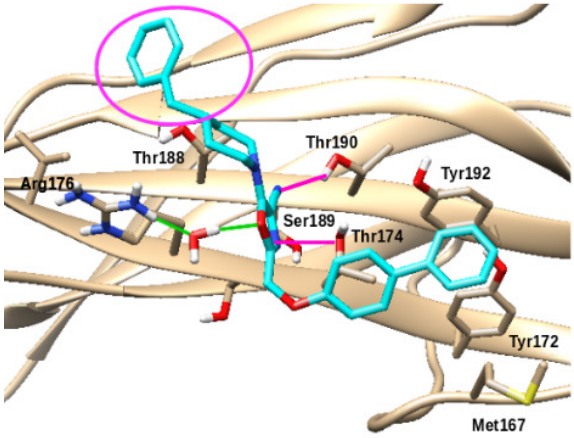
Tableau 5.7: Interactions observées dans les complexes formés entre Lu et chacun des 12 ligands restants. Les résidus qui sont impliqués dans les liaisons hydrogène directes et dans les ponts Lu-solvant-ligand sont entre parenthèses. Par exemple, le ligand L2 forme une liaison hydrogène avec le résidu Tyr172. Cette information est indiquée comme ceci dans le tableau : 1 (Tyr172). Les liaisons hydrogène particulières (liaisons hydrogène de type XH$\cdots\pi$) établies entre Lu et les ligands L5 et L9 sont en gras. Le ligand L5 adopte deux conformations au cours de la trajectoire de production qui sont les conformations Conf. 1 et Conf. 2. Dans la conformation Conf. 1, une liaison hydrogène est observée entre le cycle du chlorophényl du ligand et le résidu Thr190. Dans la conformation Conf. 2, une liaison halogène et des interactions multipolaires sont observées. La colonne « Autres » contient les différentes observations particulières réalisées pour les ligands (protubérance, liaison halogène, etc.).

	Nombre de liaisons hydrogène Lu-ligand	Nombre de ponts Lu-solvant-ligand	Contacts de van der Waals	Autres
L2	1 (Tyr172)	_	Ser134, Ala135, Met167, Asn168, Pro169	protubérance
L3	2 (Thr174, Tyr192)	_	Met167, Tyr172, Arg176, Thr188	protubérance
L5	· 3 (Ala135, Tyr172, Thr174). · **1 (entre chlorophényl du ligand et Thr190 dans la Conf. 1).**	_	Arg176, Thr188	Conf. 2 : liaisons halogène + interactions multipolaires avec Thr174 et Thr188
L6	2 (Thr174 et Thr190) importantes mais peu fréquentes	1 (Arg176)	Met167, Thr188, Ser175, Ser189, Tyr192	protubérance
L7	1 (Ala135)	_	Ser134, Ala135, Met167, Asn168, Pro169	· interactions π-π · protubérance
L8	3 (Thr140, Thr174, Thr190)	3 (Glu137, Ser175, Tyr192)	Val165, Met167, Tyr172	
L9	· 1 (Ala135) · **1 (entre diméthylphényl du ligand et Tyr172)**	1 (Ser134)	Gln136, Glu137, Met167, Thr174	
L11	2 (Ala135, Tyr172)	_	Ser134, Ala135, Met167, Asn168, Pro169	· interactions π-π · protubérance
L14	1 (Tyr172)	1 (Tyr192)	Ser134, Glu137, Met167, Asn168	
L16	2 (Tyr172, Tyr192)	1 (Ser134)	Ala135, Met167, Asn168	
L17	3 (Thr174, Thr190)	2 (Arg144)	Arg176, Val178, Thr188, Ser189	
L20	3 (Thr190, Arg176, Thr174)	1 (Tyr192)	Glu137, Thr140, Thr188	protubérance
ligA	3 (Thr174, Thr190, Tyr192)	1 (Thr174)	Arg176, Met167, Ala135	
ligB	3 (Thr174, Arg176, Thr188)	1 (Ser175)	Met167, Tyr172, Tyr192	

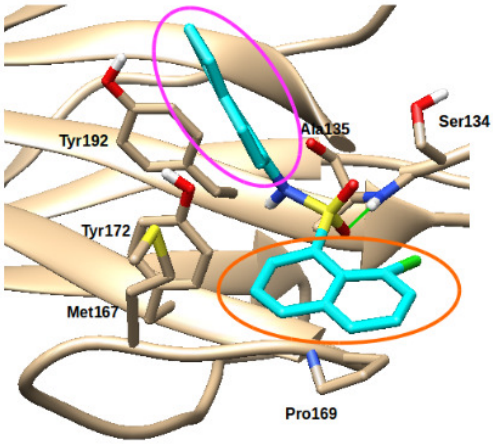
Figure 5.10: Snapshots des 12 ligands restants. Les groupements qui font protubérance dans le solvant sont entourés en magenta pour les ligands L2, L3, L6, L7 et L11 et ceux qui font contacts avec la protéine sont entourés en orange. Le ligand L5 possède deux conformations: Conformation 1 et Conformation 2. Ces conformations diffèrent par la position du chlorophényl entouré en rouge. Les liaisons hydrogène sont représentées par des traits verts. La liaison hydrogène particulière ainsi que les liaisons hydrogène peu fréquentes établies entre le ligand L9 et la Tyr172 sont respectivement montrées en trait pointillé bleu et en trait plein magenta.



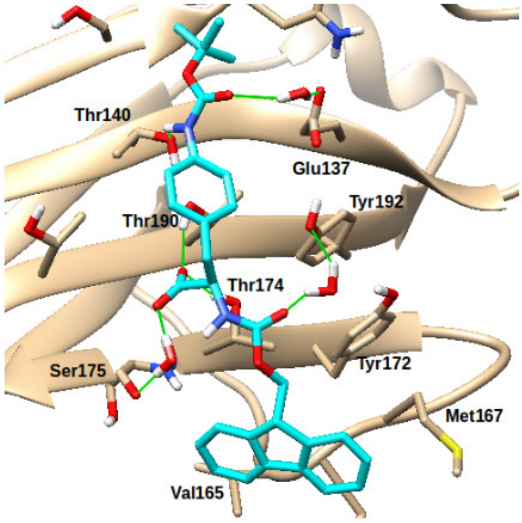
L6


Thr188
Thr190
Tyr192
Arg176
Ser189
Thr174
Tyr172
Met167

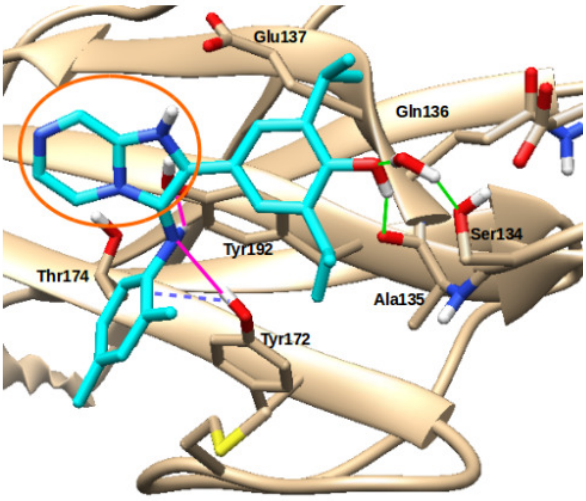
L7


Ala135
Ser134
Tyr192
Tyr172
Met167
Pro169

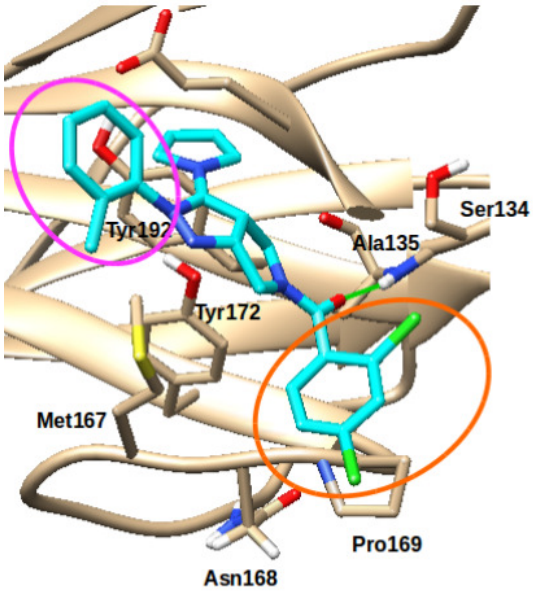
L8


Thr140
Glu137
Tyr192
Thr190
Thr174
Ser175
Tyr172
Met167
Val165

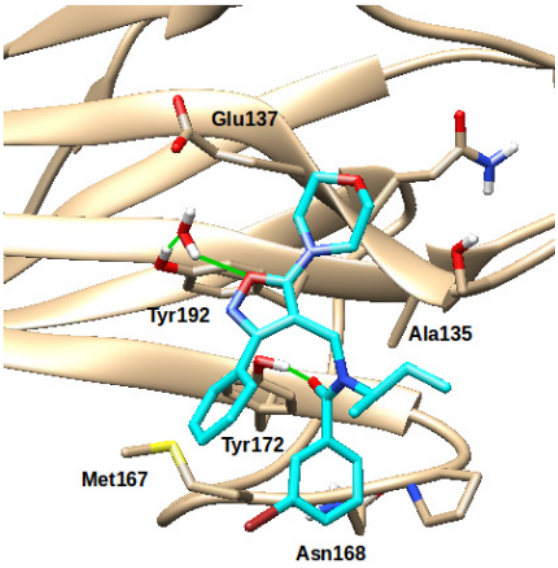
L9


Glu137
Gln136
Ser134
Tyr192
Thr174
Ala135
Tyr172

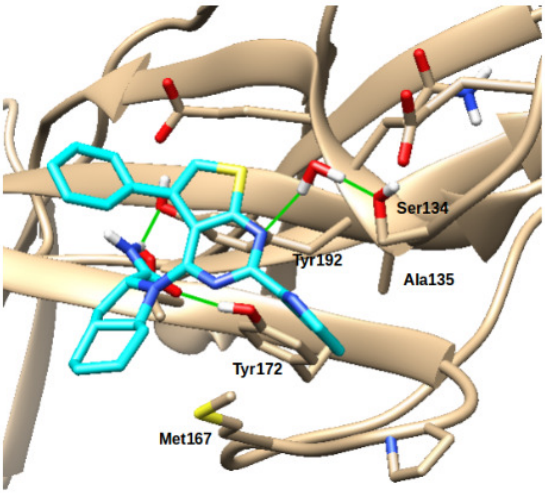
L11



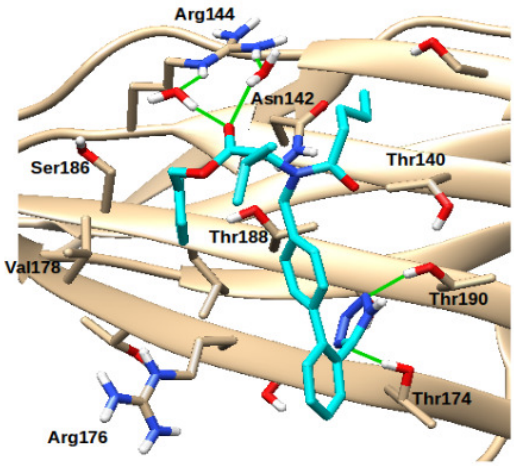
L14

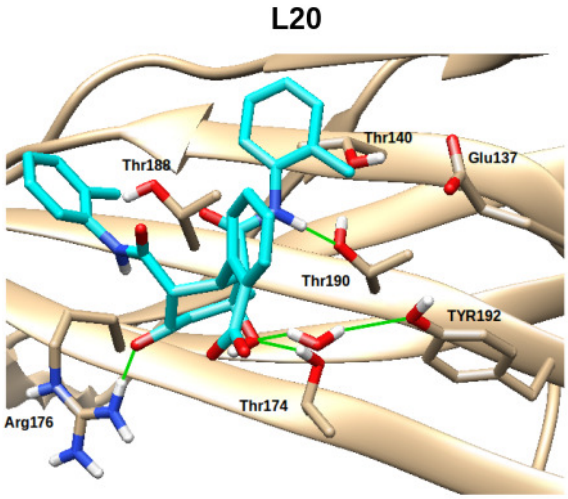


L16



L17





Au cours de la trajectoire de production, nous avons constaté que le ligand L5 adoptait deux conformations : la conformation 1 (L5, conformation 1 de la Figure 5.4.2) qui est observée sur les 86 premières picosecondes de la trajectoire de production puis la conformation 2 (L5, conformation 2 de la Figure 5.4.2) qui est observée sur le reste de la trajectoire de production. Dans la conformation 1 de L5, (Figure 5.4.2), le chlorophényl fait des contacts de van der Waals avec les résidus Arg176 et Thr188. On peut aussi observer une liaison hydrogène de type XH$\cdots\pi$ (liaison hydrogène particulière) entre l'atome OG1 du résidu Thr190 et le centre du cycle chlorophényl du ligand [118]. Dans la conformation 2 de L5, le chlore du chlorophényl du ligand pointe vers l'intérieur de la protéine et forme une liaison halogène avec les oxygènes des groupements C=O des résidus Thr188 et Thr174 et des interactions multipolaires avec les carbones de ces mêmes groupements. Cette orientation orthogonale du chlore par rapport au groupement carbonyle est le plus souvent visualisée dans les interactions multipolaires réalisées par le fluor [113, 119]. Cependant, les calculs de distances entre le chlore et les groupements carbonyles des résidus Thr174 et Thr188 correspondent à des contacts de van der Waals. Il est possible que le champ de force utilisé ne reproduise pas ces types d'interactions. Enfin, la conformation 2 favorise les interactions $\pi - \pi$ avec la chaîne

latérale de Arg176.

Le ligand L9 forme aussi une liaison hydrogène XH$\cdots\pi$ avec Lu au cours de la trajectoire de production. Cette liaison hydrogène est formée entre le groupement OH du résidu Tyr172 et le cycle aromatique du groupement diméthylphényl du ligand (liaison hydrogène représentée en trait pointillé bleu sur le snapshot nommé L9 de la Figure 5.4.2, p. 173). En effet, les groupements phényl jouent le rôle d'accepteur de liaison hydrogène lorsqu'un groupement donneur (le cas du OH du résidu Tyr172 ici) pointe vers le centre ou à proximité du centre de celui-ci [118]. Comme indiqué dans le Tableau 5.7 (p. 169), le ligand L9 forme aussi une liaison hydrogène directe avec le CO du squelette peptidique du résidu Ala135 ainsi qu'un pont Lu-solvant-ligand avec le résidu Ser134. Bien que ce soient les seules liaisons hydrogène présentes sur plus de 60 % de la trajectoire, il est à noter que deux autres liaisons hydrogène maintenues sur 46 % et 59 % de la production (absents du Tableau 5.7) pourraient également contribuer à la stabilité de ce ligand par ailleurs très hydrophobe : les liaisons hydrogène respectives formées avec le résidu Tyr192 et le résidu Tyr172 (liaisons hydrogène représentées en traits roses dans le snapshot nommé L9 dans la Figure 5.4.2, p. 173).

Bien qu'ayant des structures très différentes, les ligands L2, L7 et L11 arborent des poses similaires dans lesquelles une partie des ligands (entouré en magenta dans les snapshots correspondants de la Figure 5.4.2, p. 173) fait protubérance dans le solvant alors que l'autre partie (entouré en orange dans les snapshots correspondants de la Figure 5.4.2, p. 173) se trouve dans un environnement hydrophobe constitué des résidus Ser134, Ala135, Met167, Asn168 et Pro169 avec lesquels les ligands font des contacts de van der Waals sur plus de 80 % de la trajectoire. Pour chacun de ces ligands, la partie qui fait contact avec la région hydrophobe contient une structure cyclique (simple ou double) liée à un ou deux atomes de chlore. Pour les ligands L7 et L11, des interactions $\pi - \pi$ sont observées avec la liaison peptidique formée entre les résidus Asn168 et Pro169. Ces interactions viennent s'ajouter aux liaisons hydrogène formées entre le

N de Ala135 et ces ligands. Ce dernier phénomène ne concerne pas le ligand L2 pour lequel le chlorophényl ne fait que de simples contacts de van der Waals avec Pro169. Les interactions de van der Waals, les interactions $\pi - \pi$ et les liaisons hydrogène semblent permettre, ensemble, de renforcer la stabilité des ligands L7 et L11 au cours de la trajectoire de production malgré leur partie flexible qui fait protubérance (entouré en rose, dans les snapshots correspondants de la Figure 5.4.2, p. 173).

Enfin, comme indiqué dans le Tableau 5.7 (p. 169), le ligand L20 fait trois liaisons hydrogène directes avec les résidus Thr174, Arg176 et Thr190 et un pont Lu-solvant-ligand avec le résidu Tyr192 par l'intermédiaire d'une molécule d'eau. Bien que ce ligand possède une charge globale de -1 (situé sur le groupement COO^-, Figure 5.9, p. 157), aucune liaison hydrogène ionique n'est formée avec la protéine: le groupement COO^-, lié à un groupement phényle, fait protubérance dans le solvant ne contribuant donc pas à pénaliser la liaison de ce ligand sur Lu du fait du phénomène de désolvatation très défavorable énergétiquement dans le cas de ce type de groupement chargé. Néanmoins, ce groupement chargé peut faire des interactions électrostatiques avec le résidu Arg176 tous deux distants de moins de 7 Å. Cette interaction électrostatique pourrait favoriser la liaison hydrogène qui est formée entre le ligand et le résidu Arg176.

Dans l'analyse précédente, nous avons pu observer que les 12 ligands restants se lient à Lu grâce à des liaisons hydrogène directes établies avec les résidus et/ou grâce à des ponts Lu-solvant-ligand ainsi que par des contacts de van der Waals, mais aussi grâce à des interactions électrostatiques (ce dernier type d'interaction de concerne que le ligand L8, Figure 5.4.2, p. 173). Dans les complexes analysés, l'eau joue souvent un rôle d'intermédiaire pour la liaison des ligands et favorise donc le maintien des ligands sur Lu pendant les simulations. Cependant, ces interactions ne sont pas prises en compte dans les calculs d'énergie de liaison avec la méthode de calcul XSCORE puisque ce dernier ne considère que les liaisons hydrogène directes entre la protéine et le ligand. Par conséquent, les énergies des complexes formés par les ligands L6, L8, L9, L14, L16, L17 et

L20 qui se lient à Lu par l'intermédiaire d'au moins une molécule d'eau sont vraisemblablement sous-évaluées par XSCORE.

5.4.3 Différences entre les résultats DOCK6 et les résultats rDOCK

Les ligands ligA et ligB ont été obtenus en utilisant le programme DOCK6 à l'étape de docking. Pour autant, dans le top-20 du classement final, les ligands DOCK6 ne sont qu'au nombre de trois : les ligands L1b, L16 et L17 qui sont respectivement classés aux rangs 4, 16 et 17 (Tableau 5.3, p. 156). Afin de comprendre à quoi est due cette faible proportion de ligands DOCK6 dans le top-20 du classement final, nous avons analysé les ligands de la liste DOCK6 et ceux de la liste rDOCK.

En analysant le top-50 et le top-100 du classement final (obtenu sur toutes les listes confondues), nous y avons trouvé respectivement les 12 et les 20 meilleurs ligands de la liste DOCK6. Ces ligands DOCK6 sont plus flexibles (en moyenne 8 liaisons flexibles) que les ligands des listes rDOCK (en moyenne 5 liaisons flexibles). De plus, parmi les ligands DOCK6 retenus, les ligands L5 et L8 possèdent chacun un groupement chargé (soit les groupements respectifs NH^+ et COO^-) avec lequel ils forment une liaison hydrogène ionique avec les résidus respectifs Ala135 (groupement du squelette CO) et Thr174 et Thr190 (chaînes latérales). De telles liaisons hydrogène ioniques sont aussi observés dans les complexes formés par les ligands actifs ligA et ligB. Dans DOCK6, l'absence de prise en compte des phénomènes de solvatation (sauf au travers de la permittivité relative qui permet d'atténuer les interactions électrostatiques, Tableau 1.1, p. 49) semble donc assez logiquement favoriser les ligands qui peuvent former des interactions ioniques. Ce n'est pas le cas de rDOCK qui prend mieux en compte la solvatation au travers de calculs faits à partir de la surface accessible au solvant (Tableau 1.1, p. 49 ; *vide infra*).

Les ligands rDOCK prédominent en nombre dans le top-20 du classement final. Ces ligands sont moins flexibles que les ligands DOCK6 et ne présentent aucune charge à

l'exception du ligand L20. De plus, cinq, trois et un ligands rDOCK forment respectivement une, deux et trois liaisons hydrogène directes avec Lu. Le nombre de liaisons hydrogène direct formé entre les ligands rDOCK et Lu est donc plus faible par rapport à ce qu'on observe entre les ligands DOCK6 (L5, L8 ainsi que les ligands actifs ligA et ligB) et Lu qui forment trois liaisons hydrogène avec la protéine.

Étant donné que le nombre de liaisons hydrogène formé avec Lu ne semble pas expliquer les différences de proportion entre les ligands DOCK6 et les ligands rDOCK dans le top-20 du classement final, nous avons ensuite analysé les régions d'interaction occupées par chaque type de ligand. La région occupée par les poses générées par DOCK6 est similaire à celle occupée par le programme rDOCK (Figure 5.11). Le programme de docking rDOCK a tendance à positionner les ligands un peu plus vers le résidu Val178 alors que la région d'interaction occupée par les ligands DOCK6 (L5, L8, ligA et ligB) est limitée par les résidus Arg176 et Thr188. Parmi les ligands rDOCK-1 qui ont été retenus à la fin du criblage (L7, L16 et L17), on constate que les ligands L7 et L16 occupent une région d'interaction limitée par la zone en magenta alors que le ligand L17 occupe une région bien différente (zone en cyan). Cette différence est due à la taille des ligands et à leur structure : comparé aux ligands L7 et L16, le ligand L17 est de plus grande taille et présente une plus grande flexibilité (13 liaisons flexibles pour le ligand L17 contre trois et quatre liaisons flexibles respectivement pour les ligands L7 et L16 ; (Tableau 5.3, p. 156 et Figure 5.9, p. 157). Le ligand L17 peut donc être « étendu » afin de faire des contacts sur une plus grande région de la protéine. Les ligands issus des listes DOCK6 et rDOCK font donc en général des interactions avec les mêmes résidus.

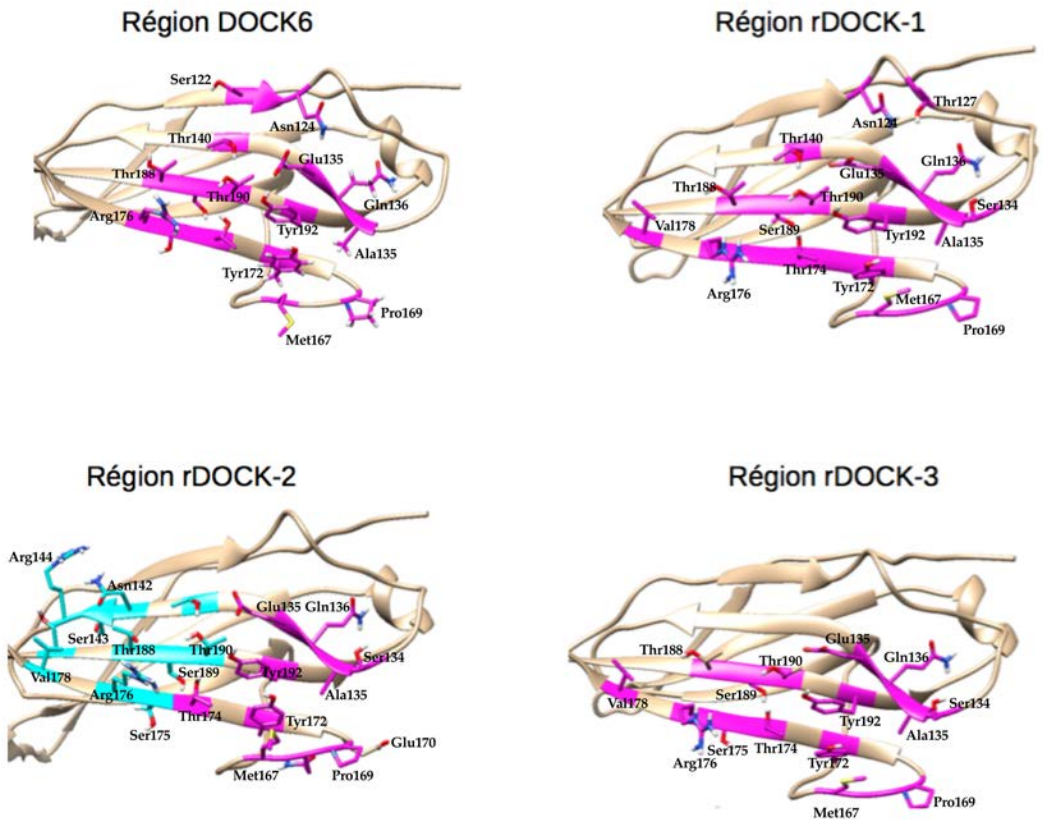


Figure 5.11: Régions occupées par les ligands DOCK6 et rDOCK (en magenta et en cyan). La région colorée en cyan correspond à la région de liaison du ligand très flexible L17 (ligand rDOCK-2).

Le grand nombre de ligands rDOCK trouvé dans le top-20 du classement final montre que le programme rDOCK a permis de mieux estimer l'énergie des complexes dès la première étape de criblage. Il semble donc que la fonction de scoring primaire rDOCK évalue mieux les complexes que la fonction de scoring primaire DOCK6.

Ligand	Liste	MW (g/mol)	HBD	HBA	LogP	PSA (Å^2)	Nombre de liaison flexible	charge globale	Score	Rang
L2	rDOCK-1	494.422	0	5	6.08	47	7	0	7.274	1
L3	rDOCK-3	440.89	1	6	6.61	72	5	0	7.248	2
ligA	**DOCK6**	**474.3**	**2**	**8**	**1.83**	**112**	**6**	**-1**	**7.110**	**3**
L5	DOCK6	522.069	4	7	3.49	92	12	+1	7.085	4
L6	rDOCK-3	449.554	0	5	6.71	62	7	0	7.025	5
L7	rDOCK-2	367.857	1	3	5.68	46	3	0	7.014	6
L8	DOCK6	501.559	2	8	5.57	116	10	-1	7.002	7
L9	rDOCK-1	414.553	2	5	6.73	62	5	0	6.977	8
L11	rDOCK-1	437.33	0	5	4.53	43	3	0	6.930	9
ligB	**DOCK6**	**440.4**	**1**	**7**	**3.76**	**102**	**7**	**-1**	**6.884**	**10**
L14	rDOCK-1	498.421	0	6	4.45	59	7	0	6.865	11
L16	rDOCK-2	455.587	2	6	6.30	85	4	0	6.857	12
L17	rDOCK-2	524.645	0	8	6.75	99	13	0	6.829	13
L20	rDOCK-1	513.57	3	8	3.68	136	6	-1	6.750	14

Tableau 5.8: Caractéristiques physico-chimiques des 12 ligands retenus après analyse et des ligands actifs ligA et ligB (en gras). Les scores de chaque ligand sont montrés dans la colonne Score. La deuxième colonne est la liste dans laquelle est issu chaque ligand. Le nombre de donneur de liaison hydrogène (HBD) et d'accepteur de liaison hydrogène (HBA) pour chaque ligand est aussi montré. Le rang de chacun de ces ligands dans ce nouveau classement apparaît dans la dernière colonne.

Enfin, lorsqu'on réalise un classement avec les 12 ligands restants et les ligands actifs ligA et ligB, on constate que les ligands L2 et L3 ont un rang supérieur à celui du ligand ligA, classé en 3^e position. Le ligand actif ligB est quant à lui classé au rang 10, avec les ligands L5-L9 et L11 qui ont une affinité prédite meilleure que celle de ligB. Ces ligands L2, L3, L5-L9 et L11 vont prochainement être testés *in vitro*.

Conclusion et perspectives

Dans cette thèse, nous avons recherché des inhibiteurs de l'interaction Lu-Lnα5 grâce à un protocole de criblable virtuel qui a préalablement été validé sur la protéine CD80. Nous avons pu identifier sur cette protéine un site de liaison similaire à celui prédit sur le second domaine N-terminal de Lu. Ce protocole de criblage contient une étape de docking, une étape de relaxation des complexes et enfin une étape d'évaluation des affinités des ligands pour la protéine cible.

Plusieurs conclusions intéressantes ont pu être obtenues à partir des travaux menés sur CD80 et sur Lu. Nous avons identifié un site de liaison précis sur CD80 ainsi qu'une pose associée aux 17 ligands étudiés sur cette protéine. Cette pose a été validée en comparant les affinités calculées par plusieurs méthodes de calculs (XSCORE, MM-PBSA et FMO) aux affinités expérimentales de ces ligands. Nous avons notamment constaté que les différents types de calcul de chimie quantique ne sont pas équivalents du point de vue de la pertinence des résultats associés puisque des calculs d'énergie d'interaction CD80-ligand avec une méthode semi-empirique (PM6-DH2X dans MOPAC) donnent des résultats insatisfaisants comparés aux résultats obtenus avec des calculs *ab initio* (FMO). Les calculs FMO nous ont notamment permis d'obtenir un découpage des énergies d'interaction par résidu, ce qui nous a renseigné sur les régions d'interaction les plus fortes et les plus faibles sur CD80. Ces résultats obtenus pour CD80 sont des résultats très importants pour la recherche d'inhibiteurs de l'interaction de cette protéine avec chacun de ses partenaires biologiques CTLA-4 et CD28, interactions impliquées dans de nombreuses maladies telles que les maladies auto-immunes et les cancers. En effet, la connaissance du site de liaison de petites molécules sur CD80 ainsi que leur pose sont des informations importantes qui permettent d'orienter les étapes de docking et de sélection de ces molécules dans un criblage virtuel. D'autres résultats importants ont été obtenus lors de l'élaboration du protocole de scoring sur CD80. D'abord, nous avons constaté que les résultats de scoring

obtenus avec FMO dépendent de la méthode d'échantillonnage utilisée. En effet, dans le système CD80-ligand que nous avons étudié, les résultats de scoring sont moins bons si on réalise un échantillonnage par clusterisation plutôt qu'un échantillonnage systématique. Puis, nous avons constaté que FMO est très sensible à la géométrie. En revanche, nos résultats ne sont pas différents si les calculs FMO sont réalisés sur des géométries minimisées avec CHARMM ou par chimie quantique. De cette observation, nous pouvons dire que la géométrie minimisée avec CHARMM ne varie pas de celle minimisée par chimie quantique et que la méthode de minimisation par mécanique moléculaire apporte, dans le cas du système CD80-ligand, un niveau de précision comparable à la méthode de minimisation par chimie quantique.

Pour la recherche de potentiels inhibiteurs de l'interaction Lu-Lnα5, nous avons utilisé un protocole légèrement différent de celui qui a permis de reproduire les affinités expérimentales des ligands sur CD80. En effet, dans le protocole appliqué à Lu, nous avons réalisé des calculs de docking avec et sans molécules d'eau explicite (programmes respectifs rDOCK et DOCK6). Puis, après avoir relaxé les complexes par une simulation DM, ce sont deux étapes d'évaluation des énergies de complexation (avec XSCORE) qui ont été réalisées (au lieu d'une seule étape dans le protocole validé sur CD80) afin de permettre la sélection des ligands. Ces étapes d'évaluation des énergies de complexation ont été réalisées sur des géométries non minimisées puisque nous avons constaté que l'étape préalable de minimisation des géométries n'est pas nécessaire lorsque les calculs d'énergie sont réalisés avec XSCORE. En revanche, les résultats de scoring obtenus avec XSCORE sont fortement dépendants du nombre de géométries pour lesquelles les énergies d'interaction sont calculées. En effet, les valeurs d'énergie de complexation ainsi que le classement des ligands qui en découle seront très différents selon que les calculs sont réalisés sur la seule géométrie du complexe protéine-ligand issue du docking ou sur un ensemble de géométries d'un même complexe extraites d'une trajectoire de simulation DM dans lequel ce complexe est relaxé. En considérant plusieurs géométries issues

d'une trajectoire de DM, les résultats de scoring sont plus précis puisque la flexibilité des complexes est prise en compte.

Trois molécules, issues d'un premier criblage de molécules ZINC, ont été testées par deux membres de l'équipe 1 expérimentale de notre UMR, Wassim El Nemer et Sylvie Cochet. Spécialiste de Lu, Wassim El Nemer a testé l'efficacité de ces molécules en utilisant la plateforme Venaflux qui permet de mimer le flux sanguin et ainsi de mesurer l'adhérence des globules rouge drépanocytaires à Lnα5 préalablement fixée. De cette étude expérimentale, deux molécules (ligA et ligB) se sont révélées actives et font actuellement l'objet de deux dépôts de brevet auprès du Bureau Européen des Brevets.

Dans le criblage réalisé dans cette étude, nous avons retenu 12 ligands. Lors de l'analyse des complexes formés entre ces ligands et Lu, nous avons constaté que la liaison de la plupart de ces ligands est renforcée par au moins un pont Lu-solvant-ligand. Enfin sur une région de liaison plane telle que celle observée pour le domaine D2 de Lu, la présence de liaison hydrogène entre les ligands et la protéine semble renforcer la stabilité des ligands dans les simulations DM. Parmi ces ligands, neuf ont un meilleur score que le ligand actif ligB ; ce dernier occupant le rang 10 dans le classement des 12 ligands restants. Parmi ces neuf premiers ligands, deux ligands ont un meilleur score que le ligand ligA qui occupe le rang 3 dans ce même classement. Ces ligands seront prochainement testés dans les mêmes conditions que pour les ligands ligA et ligB par notre équipe expérimentale.

Parmi les perspectives de ce travail, la détermination de la structure des complexes formés entre Lu et les ligands actifs ligA et ligB par cristallographie aux rayons X constituera ensuite une étape essentielle afin de valider nos prédictions. Cette perspective permettra aussi d'optimiser l'affinité des ligands ligA et ligB sur Lu après modifications que nous pourrions proposer sur leur structure. Dans cette phase d'optimisation de ces ligands, les modifications structurales proposées pourront être réalisées par l'équipe MEDCHEM du Département de Pharmacochimie Moléculaire de l'Université Joseph Four-

rier (Grenoble).

Dans une stratégie globale de développement d'une molécule à vertu thérapeutique, la condition d'une affinité élevée pour la protéine cible n'est pas une condition suffisante. L'échec de la plupart des molécules dans les phases cliniques est dû à leurs mauvaises propriétés ADME. Dans le cadre d'une collaboration qui resterait à établir, des tests précliniques pourraient être envisagés sur des modèles cellulaires afin de déterminer les propriétés d'absorption par l'intestin de ces molécules, de métabolisme par les enzymes de dégradation du foie ainsi que leur toxicité éventuelle (effets secondaires à court terme). En particulier, certaines des propriétés pharmacocinétiques des molécules pourraient être déterminées par des tests sur des microsomes hépatiques qui contiennent la plupart des enzymes impliquées dans le métabolisme des xénobiotiques (cytochromes P450 ou flavin monooxygénases). La détermination des propriétés d'absorption peut être réalisée grâce à des modèles in vitro (Caco-2 pour les tests de perméabilité).

Bibliographie

[1] N. Madeleine, F. Gardebien, Identification of Inhibitors for the Lutheran Blood Group Glycoprotein - Laminin 511/521 Interaction by Molecular Modeling and Simulation Techniques, Curr Comput Aided Drug Des.

[2] M. H. Odievre, E. Verger, A. C. Silva-Pinto, J. Elion, Pathophysiological insights in sickle cell disease, Indian J. Med. Res. 134 (2011) 532–537.

[3] R. Girot, P. Bégué, F. Galacteros, in: J. L. EUROTEXT (Ed.), La drépanocytose, 2003.

[4] M. Udani, Q. Zen, M. Cottman, N. Leonard, S. Jefferson, C. Daymont, G. Truskey, M. Telen, Basal cell adhesion molecule Lutheran protein - The receptor critical for sickle cell adhesion to laminin , J. Clin. Invest. 101 (1998) 2550–2558.

[5] W. El Nemer, P. Gane, Y. Colin, V. Bony, C. Rahuel, F. Galacteros, J. Cartron, C. Le Van Kim, The Lutheran blood group glycoproteins, the erythroid receptors for laminin, are adhesion molecules, J. Biol. Chem. 273 (1998) 16686–16693.

[6] T. Mankelow, N. Burton, F. Stefansdottir, F. A. Spring, S. F. Parsons, J. S. Pedersen, C. L. P. Oliveira, D. Lammie, T. Wess, N. Mohandas, J. A. Chasis, R. L. Brady, D. J. Anstee, The laminin 511/521-binding site on the lutheran blood group glycoprotein is located at the flexible junction of Ig domains 2 and 3, Blood 110 (2007) 3398–3406.

[7] C. Rahuel, C. L. Kim, M. G. Mattei, J. P. Cartron, Y. Colin, A unique gene encodes spliceoforms of the B-cell adhesion molecule cell surface glycoprotein of epithelial cancer and of the Lutheran blood group glycoprotein, Blood 88 (1996) 1865–1872.

[8] Y. Kikkawa, C. L. Moulson, I. Virtanen, J. H. Miner, Identification of the binding site for the Lutheran blood group glycoprotein on laminin alpha 5 through expression of chimeric laminin chains in vivo, J. Biol. Chem. 277 (2002) 44864–44869.

[9] S. F. Parsons, G. Lee, F. A. Spring, T. N. Willig, L. L. Peters, J. A. Gimm, M. J. Tanner, N. Mohandas, D. J. Anstee, J. A. Chasis, Lutheran blood group glycoprotein and its newly characterized mouse homologue specifically bind alpha5 chain-containing human laminin with high affinity, Blood 97 (1) (2001) 312–320.

[10] C. L. Moulson, C. Li, J. H. Miner, Localization of Lutheran, a novel laminin receptor, in normal, knockout, and transgenic mice suggests an interaction with laminin alpha5 in vivo, Dev. Dyn. 222 (1) (2001) 101–114.

[11] C. Spenle, P. Simon-Assmann, G. Orend, J. H. Miner, Laminin alpha 5 guides tissue patterning and organogenesis, Cell Adhes. Migr. 7 (2013) 90–100.

[12] P. Bartolucci, V. Chaar, J. Picot, D. Bachir, A. Habibi, C. Fauroux, F. Galacteros, Y. Colin, C. Le Van Kim, W. El Nemer, Decreased sickle red blood cell adhesion to laminin by hydroxyurea is associated with inhibition of Lu/BCAM protein phosphorylation, Blood 116 (2010) 2152–2159.

[13] E. Gauthier, C. Rahuel, M. P. Wautier, W. El Nemer, P. Gane, J. L. Wautier, J. P. Cartron, Y. Colin, C. Le Van Kim, Protein kinase A-dependent phosphorylation of Lutheran/basal cell adhesion molecule glycoprotein regulates cell adhesion to laminin alpha5, J. Biol. Chem. 280 (34) (2005) 30055–30062.

[14] R. K. Agrawal, R. K. Patel, V. Shah, L. Nainiwal, B. Trivedi, Hydroxyurea in sickle cell disease: drug review, Indian J Hematol Blood Transfus 30 (2) (2014) 91–96.

[15] J. A. Wells, C. L. McClendon, Reaching for high-hanging fruit in drug discovery at protein-protein interfaces, Nature 450 (7172) (2007) 1001–1009.

[16] S. Jones, J. M. Thornton, Principles of protein-protein interactions, Proc. Natl. Acad. Sci. U.S.A. 93 (1) (1996) 13–20.

[17] W. Guo, J. A. Wisniewski, H. Ji, Hot spot-based design of small-molecule inhibitors for protein-protein interactions, Bioorg. Med. Chem. Lett. 24 (11) (2014) 2546–2554.

[18] A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces, J. Mol. Biol. 280 (1) (1998) 1–9.

[19] C. Drewniok, B. G. Wienrich, M. Schon, J. Ulrich, Q. Zen, M. J. Telen, R. J. Hartig, I. Wieland, H. Gollnick, M. P. Schon, Molecular interactions of B-CAM (basal-cell adhesion molecule) and laminin in epithelial skin cancer, Arch. Dermatol. Res. 296 (2) (2004) 59–66.

[20] I. G. Campbell, W. D. Foulkes, G. Senger, J. Trowsdale, P. Garin-Chesa, W. J. Rettig, Molecular cloning of the B-CAM cell surface glycoprotein of epithelial cancers: a novel member of the immunoglobulin superfamily, Cancer Res. 54 (22) (1994) 5761–5765.

[21] D. Ansari, L. Aronsson, A. Sasor, C. Welinder, M. Rezeli, G. Marko-Varga, R. Andersson, The role of quantitative mass spectrometry in the discovery of pancreatic cancer biomarkers for translational science, J Transl Med 12 (2014) 87.

[22] Y. Kikkawa, T. Miwa, Y. Tohara, T. Hamakubo, M. Nomizu, An Antibody to the Lutheran Glycoprotein (Lu) Recognizing the LU4 Blood Type Variant Inhibits Cell Adhesion to Laminin alpha 5, PLOS One 6 (2011) e23329–e23339.

[23] Y. Kikkawa, T. Sasaki, M. T. Nguyen, M. Nomizu, T. Mitaka, J. H. Miner, The LG1-3 tandem of laminin alpha5 harbors the binding sites of Lutheran/basal cell adhesion molecule and alpha3beta1/alpha6beta1 integrins, J. Biol. Chem. 282 (20) (2007) 14853–14860.

[24] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, ZINC: A free tool to discover chemistry for biology, J. Chem. Inf. Model. 52 (2012) 1757–1768.

[25] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[26] A. P. Higueruelo, A. Schreyer, G. R. J. Bickerton, W. R. Pitt, C. R. Groom, T. L. Blundell, Atomic interactions and profile of small molecules disrupting protein–protein interfaces: the TIMBAL database, Chem. Biol. Drug. Des. 74 (2009) 457–467.

[27] R. X. Wang, L. H. Lai, S. M. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, J. Comput-Aid. Mol. Des. 16 (2002) 11–26.

[28] Y. Cao, L. Li, Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model, Bioinformatics 30 (12) (2014) 1674–1680.

[29] G. B. Li, L. L. Yang, W. J. Wang, L. L. Li, S. Y. Yang, ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions, J Chem Inf Model 53 (3) (2013) 592–600.

[30] S. Y. Huang, X. Zou, An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials, J Comput Chem 27 (15) (2006) 1866–1875.

[31] G. Neudert, G. Klebe, DSX: A knowledge-based scoring function for the assessment of protein–ligand complexes, J. Chem. Inf. Model. 51 (2011) 2731–2745.

[32] T. Hou, J. Wang, Y. Li, , W. Wang, Assessing the performance of the MM/PBSA and MM/GBSA methods: II. the accuracy of ranking poses generated from docking, J. Comput. Chem. 32 (2011) 866–877.

[33] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, M. Uebayasi, Fragment molecular orbital method: an approximate computational method for large molecules, Chemical Physics Letters 313 (3) (1999) 701 – 706.

[34] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, I. D. Kuntz, DOCK 6: Combining techniques to model RNA-small molecule complexes, RNA 15 (2009) 1219–1230.

[35] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, S. D. Morley, rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids, PLoS Comput. Biol. 10 (4) (2014) e1003571.

[36] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (3) (1990) 403 – 410.

[37] H. McWilliam, F. Valentin, M. Goujon, W. Li, M. Narayanasamy, J. Martin, T. Miyar, R. Lopez, Web services at the european bioinformatics institute-2009, Nucleic Acids Res. 37 (2009) W6–W10.

[38] J. F. Gibrat, T. Madej, S. H. Bryant, Surprising similarities in structure comparison, Curr. Opin. Struct. Biol. 6 (1996) 377–385.

[39] T. Madej, C. J. Lanczycki, D. Zhang, P. A. Thiessen, R. C. Geer, A. Marchler-Bauer, S. H. Bryant, MMDB and VAST+: tracking structural similarities between macromolecular complexes, Nucleic Acids Res. 42 (2014) D297–303.

[40] L. Holm, P. Rosenström, DALI server: conservation mapping in 3d, Nucleic Acids Res. 38 (2010) W545–549.

[41] J. I. Ito, Y. Tabei, K. Shimizu, K. Tsuda, K. Tomii, PoSSuM: a database of similar protein-ligand binding and putative pockets, Nucleic Acids Res. 40 (2012) D541–D548.

[42] R. K. Akira, N. Haruki, Similarity search for local protein structures at atomic resolution by exploiting a database management system, Biophysics 3 (2007) 75–84.

[43] L. A. Kelley, M. J. E. Sternberg, Protein structure prediction on the WEB: a case study using the Phyre server, Nature Prot. 4 (2009) 363–371.

[44] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, Computational methods in drug discovery, Pharmacol. Rev. 66 (1) (2014) 334–395.

[45] A. T. Laurie, R. M. Jackson, Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites, Bioinf. 21 (2005) 1908–1916.

[46] C. H. Ngan, D. R. Hall, B. Zerbe, L. E. Grove, D. Kozakov, S. Vajda, FTSite: high accuracy detection of ligand binding sites on unbound protein structures, Bioinf. 28 (2012) 286–287.

[47] M. Hernandez, D. Ghersi, R. Sanchez, SITEHOUND-WEB: a server for ligand binding site identification in protein structures, Nucleic Acids Res. 37 (2009) W413–W416.

[48] M. Brylinski, W. P. Feinstein, eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands, J. Comput-Aid. Mol. Des. 27 (2013) 551–567.

[49] R. A. Laskowski, SURFNET - a program for visualizing molecular-surfaces, cavities, and intermolecular interactions, J. Mol. Graph. 13 (1995) 323–330.

[50] N. J. Green, J. Xiang, J. Chen, L. Chen, A. M. Davies, D. Erbe, S. Tam, J. F. Tobin, Structure-activity studies of a series of dipyrazolo[3,4-b:3',4'-d]pyridin-3-ones binding to the immune regulatory protein B7.1, Bioorg. Med. Chem. 11 (2003) 2991–3013.

[51] R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, PROCHECK - a program to check the stereochemical quality of protein structures, J. Appl. Cryst. 26 (1993) 47–60.

[52] G. Vriend, WHAT IF: a molecular modelling and drug design program, J. Mol. Graph. 8 (1990) 52–56.

[53] B. R. Brooks, R. Bruccoleri, B. Olafson, D. J. States, S. Swaminathan, M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, J. Comput. Chem. 4 (1983) 187–217.

[54] D. V. Erbe, S. Wang, Y. Xing, J. F. Tobin, Small molecule ligands define a binding site on the immune regulatory protein B7.1, J. Biol. Chem. 277 (2001) 7363–7368.

[55] K. Uvebrant, T. D. da Graça, A. Rosén, M. Akesson, H. Berg, B. Walse, P. Björk, Discovery of selective small-molecule CD80 inhibitors, J. Biomol. Screen. 12 (2007) 464–472.

[56] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A. D. Mackerell, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, J Comput Chem 31 (4) (2010) 671–690.

[57] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian 09 Revision C.01, Gaussian, Inc., Wallingford, CT (2009).

[58] J. J. P. Stewart, *MOPAC2012*. Version 1.0, Fujitsu Limited, Tokyo, Japan (2012).

[59] Lemons, D. S. and Gythiel, A., Paul langevin's 1908 paper 'On the theory of brownian motion', Am. J. Phys. 65 (1997) 1079–1081.

[60] A. Brünger, C. L. Brooks III, M. Karplus, Stochastic boundary conditions for molecular dynamics simulations of ST2 water, Chem. Phys. Lett. 105 (1984) 495–500.

[61] J. Ryckaert, G. Ciccotti, H. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, J. comput. Phys 23 (1977) 327–341.

[62] D. G. Fedorov, T. Nagata, K. Kitaura, Exploring chemistry with the fragment molecular orbital method, Phys Chem Chem Phys 14 (21) (2012) 7562–7577.

[63] T. Nemoto, D. G. Fedorov, M. Uebayasi, K. Kanazawa, K. Kitaura, Y. Komeiji, Ab initio fragment molecular orbital (fmo) method applied to analysis of the ligand–protein interaction in a pheromone-binding protein, Computational Biology and Chemistry 29 (6) (2005) 434 – 439.

[64] D. G. F. Michael P. Mazanetz, Ewa Chudyk, Y. Alexeev, Applications of the fragment molecular orbital method to drug research, Computer-Aided Drug Discovery 29 (2016) 217–255.

[65] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, General atomic and molecular electronic structure system, J. Comput. Chem. 14 (1993) 1347–1363.

[66] T. Ishikawa, K. Kuwata, RI-MP2 gradient calculation of large molecules using the fragment molecular orbital method, J. Phys. Chem. Lett. 3 (3) (2012) 375–379.

[67] Parallelized ab initio calculation system based on FMO, http://www.paics.net/index_e.html.

[68] J. D. Yesselman, D. J. Price, J. L. Knight, C. L. Brooks, MATCH: An atom-typing toolset for molecular mechanics force fields, J. Comput. Chem. 33 (2012) 189–202.

[69] A. R. Leach, Molecular Modelling: Principles and Applications (2nd Edition), Prentice Hall, 2001.

[70] Coulomb, C.-A., Premier mémoire sur l'électricité et le magnétisme, Prentice Hall, 1785.

[71] W. Pauli, Nobel Lectures, Physics, Exclusion Principle and Quantum Mechanics, Dec. 13, 1946, in: Nobel Lectures, 1946.

[72] Guvench, O. and MacKerell, A. D. Jr., Comparison of protein force fields for molecular dynamics simulations, in: Walker, J. M. and Kukol, A. (Ed.), Methods in Molecular Biology, Humana Press, Totowa, NJ, 2008, pp. 63–88.

[73] H. A. Lorentz, Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase, Annalen der Physik 248 (1881) 127–136.

[74] G. Sutmann, Classical molecular dynamics, in: J. Grotendorst, D. Marx, A. Muramatsu (Eds.), Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms. Lecture Notes, NIC Series Volume 10, 2002, pp. 211–254.

[75] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, A geometric approach to macromolecule-ligand interactions, J. Mol. Biol. 161 (2) (1982) 269–288.

[76] N. Brooijmans, I. D. Kuntz, Molecular recognition and docking algorithms, Annu Rev Biophys Biomol Struct 32 (2003) 335–373.

[77] F. Lopez-Vallejo, T. Caulfield, K. Martinez-Mayorga, M. A. Giulianotti, A. Nefzi, R. A. Houghten, J. L. Medina-Franco, Integrating virtual screening and combinatorial chemistry for accelerated drug discovery, Comb. Chem. High Throughput Screen. 14 (6) (2011) 475–487.

[78] M. P. Repasky, M. Shelley, R. A. Friesner, Flexible ligand docking with Glide, Curr Protoc Bioinformatics Chapter 8 (2007) Unit 8.12.

[79] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (3) (1996) 470–489.

[80] C. Darwin, On the origin of species 6.

[81] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, C. R. Corbeil, Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go, Br. J. Pharmacol. 153 Suppl 1 (2008) 7–26.

[82] L. G. Ferreira, R. N. Dos Santos, G. Oliva, A. D. Andricopulo, Molecular docking and structure-based drug design strategies, Molecules 20 (7) (2015) 13384–13421.

[83] R. Dias, W. F. de Azevedo, Molecular docking algorithms, Curr Drug Targets 9 (12) (2008) 1040–1047.

[84] S. Y. Huang, S. Z. Grinter, X. Zou, Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions, Phys Chem Chem Phys 12 (40) (2010) 12899–12908.

[85] P. J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, J. Med. Chem. 28 (7) (1985) 849–857.

[86] L. Zhang, J. Skolnick, How do potentials derived from structural databases relate to "true" potentials ?, Protein Sci. 7 (1) (1998) 112–122.

[87] P. D. Thomas, K. A. Dill, Statistical potentials extracted from protein structures: how accurate are they ?, J. Mol. Biol. 257 (2) (1996) 457–469.

[88] I. Newton, Mathematical Principles of Natural Philosophy (1729).

[89] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflish, 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, M. Karplus, CHARMM: The biomolecular simulation program, J. Comput. Chem. 30 (2009) 1545–1614.

[90] R. Hockney, The potential calculation and some applications, Meth. Comp. Phys. 9 (1970) 136–211.

[91] H. Fan, A. E. Mark, J. Zhu, B. Honig, Comparative study of generalized Born models: protein dynamics, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 6760–6764.

[92] R. Kubo, The fluctuation-dissipation theorem, Rep. Prog. Phys. 29 (1966) 255–284.

[93] C. Brooks III, Deformable stochastic boundaries in molecular dynamics, J. Chem. Phys. 79 (1983) 6312–6325.

[94] A. Brünger, C. L. Brooks, M. Karplus, Molecular dynamics with stochastic boundaries: application to the active site of proteins in solution, in: Molecular Dynamics and Protein Structure, Proceedings of a Workshop, J. Hermans, University of North Carolina, 1985, pp. 16–17.

[95] A. T. Brünger, C. L. Brooks, M. Karplus, Active site dynamics of ribonuclease, Proc. Natl. Acad. Sci. U. S. A. 82 (1985) 8458–8462.

[96] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, ChEMBL: a large-

scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107.

[97] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, BindingDB: a WEB-accessible database of experimentally determined protein-ligand binding affinities, Nucleic Acids Res. 35 (2007) D198–D201.

[98] G. M. Boratyn, A. A. Schaffer, R. Agarwala, S. F. Altschul, D. J. Lipman, T. L. Madden, Domain enhanced lookup time accelerated BLAST, Biol. Direct 7 (2012) 12.

[99] A. Prlic, S. Bliven, P. W. Rose, W. F. Bluhm, C. Bizon, A. Godzik, P. E. Bourne, Pre-calculated protein structure alignments at the RCSB PDB website, Bioinformatics 26 (23) (2010) 2983–2985.

[100] G. H. Li, J. F. Huang, CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation, BMC Bioinf. 11 (2010) 419–451.

[101] Y. Lin, S. Yoo, R. Sanchez, SiteComp: a server for ligand binding site analysis in protein structures, Bioinf. 28 (2012) 1172–1173.

[102] M. J. Basse, S. Betzi, R. Bourgeas, S. Bouzidi, B. Chetrit, V. Hamon, X. Morelli, P. Roche, 2P2Idb: A structural database dedicated to orthosteric modulation of protein-protein interactions., Nucleic Acids Res. 41 (2013) 824–827.

[103] C. Stamper, Y. Zhang, J. Tobin, D. Erbe, S. Ikemizu, S. Davis, M. Stahl, J. Seehra, W. Somers, L. Mosyak, Crystal structure of the B7-1/CTLA-4 complex that inhibits human immune responses, Nature 410 (2001) 608–611.

[104] J. Bostrom, A. Hogner, S. Schmitt, Do structurally similar ligands bind in a similar fashion ?, J. Med. Chem. 49 (23) (2006) 6716–6725.

[105] Accelrys Software Inc., Discovery Studio Modeling Environment, Release 4.0, San Diego: Accelrys Software Inc., 2013.

[106] J. Wang, P. Morin, W. Wang, P. A. Kollman, Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA, J. Am. Chem. Soc. 123 (22) (2001) 5221–5230.

[107] COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient.

[108] M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, F. Fanelli, Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces, J Comput Chem 32 (6) (2011) 1183–1194.

[109] B. Mennucci, Polarizable continuum model, Wiley Interdisciplinary Reviews: Computational Molecular Science 2 (3) (2012) 386–404.

[110] D. G. Fedorov, K. Kitaura, Pair interaction energy decomposition analysis, J. Comp. Chem. 28 (2007) 222–237.

[111] S. Grimme, Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies, J. Chem. Phys. 118 (2003) 9095–9102.

[112] P. Zhou, J. Zou, F. Tian, Z. Shang, Fluorine bonding–how does it work in protein-ligand interactions ?, J Chem Inf Model 49 (10) (2009) 2344–2355.

[113] C. Bissantz, B. Kuhn, M. Stahl, A medicinal chemist's guide to molecular interactions, J. Med. Chem. 53 (14) (2010) 5061–5084.

[114] C. C. Valley, A. Cembran, J. D. Perlmutter, A. K. Lewis, N. P. Labello, J. Gao, J. N. Sachs, The methionine-aromatic motif plays a unique role in stabilizing protein structure, J. Biol. Chem. 287 (42) (2012) 34979–34991.

[115] J. C. Faver, M. N. Ucisik, W. Yang, K. M. Merz, Computer-aided drug design: Using numbers to your advantage, ACS Med. Chem. Lett. 4 (2013) 812–814.

[116] B. O. Villoutreix, M. A. Kuenemann, J. L. Poyet, H. Bruzzoni-Giovanelli, C. Labbe, D. Lagorce, O. Sperandio, M. A. Miteva, Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology, Mol Inform 33 (6-7) (2014) 414–437.

[117] T. Tuccinardi, G. Poli, V. Romboli, A. Giordano, A. Martinelli, Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies, J. Chem. Inf. Model. 54 (2014) 2980–2986.

[118] M. Levitt, M. F. Perutz, Aromatic rings act as hydrogen bond acceptors, J. Mol. Biol. 201 (4) (1988) 751–754.

[119] S. Sirimulla, J. B. Bailey, R. Vegesna, M. Narayan, Halogen interactions in protein-ligand complexes: implications of halogen bonding for rational drug design, J Chem Inf Model 53 (11) (2013) 2781–2791.

Annexes

Identification of inhibitors for the Lutheran blood group glycoprotein – Laminin 511/521 interaction by molecular modelling and simulation techniques

Noëlly Madeleine, Fabrice Gardebien*
DSIMB, INSERM, U1134, Paris, F-75015, France
and
Université de la Réunion, UMR_S 1134
Faculté des Sciences et Technologies, 15, avenue René Cassin, BP 7151
97715 Saint Denis Messag Cedex 09, La Réunion, France
and
Institut National de la Transfusion Sanguine, F-75015 Paris, France
and
Laboratory of Excellence GR-Ex

Abstract

In the pathogenesis of vaso-occlusive crises of sickle cell disease, red blood cells bind to the endothelium and promote vaso-occlusion. At the surface of these sickle red blood cells, the overexpressed protein Lutheran strongly interact with the Laminin 511/521. To hinder this protein-protein interaction, a virtual screening was performed with 395 601 coumponds that target Lutheran. Prior validation of a robust docking and scoring protocol was considered on the protein CD80 because this protein has a binding site with similar topological and physico-chemical characteristics and it also has a series of ligands with known affinity constants. This protocol consisted of multiple filtering steps based on docked scores, molecular dynamics simulations, post-screening scores, and molecular properties. We identified four molecules that have good structural and physico-chemical properties. These four molecules thus represent promising candidates to hinder this protein-protein interaction.

Keywords: Drepanocytosis; Lutheran protein; protein-protein interaction; molecular docking; molecular dynamics simulations; scoring function.

1 Introduction

Drepanocytosis is a genetic blood disorder characterized by red blood cells that assume an abnormal, rigid, sickle shape. Sickling decreases the cells' flexibility and results in a risk of various

*To whom correspondence may be addressed. E-mail: Fabrice.Gardebien@univ-reunion.fr Fax: +262(0)262-93-82-37.

complications. The hallmark of sickle cell disease is episodic and painful vaso-oclusion. Vaso-occlusion is caused by a strong adhesion of sickle red blood cells and leukocytes to the vascular endothelium. By adhering to the vascular endothelium, erythrocytes and leukocytes reduce the lumen of the vessels and reduce the blood flow. *In vitro* studies indicate that the adhesion of sickle erythrocytes to the vascular endothelium is important in the pathogenesis of vaso-occlusive crises of sickle cell disease [1, 2].

This phenomenon is explained by an abnormal expression and activation of several membrane proteins on the surface of sickle red blood cells that strongly interact with the components of the endothelium and the sub-endothelial matrix. Adhesion proteins which are responsible for the vaso-occlusion are both Lu/BCAM and ICAM-4 expressed by the mature red blood cells and reticulocytes. In the case of sickle cell disease, these cells overexpress Lu/BCAM. Both Lu/BCAM and ICAM-4 promotes adhesion of sickle erythrocytes to the endothelium through the interaction with Laminin 511/521 of the sub-endothelial extracellular matrix and endothelial integrin αVβ3, respectively [3].

Lutheran (Lu) blood group antigens and the basal cell adhesion molecule antigen (BCAM) are carried out by two glycoprotein isoforms of 85 and 78 kDa, respectively, and differ in the length of their C-terminal cytoplasmic domain. Therefore, Lu and BCAM share identical extracellular domains composed of one V-set and four C2-set domains [4]. Since these extracellular domains are identical and are those of interest in this study, this protein will be designated as Lu in the following. In the case of sickle cell disease, the glycoprotein Lu is the only receptor for Laminin 511/521 [5, 6]. Laminins (Ln) are major protein components of the basal lamina and are heterotrimers formed by three subunits designed α, β and γ chains. These subunits form 16 isoforms of Ln with Ln 511 and 521 which both contain the $\alpha5$ chain. Ln$\alpha5$ consists of five globular domains LG1-LG5 at its C-terminus which can specifically interact with few integrins, α-dystroglycan, syndecan-4 and Lu [7]. Lu seems to interact with a binding site found only in the $\alpha5$ chain. No crystallographic structure of these domains of Ln$\alpha5$ is currently available.

The only drug currently approved by the US Food and Drug Administration (FDA) for the treatment of sickle cell disease is hydroxyurea. Hydroxyurea inhibits the intracellular phosphorylation of Lu that happens before interaction with Ln$\alpha5$. This treatment significantly reduced the vaso-occlusive crises in adults with sickle cell disease, but is less effective in children [8]. Furthermore, hydroxyurea treatment causes significant side effects such as bleeding or convulsions. It is therefore important to find other drug candidates to reduce the vaso-occlusion in patients with sickle cell disease. The inhibition of the interaction between Lu and Ln$\alpha5$ is another strategy to fight the vaso-occlusion crises. This strategy is also of major interest in the research of drugs against cancer since it has been found that the binding of Lu to Ln$\alpha5$ promotes tumor cell migration [9]. The search of protein–protein interaction (PPI) inhibitors is notoriously difficult but is an important area in drug discovery because of the primary role of this type of interaction in many diseases. The search for inhibitors of the interaction Lu-Ln$\alpha5$ has already been carried out in 2011 by Kikkawa *et al.* [10]. They have found that an antibody directed against the second N-terminal domain of Lu could lead to an inhibition of the interaction by steric hindrance.

To identify a PPI inhibitor with a high probability of binding to Lu for the inhibition of the Lu-Ln$\alpha5$ interaction, a virtual screening procedure was used that consists in the following four steps: (a) identification of a ligand binding site for the compounds, (b) docking of compounds from the molecular database ZINC [11, 12], (c) geometric relaxation of the protein–compound complexes by molecular dynamics (MD) simulations, and (d) secondary scoring after the relaxation procedure

by using a more sophisticated calculation of the protein–compound interaction energy. The steps b to d will be hereafter referred to as docking and scoring protocol.

To find the binding site on Lu (step a) and to derive a reliable and robust docking and scoring protocol, we wanted to find a homology or analogy relationships between Lu and one or more proteins in interaction with a partner (either a small molecule or another protein). Indeed, in performing this search, our purpose was twofold: first, the binding site on Lu can be expected to be the same as for a homolog or similar protein that binds with a partner; second, a docking and scoring protocol can be tested and improved by using a homolog or analog of Lu that binds a series of ligands with known affinity constants. This last purpose rests on the observation that the performance of a given docking and scoring protocol is system dependent. Thus, performance is expected to be similar for similar protein system. Unfortunately, although remote homologs could be found, none of them satisfied the condition of having known ligands with known affinity constants. To derive a robust docking and scoring protocol, we therefore used the V-set domain of the protein CD80 that satisfied the three following conditions: (i) a crystal structure of this domain exists in the Protein Data Bank (PDB) [13–15]; (ii) a series of ligands with known binding constants are available; (iii) its binding site shares some structural similarities with that, predicted, of Lu. It is worth mentioning that a lot immunoglobulin domains of antibodies in complex with other molecules exist in the PDB; however, the binding region involved is always the antigen-binding sites that consist of the variable loops (the complementarity determining regions), and is thus very different in topology from the ligand binding region of CD80 that is found on top of $\beta-$strands. The protocol that was validated was applied to Lu and consisted in the three following steps: first, screening the molecular database ZINC [11, 12]; second, performing MD simulations to provide flexibility to the selected complexes; third, evaluating the interaction energy of all the complexes to find the best binders.

We shall begin by describing all the search, modelling, simulation protocols in the section 2. The following results will then be presented: in section 3.1, the docking and scoring procedure that was validated for CD80 and, in section 3.2, the results of this procedure that was then applied to Lu for the screening and filtering of the compounds extracted from the ZINC database [11, 12]. The results for the compounds that were found as good binding candidates will then be discussed.

2 Materials and Methods

2.1 Identification of similar systems

Our screening procedure for the identification of molecules with a high probability of binding to Lu comprises the following steps: (i) identification of a ligand binding site for the compounds, (ii) docking and scoring (steps b to d as above-defined).

For the identification of a ligand binding site on Lu, four different approaches were employed to find homologs or analogs. First, a BLAST analysis was performed to find homologous that have known partners (protein or compound) and for which an experimental structure of the complex exists. Second, for the proteins with low sequence identities (less than 30%), their structures were aligned with those of Lu since the homology relationship is not guaranted below this threshold; hence, relying on both the sequences and the structures allows finding remote homology relationship. Third, another finding strategy using PDBeFOLD [16], VAST [17], VAST+ [18], DALI [19], and the 3D-similarity tool in the PDB [13] consists in identifying distant homologs that cannot be

recognized by sequence comparison and that relies on protein three-dimensional structure alignment only. Fourth, local similarities with several putative binding site on Lu were searched by more local structural alignment algorithms such as POSSUM [20], GIRAF [21], Phyre2 [22] (section 3.1). Unfortunately, although remote homologs could be found, none of them satisfied the condition of having known ligands with known affinity constants. Thus, in step a, to predict the binding site on Lu, we relied on the consensus results obtained from four binding site prediction webservers: Q-SiteFinder [23], eFindSite [24], FTSite [25], and SiteHound [26]. These webservers predicted a large binding region on the second N-terminal domain of Lu that is mainly defined by charged and hydrophilic residues near residues Glu132/Asp133 (top of Figure 1).

[Figure 1 about here.]

The purpose of finding homologs or analogs was not only to find a binding site on Lu, but also to challenge a docking and scoring protocol on known protein-ligand systems. For this purpose, we finally selected the protein CD80 in the TIMBAL database [27]. This protein contains a V-set domain for which a series of compounds were designed to inhibit its interaction with the protein CTLA-4. However, the ligand binding region is unknown. To delineate this region, we therefore relied (i) on the crystallographic structure of the CD80–CTLA-4 complex (PDB code 1I8L), (ii) on binding site prediction webservers, and (iii) on mutational studies [28]. The following webservers for prediction of binding sites were used: eFind-Site [24], Surfnet in Metapocket [29], FT-Site [25], and Site-Hound [26]. Only Surfnet and eFind-Site have provided interaction sites that are in agreement with the few experimental mutation studies [28]. In the prediction result of binding pocket of CD80 by eFind-Site, the residues Tyr31 and Arg29 are predicted, as well as the residues Met43, Ser44, Val83, Leu85, and Thr41. Except for Ser44, all these residues are displayed on part (a) of Figure 2. Such residues coincide with those that interact with CTLA-4 [30]. Though not found in the interaction site of CD80 with CTLA-4, the residues Asn48 and Trp50 are predicted by Surfnet [29] and are considered to be important in the binding between CD80 and small inhibitory molecules; indeed, the CD80 mutants W50A and N48K decreased the affinity of the inhibitory molecules [28].

[Figure 2 about here.]

Despite the low percentage of sequence identity of 16 % between the C2-set domain of Lu and the V-set domain of CD80, the analysis of these two domains indicates similarities in terms of topology and polarity (Figure 2). Indeed, webserver-based prediction of the interaction site of Lu indicates that this site consists of $\beta-$strand forming residues, which is also the case of the CD80 binding site (Figure 2). For the two proteins, the $\beta-$strands thus correspond to a shallow interaction site. Moreover, the polarity of these regions is also similar (Figure 2): hydrogen bond forming residues were found in greatest proportion (Ser, Thr, Asn, and Gln) along with some hydrophobic residues (aromatic and aliphatic) and charged residues (Arg and Glu). In parts (c–e) of Figure 2, the backbone of the structure of the V-set domain of CD80 (in green) have been superimposed on that of Lu (in khaki). First, one can observe that the respective strands in the center of the binding zone (circled in purple) are almost perfectly superimposed (note that, for example in Figure 2(e), the two structures were slightly shifted for clarity purpose). Second, the charged residues are mainly found on the boundary of this binding center (Figure 2(c),(d)) while the hydrophobic residues, which include tyrosine and threonine residues, are found in large

quantities inside this binding center (Figure 2(e)). Since the performance of a given docking and scoring protocol is system dependent [31], the similarities between these two domains lend confidence that the protocol able to reproduce the experimental results for CD80 will provide reliable results for Lu. Finally, CD80 was also chosen because of the availability of more than a hundred of ligands with known affinity constants (IC50) [32]; this will be helpful when inferring the ligand binding poses from the structure-activity relationships (section 3.1).

2.2 Preparation of the protein structures

The crystal structures of Lu and CD80 were obtained from the PDB [13–15] (PDB code 2PET [33] for Lu, and 1I8L [30] and 1DR9 [34] for CD80). Only the two N-terminal domains of Lu and CD80 are available in these respective structures.

Due to the poor resolution of the two PDB structures containing CD80 (3 Å for each of the available structures with PDB codes 1DR9 [34] and 1I8L [30]), an optimization of the V-set domain structure was performed, followed by an analysis with PROCHECK and WHATIF [35, 36]. During this geometry optimization with the CHARMM program, the V-set domain was immersed in a layer of water; the force field CHARMM36 was used [37]. PROCHECK and WHATIF are programs that allow a detailed examination of the stereochemistry of a protein structure [35, 36]. As the stereochemical and packing issues were similar for both initial crystallographic structures (1DR9 and 1I8L), the structure of 1DR9 was chosen for further structure refinement and for the docking protocol. This choice was motivated by the comparison of the B-factors of the two respective crystallographic structures, the lower average values being observed for 1DR9 (41 ± 11 Å^2 for 1DR9 and 52 ± 25 Å^2 for 1I8L).

Before any docking step and in a view to improving the overall stereochemistry and the packing of the atoms of the targeted domain of CD80, a minimization procedure was considered after the immersion of the entire domain in a shell of water molecules: only the solvent molecules were first minimized for 1000 steps, followed by 5000 steps minimization of the whole system. All the minimisation steps were performed with the conjugate gradient algorithm and the root-mean-square energy gradient was less than 0.5 kcal/(mol.Å); the electrostatic and van der Waals interactions were truncated at the range of 11-14 Å using force switching. After the minimization, the sterochemical issues (bond length and torsion and valence angles for the backbone and side chains) along with the packing issue for the residues in binding site Arg29, Tyr31, Gln33, Lys36, Thr41, Met43, Asp46, Asn48, and Trp50 were solved. Among the solved issues, unusual ϕ and ψ dihedral angles were solved for Lys36; unusual χ_1 and χ_2 dihedral angles were solved for Tyr31 (for which the χ_2 value was not between -90 and 90) and for Gln33, Met43, Asp46, Asn48, and Trp50 (for which, in the crystallographic structure, the respective χ_1 value was more than 2.0 standard deviations away from the "ideal" mean value). Moreover, abnormal short interatomic distances resulting from steric hindrance were solved for Arg29 (Arg29-N $\leftrightarrow$ Thr28-CG2), Gln33 (Gln33-NE2 $\leftrightarrow$ Lys36-CA), and Met43 (Met43-SD $\leftrightarrow$ Thr41-CG2). The final minimized structure of CD80 was used in the docking step (*vide infra*). Minimizations were performed with the program CHARMM and by using the CHARMM36 force field [37].

No further relaxation of the crystallographic structure of Lu was considered given (i) its high resolution (1.7 Å) and (ii) the absence of stereochemical or packing problems in the targeted domain (residues 116–231) that were checked by the programs PROCHECK [35] and WHATIF [36].

2.3 Validation of the docking and scoring methodology

To validate the screening protocol, we first applied the above-mentioned steps b–d, as defined in the Introduction, to another protein, CD80, that contains a V-set domain for which a series of compounds exist that were designed to inhibit its interaction with the protein CTLA-4. In step b, the docking computations were performed using DOCK6 [38]. Rather than using ZINC compounds, in this validation step, nine known ligands of CD80 with a wide range of IC50 values (from 7 nM up to 6 000 nM) were chosen (Table 1). We could not perform redocking calculations to validate our protocol by reproducing an experimental protein–ligand complex since no crystallographic structure of Lu or CD80 in complex with a small ligand exists.

The docking procedure accounts for ligand flexibility while the protein is treated as rigid. One or more anchor fragments (e.g., rigid units, such as rings with six or more atoms of each compound) are overlaid in 100 orientations on the identified binding site of the protein. The remainder of the ligand is then built bond after bond, with a rotation around each added bond in 10° increments to identify the most favorable orientation based on the interaction energy (van der Waals and electrostatic contributions only) between the protein and the compound under construction [38]. Intermolecular interaction energies calculated only with van der Waals and electrostatic contributions lead to an initial ranking of the nine compounds (primary scoring after the docking step b). However, this ranking generally lacks precision, usually due to neglected solvent effect and to a rigid protein structure. Thus, in step d, the intermolecular interaction energies were further evaluated by secondary scoring methods that are more accurate such as MM-GBSA [39,40], XSCORE [41], and DrugScore [42] for comparison with the nine experimental IC50 values.

Prior to this energy re-evaluation performed in step d, these compounds are subjected to MD simulations in step c to account for protein flexibility and for the solvent effect. In this step c, both the protein and the compound can change their structure such as to increase their interaction and complementarity (*e.g.* a hydrogen bond donor in the protein can move to face an acceptor in the compound). This step therefore allows overall structure refinement for the subsequent secondary scoring procedure (step d). In this last step, the force field CHARMM36 was used for the protein and the force field CGenFF was used for the ligands (hereafter referred to as CHARMM36–CGenFF) [43]. Three alternative representations of the water surrounding were tested for the MD simulations: no water, implicit representation, and explicit representation. For the implicit solvent representation, Langevin molecular dynamics (LD) [44] simulations were performed that allow approximating the effects of solvent molecules not explicitly present in the system in terms of a frictional drag on the solute (protein and ligand), as well as random bumps associated with the thermal motions of the solvent molecules. For the explicit solvent representation, stochastic boundary molecular dynamics (SBD) [45] simulations of each protein–ligand complex were also carried out where the protein and ligand are immersed in a sphere of water molecules of 32-Å radius. This sphere size is sufficient to provide a hydration shell around the binding site of the complex. A potential was imposed at the water–vacuum boundary to avoid evaporation.

All the LD and SBD simulations were performed with the CHARMM program [37]. All the protein-ligand complexes were first energy-minimized and then heated from 0 to 300 K using a step of 10 K. Heating (60 ns) and equilibration (80 ns) periods were followed by a production period (200 ps). This relatively short simulation length will be sufficient for local geometric relaxations in the presence of water. Nevertheless, this length could not be too much extended since

our goal was to apply this same protocol to find Lu binders in a large subset of ZINC database. In the simulations, the electrostatic and van der Waals interactions were truncated at the range of 11-14 Å using force switching. An integration time step of 1 fs was used along with the SHAKE algorithm to constrain all covalent bonds involving hydrogen atoms [46]. Coordinates were saved every 250 fs, yielding a total of 800 conformations from which 40 and 100 structures that are evenly spaced were selected for the secondary scoring.

The following secondary scoring methods were evaluated on the nine complexes: two variants of the MM-GBSA approach (hereafter named GB^{OBC} [40] avalaible in the AMBER module of DOCK6 and GB^{HCT} [39] avalaible as secondary scoring function in DOCK6), DrugScore [42], and XSCORE [41].

2.4 Screening procedure for Lu

The docking and scoring protocol (steps b to d) that yielded the best results for CD80 was then applied to Lu. However, for Lu the step b here corresponded to the screening of the ZINC database [11, 12] whereas only nine known compounds were used for CD80 to set up this protocol. In this step b, the docking procedure that provided the results that are consistent with experimental data for CD80 was used for Lu.

In the step b, 395 601 compounds were docked: 265 881 compounds were extracted from the Clean Drug-like category (compounds with drug-like properties); 129 720 compounds were obtained from All-clean category (compounds that are purchasable). In the Clean Drug-like and All-clean categories, all the compounds have no functional groups known to initiate toxicity [11]. From the docking step of these 395 601 compounds using a rigid structure of the protein, 9 000 of these compounds were ranked on the basis of the primary scoring that uses van der Waals and electrostatic interaction energy only. The MATCH program [47] could derive CGenFF force field parameters for only 5 048 of the 9 000 compounds. Among these 5 048 ligands, only 82 were retained on the basis of the following criteria: hydrogen bonds should be formed with the residues Gln136 and Glu137 to replace those formed with a few crystallographic water molecules whose B-factors are lower than 25 Å^2; alternatively, the position of the ligand should not overlap with that of these water molecules.

In step c, SBD simulations were then performed for the selected ligands. The analysis of all the MD trajectories allowed us to eliminate: (i) ligands that are making only transient contacts with the protein (unstable); (ii) ligands for which displacement during the simulations shows some atoms that do not perfectly substitute the water molecules with B-factors lower than 25 Å^2; (iii) ligands for which displacement during the simulations leads to the burying of donor or acceptor groups. For each of the remaining 50 ligands at this step, 100 snapshots were extracted from the respective trajectory and finally minimized in the presence of water.

In step d, for each ligand retained, the scoring by XSCORE was evaluated for the 100 extracted snapshots and then averaged to derive the final score.

3 Results

3.1 Docking and scoring protocol validation

The program DOCK6 generates different poses for a ligand on a defined interaction site. In the case of CD80, this site is delineated by the residues shown in figure 2. However, these residues define a large region. Consequently, a huge amount of poses can be generated for the nine chosen ligands. The structures of these ligands are shown in Table 1. To obtain an unique pose for each ligand, we relied on the assumption that similar ligands exhibit similar binding poses [48] and on a parallel analysis of (i) 50 different poses that DOCK6 generated for each of the nine ligands (which amounts to 450 complexes) and of (ii) the relationship between the structures of 70 other ligands and their published affinity constants [32]. Indeed, by examining the incidence of the structure variation among this large ligand series on their measured affinities, the poses generated for the nine ligands were rejected or classified as compatible with the experimental data. It is also worth mentioning that this series of 70 ligands are structurally similar to the nine selected ligands [32]. To further help in the decision-making process, the poses were chosen or rejected based on conditions that favor a low enthalpy of interaction between the protein and the ligand. Such conditions are a correspondence between the hydrophylic moieties, on the one hand, and between the hydrophobic moities, on the other hand. Also, a maximum number of hydrogen bonds must be formed between the protein and the ligand, and there must have no buried hydrogen bond donor or acceptor in each of the binding partner. Among the hundred of inhibitors available for this protein, the nine chosen inhibitors (Table 1) were selected not only because of the wide range of binding constants covered, but also because of their diverse physico-chemical and interaction properties, therefore (i) providing many of the above conditions to satisfy and (ii) optimizing our chance to find the real poses.

[Table 1 about here.]

Validation of the scoring is an important step to validate the protocol. To this aim, not only the poses should be correct but the interaction energy should also be correctly estimated. Hence, once the poses are found to be compatible with all the experimental data and the above-defined conditions, to further validate the protocol, the following two other conditions should be met: (i) the relative order of the nine experimental IC50 values are reproduced and (ii) each of the nine ligands bound to the protein is stable during the molecular dynamics trajectory. Ideally, a linear relationship should connect the experimental affinity values to those calculated, but this third condition is almost never met. In particular, it is notoriously difficult to obtain correlation coefficients between calculated and experimental affinities that exceed 0.7 [49], whatever is the secondary scoring function.

To reproduce the ranking of the nine ligands, the method GB^{OBC} was first applied. This method allows performing MD simulations without water to relax the geometries of the poses obtained from DOCK6. When performing such MD simulations, a large amplitude movement is observed for the carboxylate group of the ligands **2** and **8**: this group is shifted by about 10 Å in the MD simulations, which is not expected to be physically meaningful. Both set of charges, AM1 and Gasteiger, were used and similar results were obtained in both cases. This large positionnal shift is the consequence of a strong charge-charge interaction between the groups COO^- of the ligand and a remote NH_3^+ group of the protein, this interaction being not screened because of the

absence of water. We then used the CHARMM program with an explicit solvent model.

Each ligand pose was validated by simulations whenever, at least, three MD simulations provided stable and consistent positions for the ligand with respect to the protein. A snapshot of the complex with the ligand **8** (ligand of higher affinity) is displayed in Figure 3. For each of the nine ligands, 40 evenly spaced geometries were selected in one of the stable trajectories. The GB^{OBC}, GB^{HCT}, DrugScore (DS), and XSCORE scoring methods were each considered for each of the 40 geometries of each ligand.

[Figure 3 about here.]

The reported calculated ranks for each methods result from averaging over these 40 score values with or without a prior energy minimization of the complex with the CHARMM36–CGenFF force field combination, followed or not by a second minimization with the AMBER force field. For the two GBSA and the DS methods, the lower the score for a ligand, the better is its affinity; the opposite is true for the XSCORE method. Hence, a good correlation with experiment is expected (i) when a low negative correlation coefficient is calculated for the GBSA and DS methods and (ii) when a high positive correlation coefficient is calculated for the XSCORE method.

For the GB^{HCT} method performed on 40 geometries, no correlation is observed between the experimental and the calculated scores, as seen in Figures 4a–c for both minimization using one force field (CHARMM36–CGenFF or AMBER) and minimization using the force field combination. The same conclusion is valid for GB^{OBC} (Figures 4d–f). Hence, it is noteworthy that the performance of the two variants of the GBSA method is similar. In contrast, a good correlation is obtained for the DS and XSCORE results (Figures 4g,h), both in terms of energy correlation and in terms of ranking with four of the experimental top-5 ligands found in the calculated top-5 rank (ligands **2**, **8**, **5**, **1** and **8**, **1**, **3**, **2**, respectively).

[Figure 4 about here.]

[Figure 5 about here.]

In an aim to improving the calculated ligand ranking by reducing the bias due to an insufficient geometry sampling, 100 evenly spaced geometries were extracted and minimized (similarly to the above 40 geometries) from each of the ligand MD trajectory. The GB^{HCT}, XSCORE, and DS scoring methods were then applied (no GB^{OBC} calculations were considered here since, as noted, its performance is very similar to the other GBSA method). The results obtained for the GB^{HCT} on the 100 geometries (Figures 5a,b) show rather poor correlation with experimental results, both in terms of energy and rank. As previously, DS and XSCORE methods perform better than GBSA (Figures 5c,d). Though, as for the 40 geometries, only four of the experimental top-5 ligands are found in the calculated top-5 rank for both methods, the energy correlation improves slightly with 100 geometries for the XSCORE method. It should be stressed that, though the correlation coefficient is only slightly higher than in the case of the 40 geometries, the ranking is better: only one permutation (between ligands **1** and **5**) is observed and the ligand **7** has an affinity that is slightly overestimated.

The XSCORE results therefore provided the best results both for energy correlation and for ranking. However, the explicit solvation that is used lengthens the computation time. To reduce the computation time and thus to increase the number of ligands that could be processed in the case of Lu, we tested the Langevin molecular dynamics simulations on these complexes (implicit

model of solvation). In Figure 5e, one can see that the correlation coefficient is high but the ranking is worst than for an explicit solvation model (Figure 5d). This last scoring method was therefore not retained as a protocol for Lu.

3.2 Docking and scoring for Lu

For the virtual screening on the predicted binding region of the second N-terminal domain of Lu, we used the same protocol that led to the best results for CD80. The flowchart for the application of this protocol to Lu is shown in Figure 6. In our screening process, the nine inhibitors of CD80 were thus substituted by 395 601 compounds that were extracted from the ZINC database; 9 000 of these compounds were ranked on the basis of the primary scoring. The MATCH program [47] could derive CGenFF force field parameters for only 5 048 of the 9 000 compounds. These 5 048 ligands were thus retained for the subsequent steps.

[Figure 6 about here.]

The structure of Lu exhibits few water molecules with low B-factors. In particular, the B-factors of water 244, 257, and 264 are 20, 24, and 25 $Å^2$, respectively. What is more, the first two of them are involved in two hydrogen bonds with the residues of Lu (Figure 7): water 244 forms hydrogen bonds with Thr140 and Glu137; water 257 forms hydrogen bonds with Gln136 and Thr127, water 264, and water 414. Water 264 forms a hydrogen bond with waters 343 and 257 and the backbone nitrogen of Glu137 (PDB code 2PET). The combination of low B-factors and a high number of hydrogen bonds formed implies that these water molecules should not be easily substituted by a ligand without an enthalpy penalty, unless the ligand reproduces this network of hydrogen bonds. By analyzing the hydrogen bonds that the 5 048 ligands make with the two residues Gln136 and Glu137, only 34 ligands, whose atoms can reproduce the above hydrogen bond network, were retained. To increase the number of possible ligands of Lu, we also sought ligands that do not displace the waters 244, 257, and 264; this requires that any atoms of the ligand being more than 4 Å from the atoms O^{ϵ}, nitrogen backbone, and O^{γ} of the residues Gln136, Glu137, and Thr140, respectively. Following this filtering, 48 other ligands matched these criteria.

[Figure 7 about here.]

Before running the MD simulations, a further analysis of the protein-ligand complexes allowed to discard those that bury donor or acceptor groups. It is noteworthy that most rejected ligands at this stage buried the NH or CO backbone bonds of the residue Ala135. The burying of such groups may cause an unfavorable enthalpy balance during the binding of these ligands. On the other hand, the structures of four ligands selected at this stage were further modified to improve the correspondence between acceptor and donor groups of hydrogen bonds and the correspondence between the respective hydrophobic groups (for example, by adding a hydroxyl group on the ligand to match a protein hydrogen bond acceptor). Following this analysis and performing these few structural modifications, we ended with only 39 ligands that are more than 4 Å from Gln136, Glu137, and Thr140 and 16 other ligands that make contacts with Gln136 and Glu137 (Figure 6).

MD Simulations were then performed for the remaining 55 ligands. An analysis of all the MD trajectories allowed us to reject (i) ligands that are making only transient contacts with the protein (unstable), (ii) ligands whose rearrangement during the simulations causes some of its atoms to not perfectly reproduce the hydrogen bonds found with the water 244, 257, and 264 in

the crystallographic structure, and (iii) ligands whose rearrangement during the simulations leads to the burying of donor or acceptor groups. After the pruning of five more ligands in following the previous analysis, for each of the remaining 50 ligands, 100 geometries were extracted from the respective trajectory and finally minimized in the presence of water. For each ligand retained, the scoring by XSCORE was evaluated for the 100 geometries and then averaged to derive the final score reported in Table 2 for some of the top hits.

[Table 2 about here.]

The ligands that do not form hydrogen bonds with residues Gln136, Glu137, and Thr140 form the first category and will be referred to as $1x$ where x is a letter that is related to the rank calculated (Table 2). In this category, only the top 10 scored ligands were analyzed. Several of the first ten ligands make at least two hydrogen bonds with residues Thr188, Thr190, Thr174, Met167, Tyr192, and Tyr172 or with the backbone of residue Ser175. These contacts are supplemented by van der Waals interactions with Met167, Thr188, Thr140, Tyr192, Thr190, Arg176, Val165, Glu166, and Tyr172 for most ligands (Figure 8). The following ligands with special features will be described hereafter, namely the ligands $1a$, $1b$, $1c$, $1e$, $1f$, and $1i$. The 2D representations of the top 10 ligands in this first category as well as those for two ligands that interact with Gln136 and Glu137 are given in Figure 9.

[Figure 8 about here.]

[Figure 9 about here.]

Ligand $1a$ forms three hydrogen bonds with residues Thr174, Thr190, and Met167 in addition to van der Waals interactions with residues Met167, Thr188, and Thr140 (Figure 8a). Through their respective negative groups COO^- and SO_3^-, the ligands $1b$ and $1c$ form three ionic hydrogen bonds with residues Thr188, Thr190, and Thr174 (Figure 8b,c). The ligand $1b$ makes two additional hydrogen bonds with Tyr172 and Tyr192. This last hydrogen bond results from our decision to add an hydroxyl group to the non-aromatic cycle of this ligand (Figure 9b). Structural modifications brought to the three other ligands (data not shown) did not lead to other top 10 scores. The ligand $1c$ makes one additional hydrogen bond with Arg176; however, apart from its sulfonyl group, the ligand barely interacts with the protein. Therefore, this ligand is expected to be of low specificity toward Lu. The ligand $1e$ form two permanent hydrogen bonds with Ser175 while a transient one is formed with Thr174 (solid and dashed lines, respectively, in Figure 8e). Van der Waals contacts are formed with residues 165 to 167 and 176. The ligand $1g$ forms also a labile hydrogen bond with Thr190, as do the ligand $1j$ (shown as dashed lines in Figure 8g,j). In the case of last three ligands $1e$, $1g$, and $1j$, at least one of the two outermost aromatic rings is highly flexible, having many alternative positions during the MD simulations, leading to transient van der Waals contacts with their protein residue partners (see also Figure 9e,g,j). The ligand $1f$ is highly hydrophobic and highly flexible with 15 rotatable bonds; it forms many intra- and intermolecular van der Waals contacts. This ligand makes two hydrogen bonds with residues Thr174 and Met167 despite a low proportion of heteroatoms.

The ligands that form hydrogen bonds with residues Gln136, Glu137, and Thr140 form the second category and will be referred to as $2x$ where x is a letter (Table 2). Among the ligands in this category, a good correspondence between respective hydrophilic and hydrophobic groups of the protein and ligand is observed only for the ligand $2k$ (Figure 8k), which is also ranked first

in this category. Its high score can be related (i) to six hydrogen bonds formed with the residues Tyr172, Tyr192, Thr127, Gln136, Glu137, and Ala135 and (ii) to van der Waals interactions with the residues Met167, Thr127, Tyr192, and Glu132 (Figure 8k). Hence, since this ligand interacts with Glu132, at least part of its structure is expected to lie close to the binding region of Lnα5. However, the backbone oxygen atom of Ala135 is a hydrogen bond acceptor that remain buried in all the frames of the MD trajectory (oxygen designated by a yellow arrow in Figure 8k). The ligand $2l$, which has the rank 6 in this second category, binds to Lu with four hydrogen bonds that are formed with residues Thr127, Thr174, Thr190, and Glu137. Though no steric clash is observed between this ligand and water 257 in the docking results, one can see from the MD simulations that the simultaneous binding of this ligand and this water is not possible given the fact that the respective van der Waals spheres are overlapping in one snapshot taken from the trajectory (Figure 8l). Other ligands of this second category were discarded because, as seen from the respective MD trajectories, they do not form hydrogen bond with Gln136. For example, during the MD simulations, the hydrogen donor of most of the other ligands is reoriented and makes a contact with the carbonyl oxygen of the backbone of Thr127 (which is close to Gln136). Few other ligands were initially retained because some of their atoms perfectly replace the water 264 and form a hydrogen bond with the backbone bond NH of Glu137 while the water 257 is maintained. However, the analysis of their MD trajectories reveals a steric hindrance between the ligand and water 257, thereby compromising the hydrogen bond this water forms with Gln136 and Thr127. These ligands were thus also discarded.

It should also be mentioned that, when the scores of the two categories of ligands are combined, the ligand $2k$ is found at the seventh rank while the ligand $2l$ is found only at rank 29.

The number of hydrogen bonds observed between the ligands and the protein varies between two and six hydrogen bonds. However, this number of hydrogen bonds is not correlated with the rank calculated for the ligand. For example, the best scoring ligand $1a$ makes three hydrogen bonds while the ligand $1i$, which has one of the lowest scores, makes four hydrogen bonds. However, we observed that the presence of, at least, a few hydrogen bonds is important as they contribute to increasing the stability of the ligand during the MD simulations. These hydrogen bonds act as anchoring interactions that seem crucial for stabilizing the ligand on such rather flat interacting surface. In the second category, the ligand $2k$, which has an overall rank of 7, forms six hydrogen bonds with the protein; this represents the highest number of hydrogen bonds that is observed among all of the ligands. It is noteworthy that the ligand $2k$ perfectly replaces all the three hydrogen bonds that the water molecules 257 and 264 form with the residues Thr127, Gln136, and Glu137 in the crystallographic apo structure. Despite this exact replacement and the fact that the ligand $2k$ forms the highest number of hydrogen bonds, the calculated score for $2k$ is lower than those of the first six ranked ligands, $1a$ up to $1f$, that form fewer hydrogen bonds (between two and five). These previous observations lead to the following three conclusions: (i) the replacement of crystallographic water molecules with low B-factors by a ligand does not ensure better scores; (ii) the rank of the ligand is not correlated with the number of hydrogen bonds formed with the protein; and (iii) a few hydrogen bonds are required for the stability of the ligand during the simulations, probably due to the uncommon flat binding surface of the protein in our case.

Besides, it is difficult to speculate about the intriguing cooperativity effect since our best scoring function is empirical and, thus, allows the description of additive effects only.

4 Discussion

Lipinski has defined a set of rules (rule of five) for estimating the permeation and the absorption of a compound from its two-dimensional structure (2D) [50]. Such analyses of the structures of orally administered drugs, and of drug candidates, has so far been the primary guide to correlate physical properties with successful drug development [50, 51]. These analyses have proved to be very useful and have led to a set of rules relating the importance of lipophilicity (Log P should be less than 5), molecular weight (less than 500 Da), and the number of hydrogen bond donors (HBD) and acceptors (HBA) [50, 51]. HBD and HBA should be less than 5 and 10, respectively. In addition to these rules, Veber has introduced two additional criteria. According to the study of 1100 drug candidates, the polar surface area (PSA) of the compound should be less than 140 $Å^2$ and the number of rotatable bonds must be less than 10 for a good oral bioavailability [52]. These criteria are now widely used at a very early stage in the process of drug discovery.

Though the 500-Da rule is not a hard limit, the value of Log P should never be larger than 5 since the solubility of a compound generally decreases about ten fold when Log P increases by one unit. Hence the ligands $1a$ and $1f$ should not be considered for experimental tests (Table 2).

Since the residues Glu132/Asp133 have been found to be important for Lnα5 binding, useful criteria for selecting molecules that can hinder this PPI is the proximity with these residues. Moreover, in the study of Kikkawa *et al.* [10], an antibody whose epitope was found on the second N-terminal domain of Lu is thought to bind, at least in part, to the binding region of Lu defined in Figure 1. Hence, from these experimental data, we have inferred a useful guide to further select compounds for testing: these compounds should be geometrically as close as possible to Glu132 and/or Asp133 and should have a few atoms that protrude into the solvent.

As seen in Table 2, the ligands that are closest to Glu132 are $2l$, $2k$, $1b$, $1h$, $1d$, $1i$, and $1c$. Among these ligands, only the ligands $2l$, $2k$, $1b$, $1d$, and $1i$ exhibit a region that protrude into the solvent (Figure 10). Though the ligand $2l$ cumulates interesting structural and physico-chemical properties, it is expected to have a low affinity toward Lu, given its high overall rank, 29 (section 3.3). Moreover, the ligands $2k$ and $1b$ are not commercially available. Since the ligand $1b$ is the second best ligand, experimental binding tests would be of interest but would require prior synthesis. It should thus be mentioned that its synthesis may start from the commercially available analog where the hydroxyl group on the outermost non-aromatic cycle is missing (Figure 9b).

In contrast, the ligands $1d$ and $1i$ are among the top 10 ligands, both exhibit structural and physico-chemical properties that are expected for drug candidates and are commercially available.

[Figure 10 about here.]

5 Conclusion

To design inhibitors of the Lu-Lnα5 interactions, a virtual screening was performed targeting the second N-terminal domain of Lu. We first searched a putative binding site on Lu by comparative and predictive techniques in which both the sequence and the structure of the targeted domain of Lu were considered. We took advantage of the similarities between the identified binding site on Lu and that of another protein, CD80, to set up a robust docking and scoring protocol. This validation step has revealed poor performance of the two MM-GBSA methods that were tested based on atomic charges derived from CGenFF. In contrast, a good correlation with experiments could be

obtained for XSCORE. We also showed that both using an explicit (*vs.* an implicit) solvation model and increasing the number of snapshots taken from the MD simulations contribute to improving the ranking of the ligands. This protocol was then applied to Lu for the screening of 395 601 compounds extracted from the ZINC database. Our protocol for primary scoring filtering, molecular dynamics simulation filtering, secondary scoring filtering, and molecular property filtering allows discarding most of the ligands, with only 12 compounds retained. Four of these compounds are promising candidates for inhibiting this PPI interaction; two of them being commercially available and, thus, readily tested while an other needs only minor structural modification starting from an available analog.

List of Abbreviations

BCAM, Basal Cell Adhesion Molecule; BLAST, Basic Local Alignment Search Tool; CD80, Cluster of Differentiation 80; CGenFF, CHARMM General Force Field; CTLA-4, Cytotoxic T-Lymphocyte-Associated Protein 4; DS, DrugScore; HBD, Hydrogen Bond Donors; HBA, Hydrogen Bond Acceptors; ICAM, InterCellular Adhesion Molecule; IC50, half maximal inhibitory concentration; LD, Langevin molecular Dynamics; Lu, Lutheran; Ln, Laminin; MATCH, Multipurpose Atom-Typer for CHARMM; MD, Molecular Dynamics; GB, Generalized Born; GBSA, Generalized Born Surface Area; MM-GBSA, Molecular Mechanics–Generalized Born Surface Area; PDB, Protein Data Bank; Phyre2, Protein Homology/analogy Recognition Engine 2; PSA, Polar Surface Area; POSSUM, Pocket Similarity Search using Multiple-sketches; PPI, Protein–Protein Interaction; SBD, Stochastic Boundary molecular Dynamics; VAST, Vector Alignment Search Tool;

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

N. Madeleine was the recipient of a doctoral fellowship from the Région Réunion. This study was supported by grants from Laboratory of Excellence GR-Ex, reference ANR-11-LABX-0051. The labex GR-Ex is funded by the program 'Investissements d'avenir' of the French National Research Agency, reference ANR-11-IDEX-0005-02. All calculations were performed on the 'Centre de Calcul de l'Université de la Réunion' (CCUR). NM and FG performed study, collected data, analyzed data, and wrote paper; FG designed research.

References

[1] Connes P, Hue O, Tripette J, Hardy-Dessources MD (2008) Blood rheology abnormalities and vascular cell adhesion mechanisms in sickle cell trait carriers during exercise. Clin Hemor Microcirc 39: 179-184.

[2] Kaul DK, Finnegan E, Barabino GA (2009) Sickle Red Cell-Endothelium Interactions. Microcir 16: 97-111.

[3] Colin Y, Le Van Kim C, El Nemer W (2014) Red cell adhesion in human diseases. Curr Opin Hemat 21: 186-192.

[4] Rahuel C, Kim CL, Mattei MG, Cartron JP, Colin Y (1996) A unique gene encodes spliceoforms of the B-cell adhesion molecule cell surface glycoprotein of epithelial cancer and of the Lutheran blood group glycoprotein. Blood 88: 1865-1872.

[5] El Nemer W, Gane P, Colin Y, Bony V, Rahuel C, et al. (1998) The Lutheran blood group glycoproteins, the erythroid receptors for laminin, are adhesion molecules. J Biol Chem 273: 16686-16693.

[6] Udani M, Zen Q, Cottman M, Leonard N, Jefferson S, et al. (1998) Basal cell adhesion molecule Lutheran protein - The receptor critical for sickle cell adhesion to laminin . J Clin Invest 101: 2550–2558.

[7] Spenle C, Simon-Assmann P, Orend G, Miner JH (2013) Laminin alpha 5 guides tissue patterning and organogenesis. Cell Adhes Migr 7: 90-100.

[8] Bartolucci P, Chaar V, Picot J, Bachir D, Habibi A, et al. (2010) Decreased sickle red blood cell adhesion to laminin by hydroxyurea is associated with inhibition of Lu/BCAM protein phosphorylation. Blood 116: 2152-2159.

[9] Kikkawa Y, Ogawa T, Sudo R, Yamada Y, Katagiri F, et al. (2013) The Lutheran/Basal Cell Adhesion Molecule Promotes Tumor Cell Migration by Modulating Integrin-mediated Cell Attachment to Laminin-511 Protein. J Biol Chem 288: 30990-31001.

[10] Kikkawa Y, Miwa T, Tohara Y, Hamakubo T, Nomizu M (2011) An Antibody to the Lutheran Glycoprotein (Lu) Recognizing the LU4 Blood Type Variant Inhibits Cell Adhesion to Laminin alpha 5. PLOS One 6: e23329-e23339.

[11] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: A free tool to discover chemistry for biology. J Chem Inf Model 52: 1757-1768.

[12] Irwin JJ, Shoichet BK (2005) Zinc – a free database of commercially available compounds for virtual screening. J Chem Inf Model 45: 177–182.

[13] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.

[14] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, et al. (2002) The Protein Data Bank. Acta Cryst Sect D: Biol Cryst 58: 899–907.

[15] Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The Protein Data Bank and structural genomics. Nucleic Acids Res 31: 489–491.

[16] McWilliam H, Valentin F, Goujon M, Li W, Narayanasamy M, et al. (2009) Web services at the european bioinformatics institute-2009. Nucleic Acids Res 37: W6-W10.

[17] Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6: 377-385.

[18] Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, et al. (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res 42: D297-303.

[19] Holm L, Rosenström P (2010) DALI server: conservation mapping in 3d. Nucleic Acids Res 38: W545-549.

[20] Ito JI, Tabei Y, Shimizu K, Tsuda K, Tomii K (2012) PoSSuM: a database of similar protein-ligand binding and putative pockets. Nucleic Acids Res 40: D541-D548.

[21] Akira RK, Haruki N (2007) Similarity search for local protein structures at atomic resolution by exploiting a database management system. Biophysics 3: 75-84.

[22] Kelley LA, Sternberg MJE (2009) Protein structure prediction on the WEB: a case study using the Phyre server. Nature Prot 4: 363-371.

[23] Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinf 21: 1908-1916.

[24] Brylinski M, Feinstein WP (2013) eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. J Comput-Aid Mol Des 27: 551-567.

[25] Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, et al. (2012) FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinf 28: 286-287.

[26] Hernandez M, Ghersi D, Sanchez R (2009) SITEHOUND-WEB: a server for ligand binding site identification in protein structures. Nucleic Acids Res 37: W413-W416.

[27] Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, et al. (2009) Atomic interactions and profile of small molecules disrupting proteinprotein interfaces: the TIMBAL database. Chem Biol Drug Des 74: 457-467.

[28] Erbe DV, Wang S, Xing Y, Tobin JF (2001) Small molecule ligands define a binding site on the immune regulatory protein B7.1. J Biol Chem 277: 7363-7368.

[29] Laskowski RA (1995) SURFNET - a program for visualizing molecular-surfaces, cavities, and intermolecular interactions. J Mol Graph 13: 323-330.

[30] Stamper C, Zhang Y, Tobin J, Erbe D, Ikemizu S, et al. (2001) Crystal structure of the B7-1/CTLA-4 complex that inhibits human immune responses. Nature 410: 608–611.

[31] De Azevedo J, Walter F (2010) MolDock applied to structure-based virtual screening. Curr drug targets 11: 327–334.

[32] Green NJ, Xiang J, Chen J, Chen L, Davies AM, et al. (2003) Structure-activity studies of a series of dipyrazolo[3,4-b:3',4'-d]pyridin-3-ones binding to the immune regulatory protein B7.1. Bioorg Med Chem 11: 2991-3013.

[33] Mankelow T, Burton N, Stefansdottir F, Spring FA, Parsons SF, et al. (2007) The laminin 511/521-binding site on the lutheran blood group glycoprotein is located at the flexible junction of Ig domains 2 and 3. Blood 110: 3398-3406.

[34] Ikemizu S, Gilbert RJ, Fennelly JA, Collins AV, Harlos K, et al. (2000) Structure and dimerization of a soluble form of B7-1. Immunity 12: 51–60.

[35] Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK - a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 47-60.

[36] Vriend G (1990) WHAT IF: a molecular modelling and drug design program. J Mol Graph 8: 52-56.

[37] Brooks BR, Brooks III CL, Mackerell Jr AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: The biomolecular simulation program. J Comput Chem 30: 1545–1614.

[38] Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, et al. (2009) DOCK 6: Combining techniques to model RNA-small molecule complexes. RNA 15: 1219-1230.

[39] Tsui V, Case DA (2000) Theory and applications of the generalized born solvation model in macromolecular simulations. Biopolymers 56: 275–291.

[40] Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Structure, Function, and Bioinformatics 55: 383–394.

[41] Wang RX, Lai LH, Wang SM (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput-Aid Mol Des 16: 11–26.

[42] Neudert G, Klebe G (2011) DSX: A knowledge-based scoring function for the assessment of proteinligand complexes. J Chem Inf Model 51: 2731-2745.

[43] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, et al. (2010) Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. J Comput Chem 31: 671–690.

[44] Lemons, D S and Gythiel, A (1997) Paul langevin's 1908 paper 'On the theory of brownian motion'. Am J Phys 65: 1079-1081.

[45] Brünger A, Brooks III CL, Karplus M (1984) Stochastic boundary conditions for molecular dynamics simulations of ST2 water. Chem Phys Lett 105: 495-500.

[46] Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23: 327-341.

[47] Yesselman JD, Price DJ, Knight JL, Brooks CL (2012) MATCH: An atom-typing toolset for molecular mechanics force fields. J Comput Chem 33: 189–202.

[48] Boström J, Hogner A, Schmitt S (2006) Do structurally similar ligands bind in a similar fashion? J Med Chem 49: 6716–6725.

[49] Hou T, Wang J, Li Y, , Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods: II. the accuracy of ranking poses generated from docking. J Comput Chem 32: 866-877.

[50] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 64: 4-17.

[51] Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxic Meth 44: 235-249.

[52] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, et al. (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45: 2615–2623.

List of Figures

1 Experimental structure of the two N-terminal domains of Lu. The binding site predicted is delineated by the residues colored gold and by the residue Glu132. The residues Glu132, Asp133, Asp198, and Asp199 whose mutation decreases the binding of Ln$\alpha 5$ are colored green [33]. 20

2 Comparison between the structure of CD80 (a) and that of the second N-terminal domain of Lu (b) reveals overall similarity between the respective predicted binding sites. Moreover, the color code provided on the top right, which is valid only for parts (a) and (b), also reveals several similarities in terms of physico-chemical properties of the residues when comparing the various residues reported in parts (a) and (b). In parts (c–e), the structures of CD80 (backbone in green) and Lu (in khaki) are superimposed: in (c), the negative residues are displayed in red; in (d), the positive residues are in cyan; in (e), the hydrophobic residues are in blue. In parts (c–e), the center of the binding region of each protein is circled in purple. . . 21

3 Snapshots for the complex formed between the V-set domain of CD80 and the ligand **8**. 22

4 Results for the secondary scoring obtained for 40 geometries that were extracted from the simulations of each ligand and minimized before applying the GB^{HCT} ((a), (b), and (c)), GB^{OBC} ((d), (e), and (f)), DS, and XSCORE scoring methods. Whether the minimization of the geometries were performed with CHARMM36–CGenFF and/or AMBER force fields is indicated in the figure legends along with the number of geometries considered for each ligand. 23

5 Results for the secondary scoring obtained for 100 geometries that were extracted from the simulations of each ligand and minimized before applying the GB^{HCT} ((a) and (b)), DS, and XSCORE scoring methods. Whether the minimization of the geometries were performed with CHARMM36–CGenFF and/or AMBER force fields is indicated in the figure legends along with the number of geometries considered for each ligand. All results were obtained for the geometries extracted from the simulations with an explicit solvation, except for (e). 24

6 Flowchart that summarizes the main steps of our docking and scoring protocol applied to Lu. Compounds without contact with residues Gln136, Glu137, and Thr140 are more than 4 Å from these residues. 25

7 The crystallographic structure of Lu is shown along with the water molecules 244, 257, and 264 associated with B-factor values lower than 25 Å^2 (in cyan). The interacting water partner 343 and 414 are also shown in magenta. 26

8 Snapshots from the MD trajectories of the ligands $1a$–j and $2k, l$, colored gold. Permanent and transient hydrogen bonds are represented by plain and dashed thick green lines, respectively. For the protein, all but the hydrogen atoms on the heteroatoms were hidden for clarity purpose. 27

9 2D representation of the top 10 ligands found for the first category and the two ligands *2k* and *2l*. The respective labels a to l were used. 28

10 Snapshots of the complexes for the ligands $1a$–j and $2k, l$. Surface representation is used for the protein and stick representation for the ligand. The residue Glu132 is colored in light green. The region of the ligand into the solvent is encircled. The respective labels a to l were used. 29

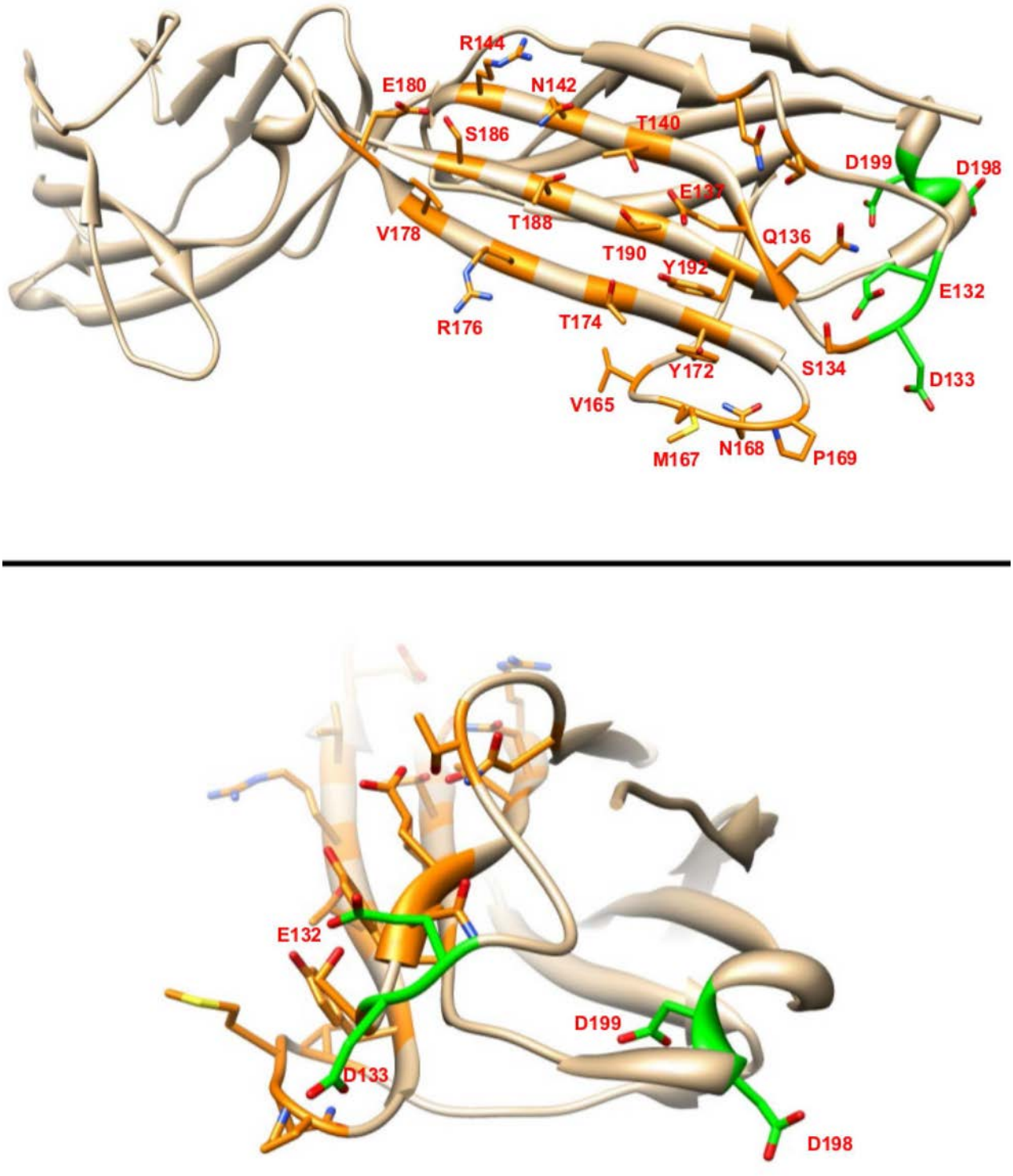


Figure 1: Experimental structure of the two N-terminal domains of Lu. The binding site predicted is delineated by the residues colored gold and by the residue Glu132. The residues Glu132, Asp133, Asp198, and Asp199 whose mutation decreases the binding of Ln$\alpha 5$ are colored green [33].

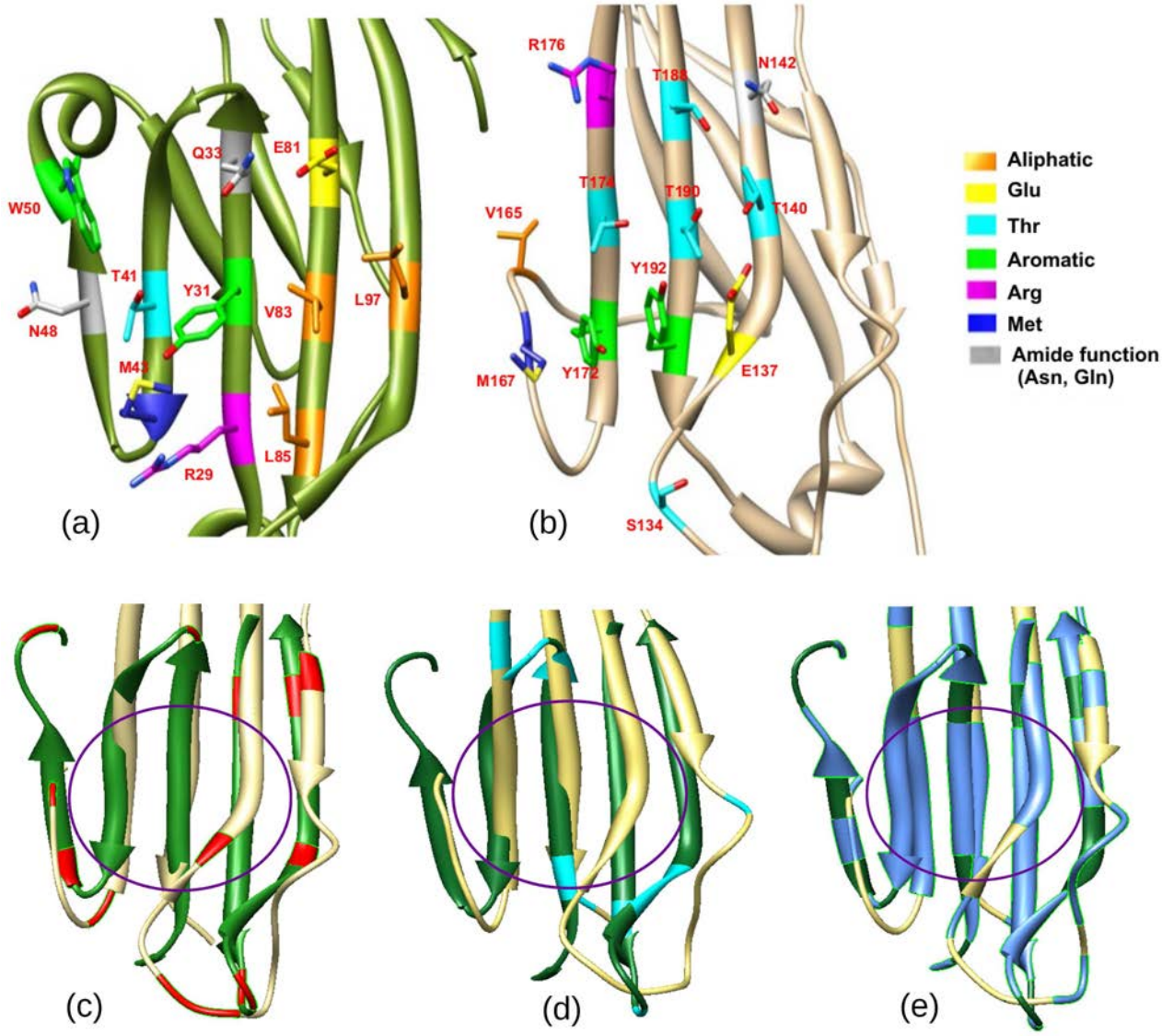


Figure 2: Comparison between the structure of CD80 (a) and that of the second N-terminal domain of Lu (b) reveals overall similarity between the respective predicted binding sites. Moreover, the color code provided on the top right, which is valid only for parts (a) and (b), also reveals several similarities in terms of physico-chemical properties of the residues when comparing the various residues reported in parts (a) and (b). In parts (c–e), the structures of CD80 (backbone in green) and Lu (in khaki) are superimposed: in (c), the negative residues are displayed in red; in (d), the positive residues are in cyan; in (e), the hydrophobic residues are in blue. In parts (c–e), the center of the binding region of each protein is circled in purple.

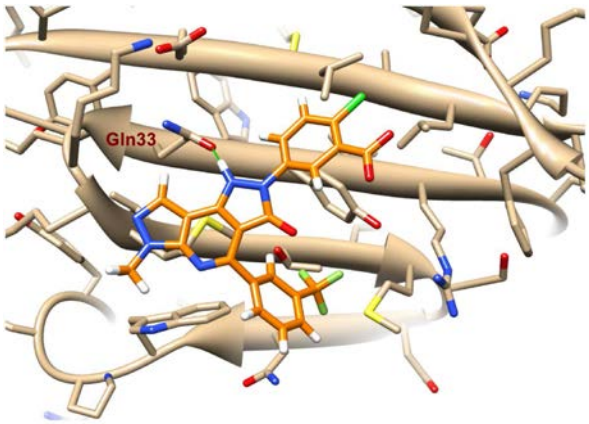
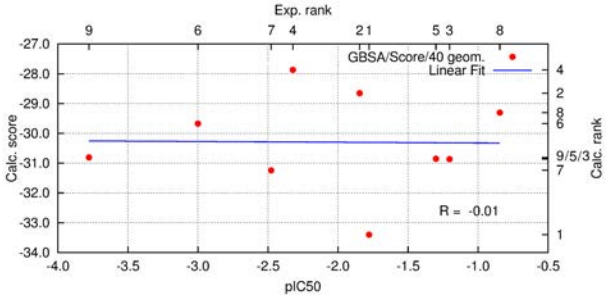
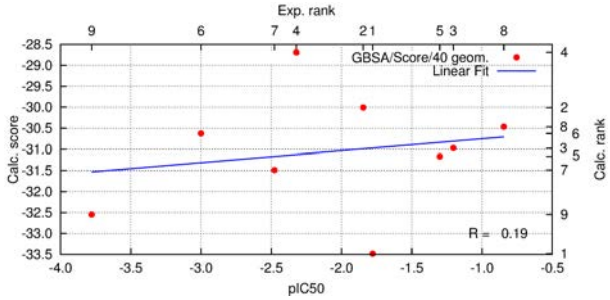


Figure 3: Snapshots for the complex formed between the V-set domain of CD80 and the ligand **8**.

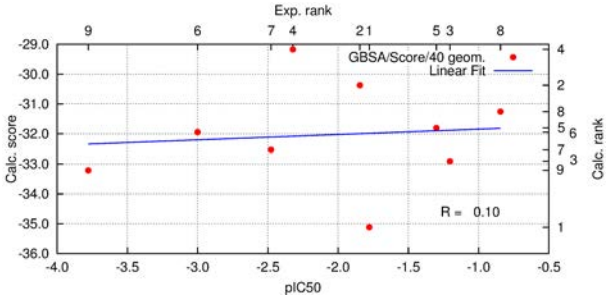
Figure 4: Results for the secondary scoring obtained for 40 geometries that were extracted from the simulations of each ligand and minimized before applying the GB^{HCT} ((a), (b), and (c)), GB^{OBC} ((d), (e), and (f)), DS, and XSCORE scoring methods. Whether the minimization of the geometries were performed with CHARMM36–CGenFF and/or AMBER force fields is indicated in the figure legends along with the number of geometries considered for each ligand.



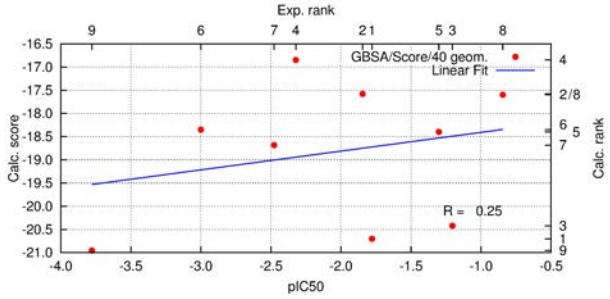
(a) 40 geom./CHARMM36–CGenFF



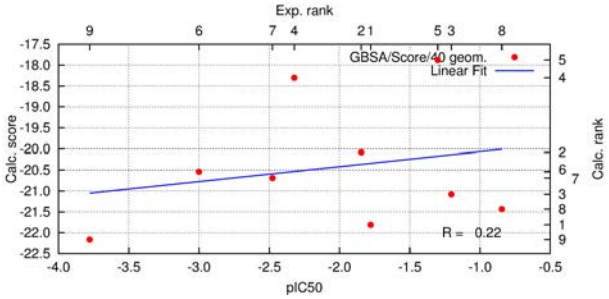
(b) 40 geom./AMBER



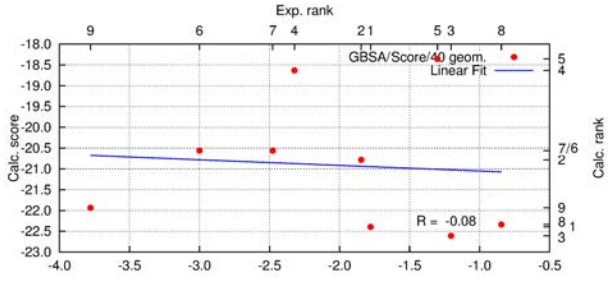
(c) 40 geom./CHARMM36–CGenFF and AMBER



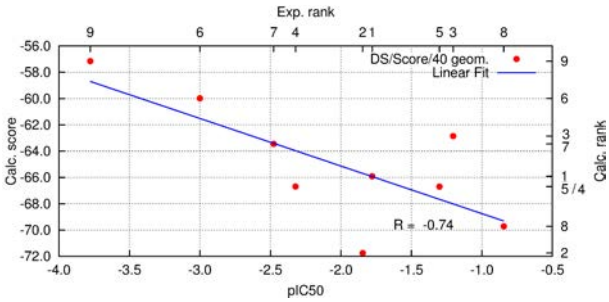
(d) 40 geom./CHARMM36–CGenFF



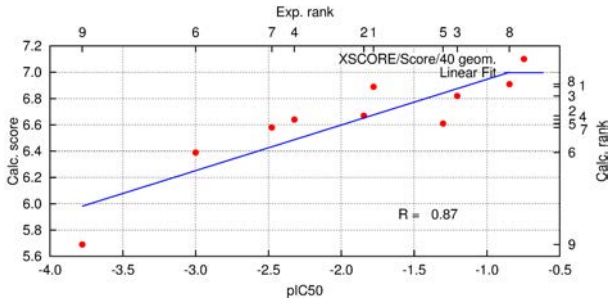
(e) 40 geom./AMBER



(f) 40 geom./CHARMM36–CGenFF and AMBER

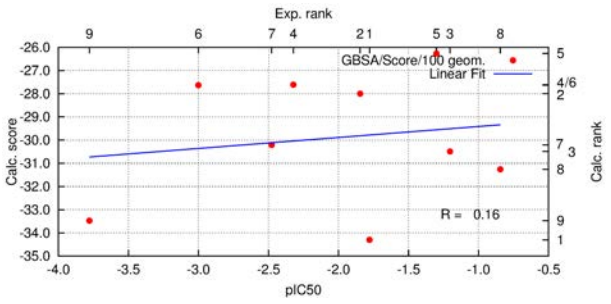


(g) 40 geom./CHARMM36–CGenFF

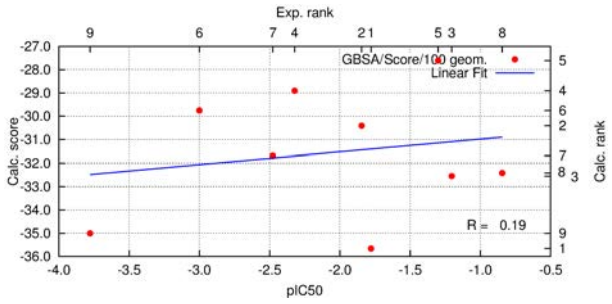


(h) 40 geom./CHARMM36–CGenFF

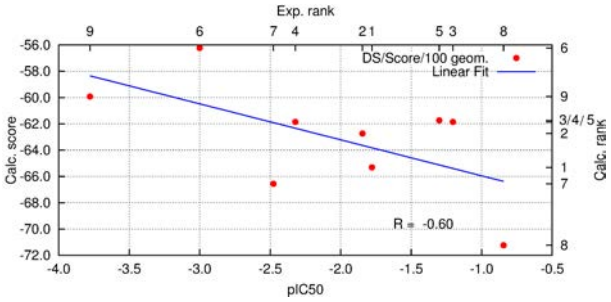
Figure 5: Results for the secondary scoring obtained for 100 geometries that were extracted from the simulations of each ligand and minimized before applying the GB^{HCT} ((a) and (b)), DS, and XSCORE scoring methods. Whether the minimization of the geometries were performed with CHARMM36–CGenFF and/or AMBER force fields is indicated in the figure legends along with the number of geometries considered for each ligand. All results were obtained for the geometries extracted from the simulations with an explicit solvation, except for (e).



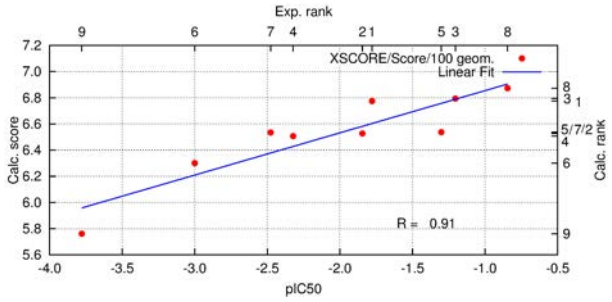
(a) 100 geom./CHARMM36–CGenFF



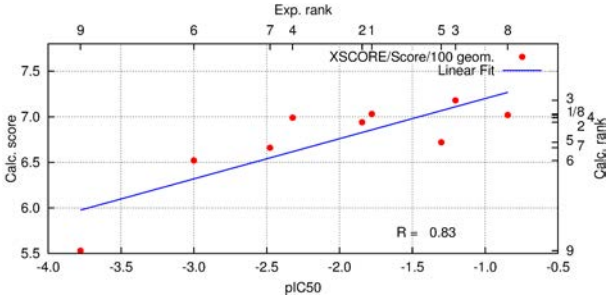
(b) 100 geom./CHARMM36–CGenFF and AMBER



(c) 100 geom./CHARMM36–CGenFF



(d) 100 geom./CHARMM36–CGenFF



(e) 100 geom./CHARMM36–CGenFF and Langevin

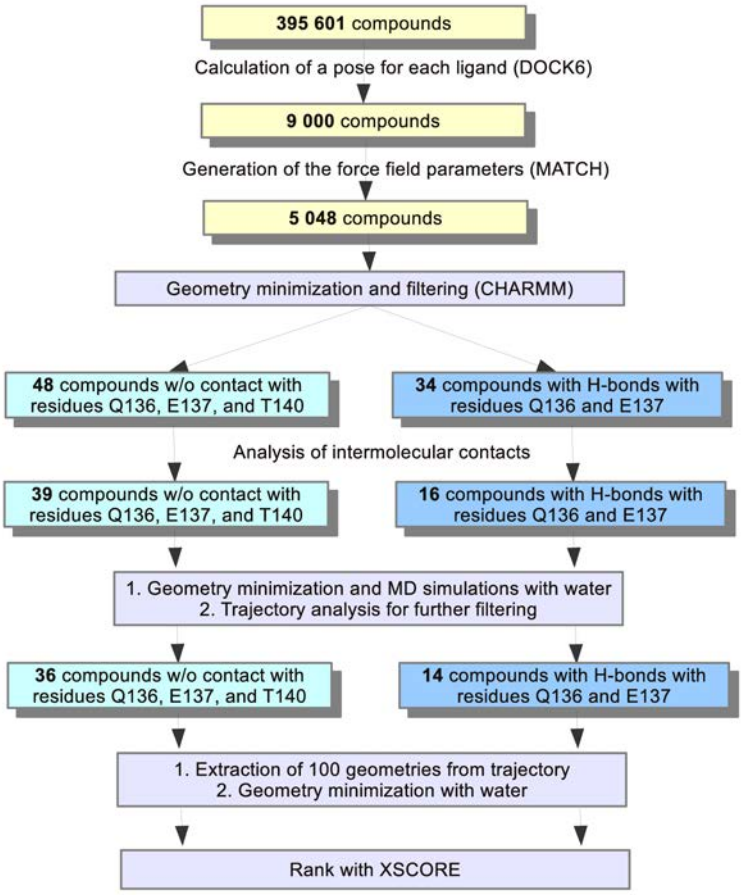


Figure 6: Flowchart that summarizes the main steps of our docking and scoring protocol applied to Lu. Compounds without contact with residues Gln136, Glu137, and Thr140 are more than 4 Å from these residues.

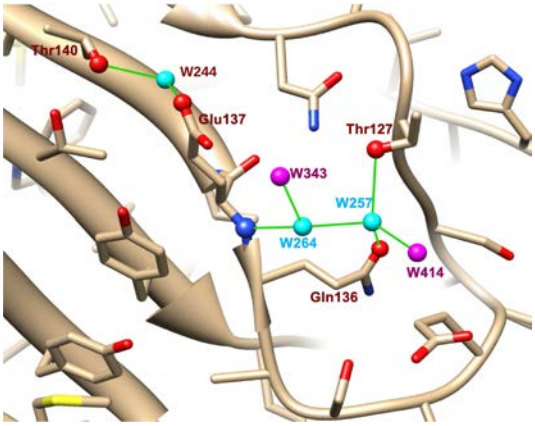


Figure 7: The crystallographic structure of Lu is shown along with the water molecules 244, 257, and 264 associated with B-factor values lower than 25 $Å^2$ (in cyan). The interacting water partner 343 and 414 are also shown in magenta.

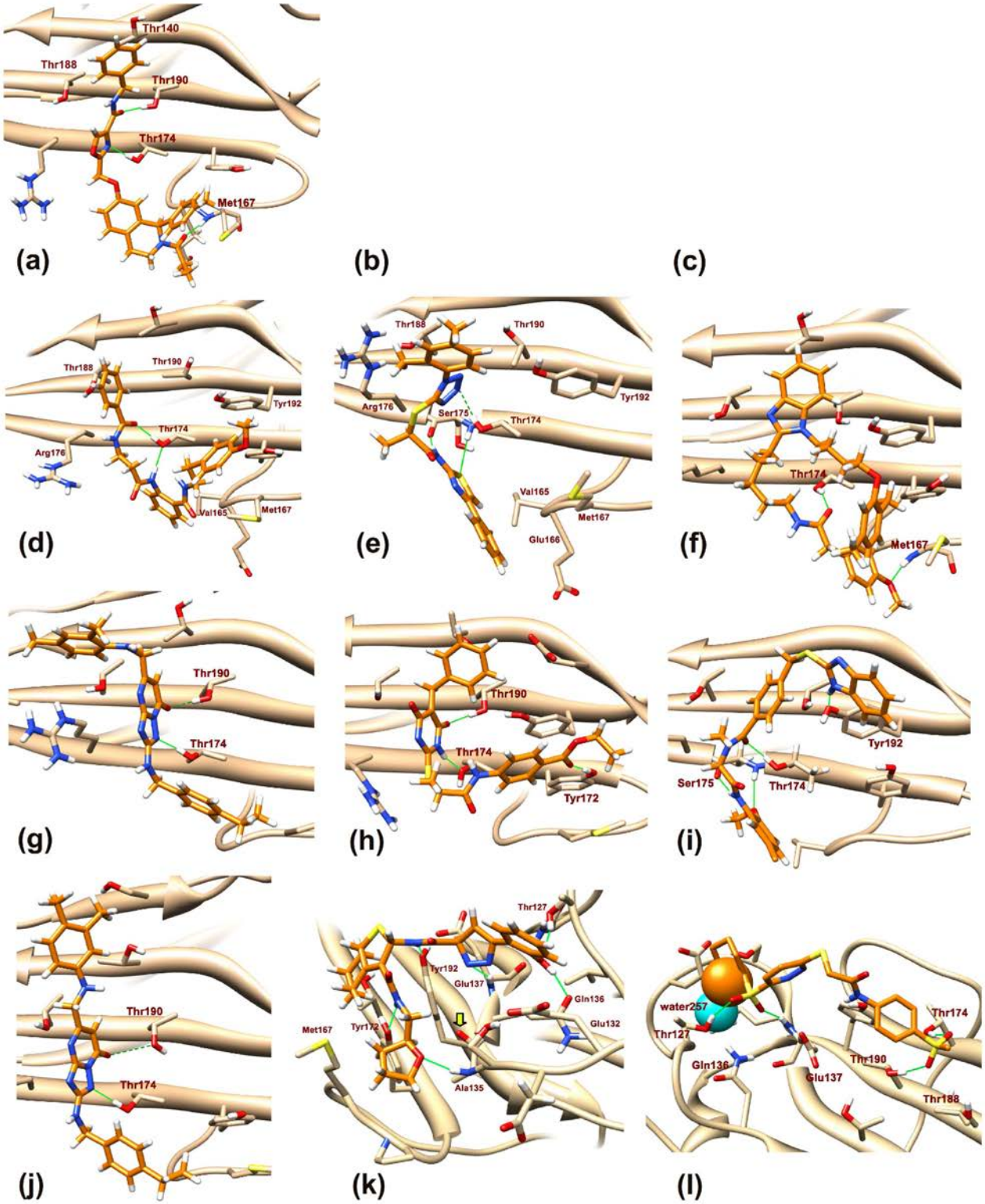


Figure 8: Snapshots from the MD trajectories of the ligands $1a$–j and $2k, l$, colored gold. Permanent and transient hydrogen bonds are represented by plain and dashed thick green lines, respectively. For the protein, all but the hydrogen atoms on the heteroatoms were hidden for clarity purpose.

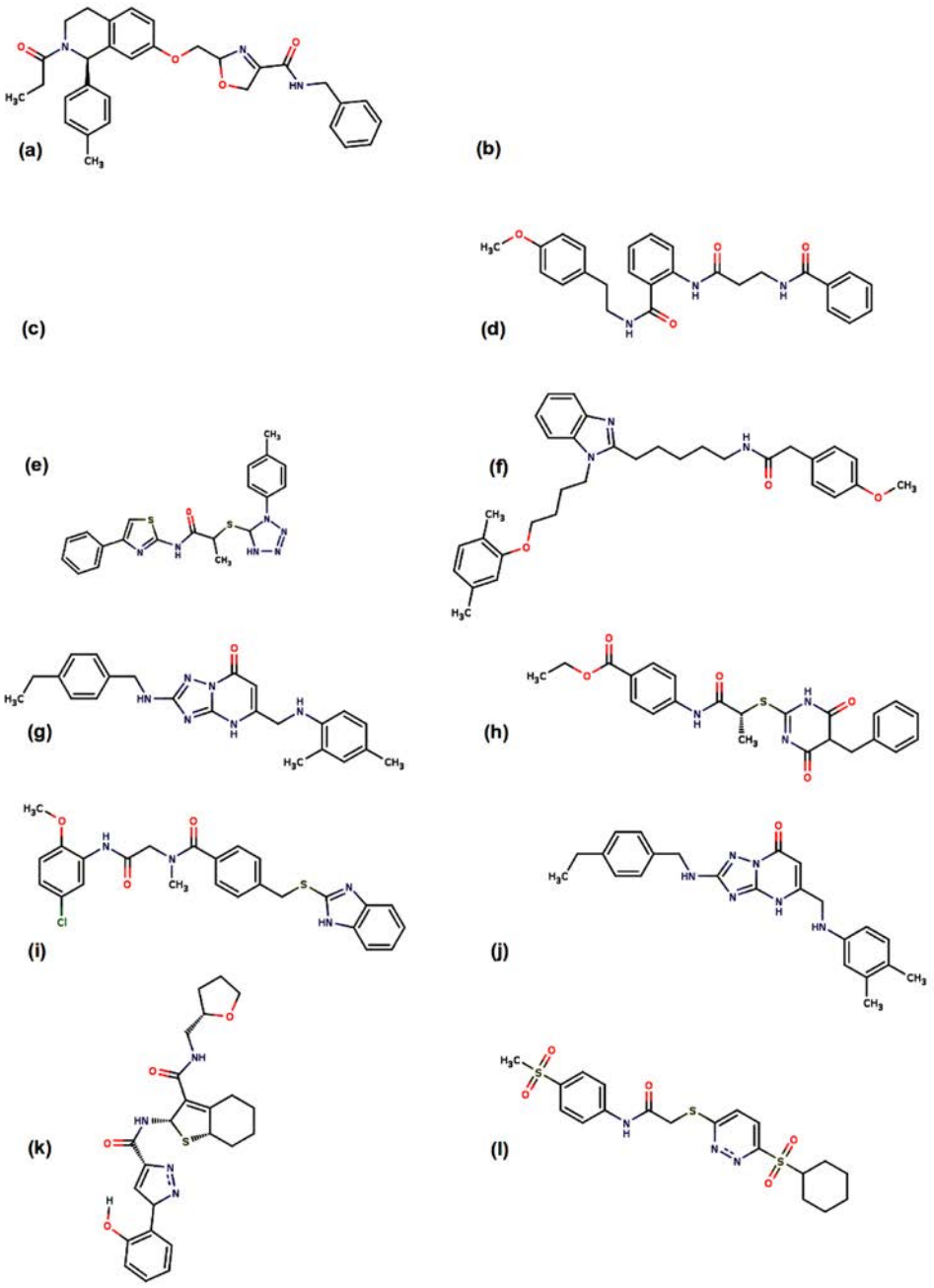


Figure 9: 2D representation of the top 10 ligands found for the first category and the two ligands *2k* and *2l*. The respective labels a to l were used.

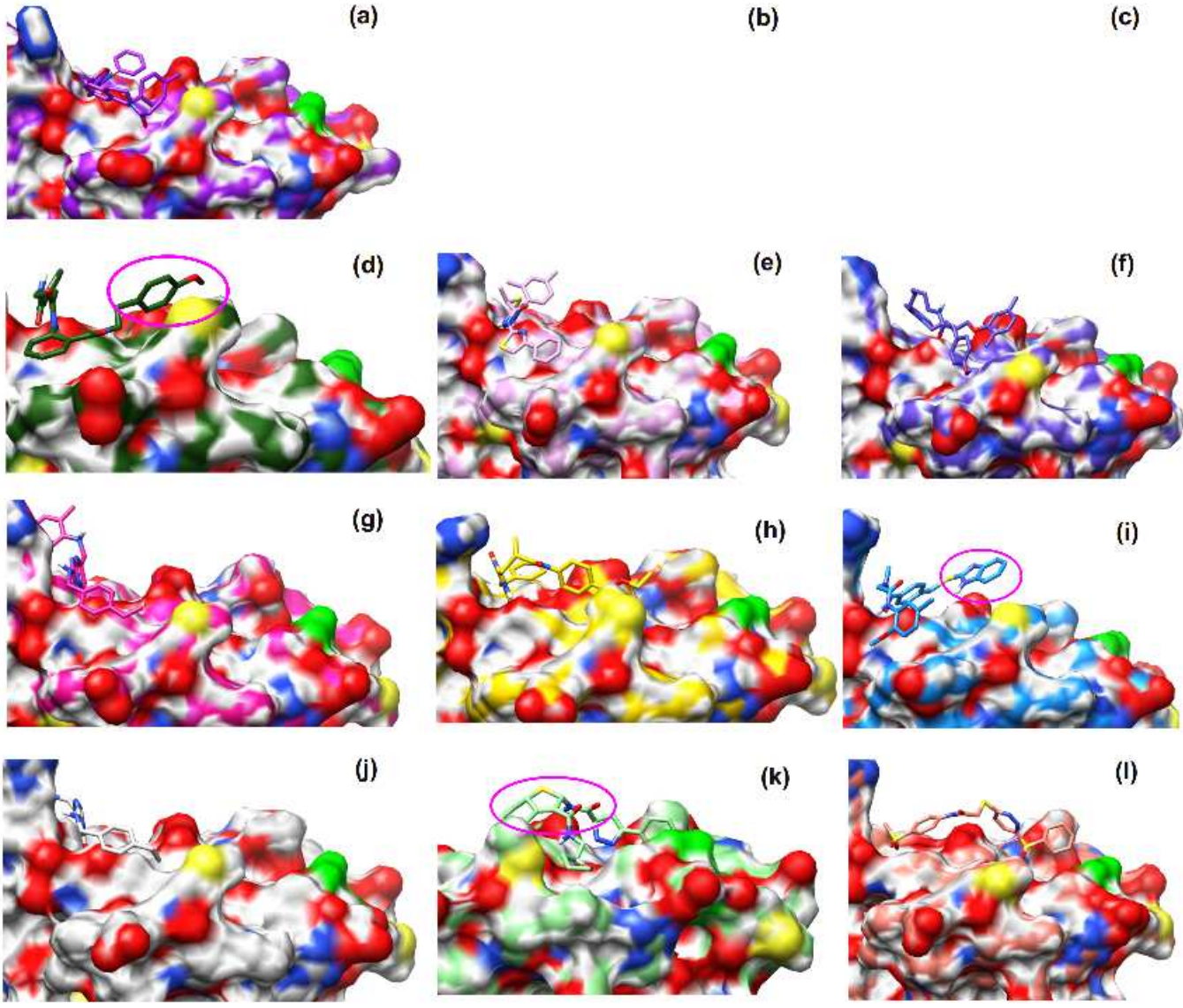
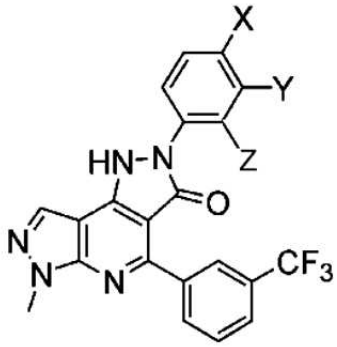


Figure 10: Snapshots of the complexes for the ligands $1a$–j and $2k, l$. Surface representation is used for the protein and stick representation for the ligand. The residue Glu132 is colored in light green. The region of the ligand into the solvent is encircled. The respective labels a to l were used.

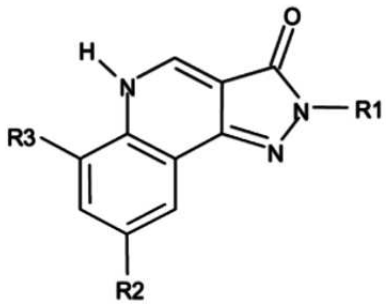
List of Tables

1 Structures of compounds (a) **1** to **8** and (b) **9** used for the validation of the docking and scoring protocol. 31
2 Score derived from XSCORE and few structural and physico-chemical properties calculated for the top 10 ligands of the first category and the ligands *2k* and *2l*. HBD and HBA stand for hydrogen bond donor and acceptor, respectively. d_{PL} measures the distance between the atom C^{δ} of Glu132 and the heavy atom of the ligand that is closest to this atom. 32

Table 1: Structures of compounds (a) **1** to **8** and (b) **9** used for the validation of the docking and scoring protocol.



Compd	Y	X	Z	IC_{50} (nM)
1	H	Cl	H	60 ± 17
2	CO_2H	F	H	70 ± 14
3	Cl	CO_2H	H	16 ± 2
4	F	CO_2H	H	210 ± 20
5	F	H	H	20 ± 3
6	H	F	F	1000 ± 250
7	H	H	F	300 ± 80
8	CO_2H	Cl	H	7 ± 3



Compd	*R1*	*R2*	*R3*	IC_{50} (nM)
9	Phenyl	H	H	6000

Table 2: Score derived from XSCORE and few structural and physico-chemical properties calculated for the top 10 ligands of the first category and the ligands *2k* and *2l*. HBD and HBA stand for hydrogen bond donor and acceptor, respectively. d_{PL} measures the distance between the atom C^{δ} of Glu132 and the heavy atom of the ligand that is closest to this atom.

Ligand	Masse (Da)	HBD	HBA	Log P	PSA (Å^2)	Rotatable bonds	d_{PL} (Å)	Score
1a	509.3	1	7	5.29	84	8	14.29	6.98
1b	474.3	2	8	1.83	112	6	8.35	6.96
1c	440.4	1	7	3.76	102	7	12.32	6.41
1d	445.3	3	7	3.94	96	10	11.65	6.35
1e	436.4	2	7	2.42	85	6	16.54	6.34
1f	527.4	1	6	5.91	65	15	12.95	6.30
1g	401.3	3	7	3.22	84	7	16.89	6.17
1h	452.3	2	8	2.30	114	9	10.19	6.15
1i	471.4	2	7	3.75	87	8	12.28	6.15
1j	401.3	3	7	3.43	84	7	15.59	5.98
2k	468.4	3	8	2.34	112	7	8.96	6.20
2l	469.4	1	8	1.98	123	7	3.89	5.36

Example 1 – Adhesion test on Ln

Red blood cell adhesion to laminin 521 was measured under flow conditions using Vena8 Endothelial+^{TM} biochips (internal channel dimensions: length 20 mm, width 0.8 mm, height 0.12 mm) and ExiGo^{TM} Nanopumps (Cellix Ltd, Dublin, Ireland). Recombinant human laminin 521 (BioLamina) at 5 ng/µl was immobilized on the internal surface of the biochips at 4°C overnight. RBCs from sickle cell disease patients were suspended at $5x10^7$ cells/ml in Hanks balanced salt solution, without calcium chloride and magnesium sulfate (Sigma-Aldrich) supplemented with 0.4% BSA, and incubated or not with molecule 1 (ligC), molecule 2 (ligA) or molecule 3 (ligB) at 0.1 mM, or DMSO as vehicle used to solubilize these molecules, for 30 min at room temperature and perfused through the biochip channels for 10 min at a shear stress of 0.5 dyn/cm^2 followed by a washout at the same shear stress with the buffer alone for 30 min. Five minutes washouts with the Hanks/0.4% BSA buffer were performed at 1, 2, 3, 4 and 5 dyn/cm^2. After each wash, adherent RBCs were counted in 7 representative areas along the centerline of each channel using the AxioObserver Z1 microscope and ZEN analysis software (Carl Zeiss, Le Pecq, France) and the mean number of adherent RBCs was determined for each condition and each washout step (Figure 1). Images of the same 7 areas were obtained throughout each experiment. Three adhesion assays were performed with 3 different blood samples. RBC adhesion to laminin was inhibited when the RBC suspension was incubated with Molecule 2 or 3, but not Molecule 1.

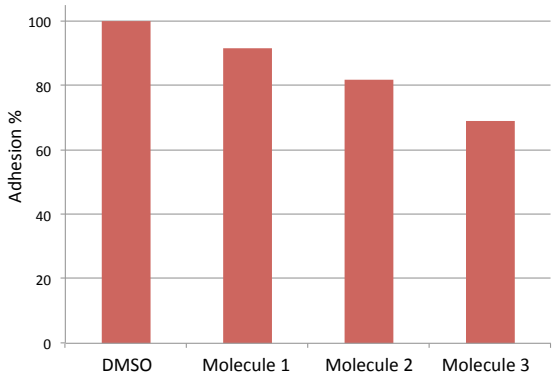


Figure 1: Adhesion of RBCs from sickle cell disease patients (N = 3) to laminin 521. Results are presented as a percentage of adhesion, 100% of adhesion being the adhesion level observed for RBCs incubated with buffer supplemented with DMSO alone, DMSO being the vehicle used to solubilize the three molecules.

Example 2 – Control of hemolysis

At the end of the 30 min of incubation with the molecules, measure by casy of the number of RBC for the control of hemolysis.

For hemolysis assays, RBCs were suspended at $6x10^6$ cells/ml in the Hanks/0.4% BSA buffer and the concentration measured accurately using the CASY cell counter. RBCs were then incubated or not with DMSO, molecule 1 (ligC), molecule 2 (ligA) or molecule 3 (ligB) at 0.1 mM for 30 min at room temperature. After this incubation time cell concentration was measured again using CASY (Table 1). If hemolysis occurs during this incubation time it will be reflected by a decrease in the RBC concentration.

Table 1.

Incubation condition	**RBCs/ml ($x10^6$)**
Buffer alone	6.125
Buffer + DMSO	6.04
Molecule 1 (ligC)	6.04
Molecule 2 (ligA)	5.55
Molecule 3 (ligB)	6.02

There was no decrease of RBC concentration in the presence of Molecule 1 or 3 compared to vehicle (DMSO) alone. A slight decrease of the concentration (< 10%) was observed in the presence of Molecule 2 suggesting a potential small hemolytic effect for this molecule.

Example 3 – Dose effect

The adhesion assays performed to address the dose effect of the molecules were performed in the same conditions as in Example 1, using the following concentrations for the 3 molecules: 50, 100 and 200 µM (Figure 2). Increasing concentrations of molecules 2 and 3 showed increasing inhibition of RBC adhesion to laminin. This inhibitory dose-dependent effect supports the specificity of both molecules to interact with Lu/BCAM and inhibit its interaction with laminin.

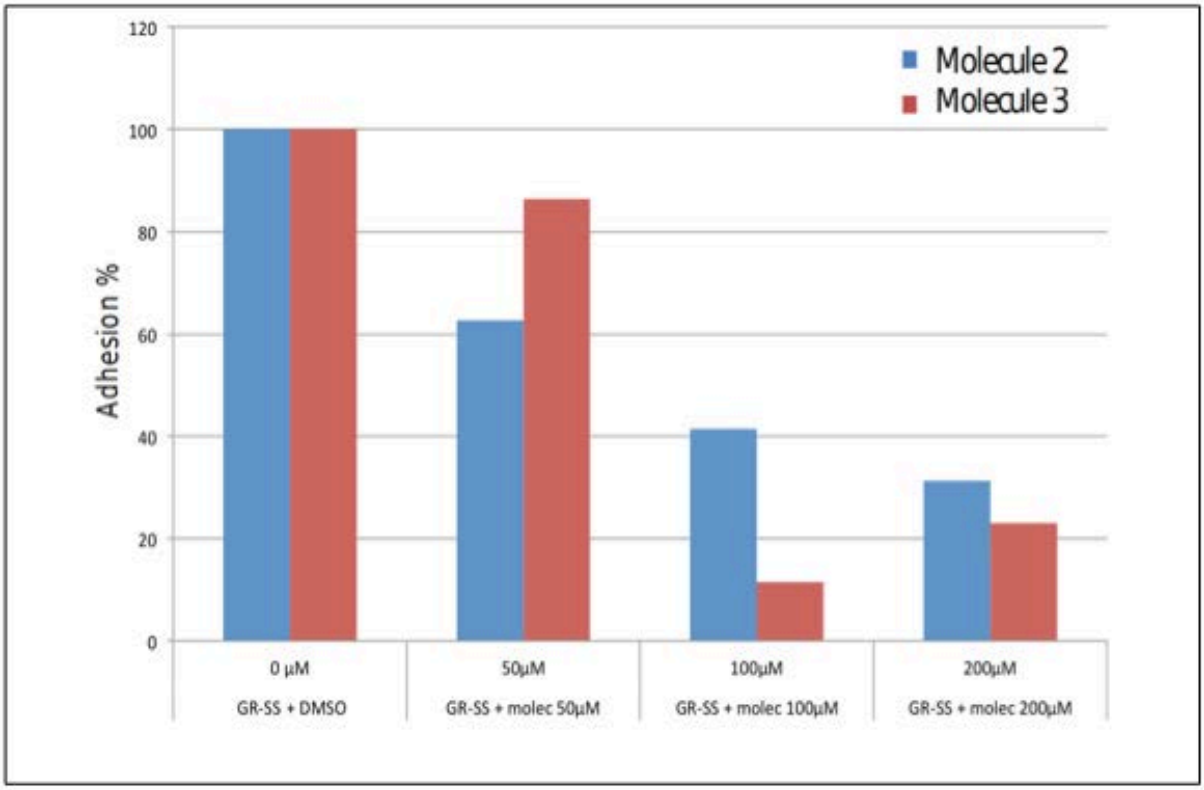


Figure 2: Adhesion of RBCs from sickle cell disease patients to laminin 521 in the presence of increasing concentrations of Molecule 2 and Molecule 3. Results are presented as a percentage of adhesion, 100% of adhesion being the adhesion level observed for RBCs incubated without the molecules.

Abstract

Drepanocytosis is a genetic blood disorder characterized by red blood cells that assume an abnormal sickle shape. In the pathogenesis of vaso-occlusive crises of sickle cell disease, red blood cells bind to the vascular endothelium and promote vaso-occlusion. At the surface of these sickle red blood cells, the overexpressed protein Lutheran (Lu) strongly interacts with the Laminin (Ln) 511/521. The aim of this study was to identify a protein-protein interaction (PPI) inhibitor with a high probability of binding to Lu for the inhibition of the Lu-Ln 511/521 interaction. A virtual screening was performed with 1 295 678 compounds that target Lu. Prior validation of a robust scoring protocol was considered on the protein CD80 because this protein has a binding site with similar topological and physico-chemical characteristics and it also has a series of ligands with known affinity constants. This protocol consisted of multiple filtering steps based on calculated affinities (scores), molecular dynamics simulations and molecular properties. A robust scoring protocol was validated on the protein CD80 with the docking program DOCK6 and the scoring functions XSCORE, MM-PBSA and the FMO method. This protocol was applied to the protein Lu and we found two compounds that were validated by *in vitro* studies. The protection of these ligands by a patent is under process. Nine other compounds were identified and seem to be promising candidates for inhibiting the Lu-Ln 511/521 interaction.

Résumé

La drépanocytose est une maladie génétique qui se caractérise par des globules rouges en forme de faucille. Chez les personnes atteintes de drépanocytose, ces globules rouges (GR) adhèrent à l'endothélium vasculaire et provoquent ainsi une vaso-occlusion. Ce phénomène s'explique par la surexpression, à la surface des globules rouges falciformes, de la protéine Lutheran (Lu) qui se lie fortement à la Laminine (Ln) 511/521 exprimée par l'endothélium vasculaire enflammé. Le but de cette étude est d'identifier des inhibiteurs d'interaction protéine-protéine (PPI) qui possèdent une forte probabilité de liaison à Lu afin d'inhiber l'interaction Lu-Ln 511/521. Un criblage virtuel de 1 295 678 composés ciblant la protéine Lu a été réalisé. La validation préalable d'un protocole de scoring a été envisagée sur la protéine CD80 qui présente un site de liaison avec des caractéristiques topologiques et physico-chimiques similaires au site de liaison prédit sur Lu ainsi que plusieurs ligands avec des constantes d'affinité connues. Ce protocole contient différentes étapes de sélection basées sur les affinités calculées (scores) et les propriétés d'interaction. Un protocole de scoring fiable a été validé sur CD80 avec le programme de docking DOCK6 et les fonctions de scoring XSCORE et MM-PBSA ainsi qu'avec la méthode de calcul FMO. L'application de ce protocole sur Lu a permis d'obtenir deux ligands validés par des tests *in vitro* qui font l'objet d'un dépôt de brevet auprès de l'Office Européen des Brevet. La fonction de scoring XSCORE a permis d'identifier neuf autres ligands qui semblent aussi être des candidats prometteurs pour inhiber l'interaction Lu-Ln 511/521.

POLE RECHERCHE
Ecoles **D**octorales

LETTRE D'ENGAGEMENT DE NON-PLAGIAT

Je, soussigné(e) .Noelly MADELEINE……,
en ma qualité de doctorant(e) de l'Université de La Réunion, déclare être conscient(e) que le plagiat est un acte délictueux passible de sanctions disciplinaires. Aussi, dans le respect de la propriété intellectuelle et du droit d'auteur, je m'engage à systématiquement citer mes sources, quelle qu'en soit la forme (textes, images, audiovisuel, internet), dans le cadre de la rédaction de ma thèse et de toute autre production scientifique, sachant que l'établissement est susceptible de soumettre le texte de ma thèse à un logiciel anti-plagiat.

Fait à Sainte Clotilde le : 31/08/2017

Signature :

Extrait du Règlement intérieur de l'Université de La Réunion
(validé par le Conseil d'Administration en date du 11 décembre 2014)

Article 9. Protection de la propriété intellectuelle – Faux et usage de faux, contrefaçon, plagiat

L'utilisation des ressources informatiques de l'Université implique le respect de ses droits de propriété intellectuelle ainsi que ceux de ses partenaires et plus généralement, de tous tiers titulaires de tels droits.
En conséquence, chaque utilisateur doit :
- utiliser les logiciels dans les conditions de licences souscrites ;
- ne pas reproduire, copier, diffuser, modifier ou utiliser des logiciels, bases de données, pages Web, textes, images, photographies ou autres créations protégées par le droit d'auteur ou un droit privatif, sans avoir obtenu préalablement l'autorisation des titulaires de ces droits.

La contrefaçon et le faux
Conformément aux dispositions du code de la propriété intellectuelle, toute représentation ou reproduction intégrale ou partielle d'une œuvre de l'esprit faite sans le consentement de son auteur est illicite et constitue un délit pénal.
L'article 444-1 du code pénal dispose : « Constitue un faux toute altération frauduleuse de la vérité, de nature à causer un préjudice et accomplie par quelque moyen que ce soit, dans un écrit ou tout autre support d'expression de la pensée qui a pour objet ou qui peut avoir pour effet d'établir la preuve d'un droit ou d'un fait ayant des conséquences juridiques ».
L'article L335_3 du code de la propriété intellectuelle précise que : « Est également un délit de contrefaçon toute reproduction, représentation ou diffusion, par quelque moyen que ce soit, d'une œuvre de l'esprit en violation des droits de l'auteur, tels qu'ils sont définis et réglementés par la loi. Est également un délit de contrefaçon la violation de l'un des droits de l'auteur d'un logiciel (…) ».

Le plagiat est constitué par la copie, totale ou partielle d'un travail réalisé par autrui, lorsque la source empruntée n'est pas citée, quel que soit le moyen utilisé. Le plagiat constitue une violation du droit d'auteur (au sens des articles L 335-2 et L 335-3 du code de la propriété intellectuelle). Il peut être assimilé à un délit de contrefaçon. C'est aussi une faute disciplinaire, susceptible d'entraîner une sanction.
Les sources et les références utilisées dans le cadre des travaux (préparations, devoirs, mémoires, thèses, rapports de stage…) doivent être clairement citées. Des citations intégrales peuvent figurer dans les documents rendus, si elles sont assorties de leur référence (nom d'auteur, publication, date, éditeur…) et identifiées comme telles par des guillemets ou des italiques.

Les délits de contrefaçon, de plagiat et d'usage de faux peuvent donner lieu à une sanction disciplinaire indépendante de la mise en œuvre de poursuites pénales.